

Patrick R. LeClair

Electricity, Magnetism & Optics

– an algebra-based introduction –

October 12, 2009

The University of Alabama
Tuscaloosa, Alabama

Contents

Part I Relativity

1	Relativity	3
1.1	Frames of Reference	3
1.2	Moving Frames of Reference	5
1.2.1	Lack of a Preferred Reference Frame	6
1.2.2	Relative Motion	7
1.2.3	Invariance of the Speed of Light	9
1.2.4	Principles of special relativity	10
1.3	Consequences of Relativity	11
1.3.1	Lack of Simultaneity	11
1.3.2	Time Dilation	12
1.3.3	Length Contraction	18
1.3.4	Time and position in different reference frames	21
1.3.5	Addition of Velocities in Relativity	24
1.3.6	Space-time Intervals	28
1.4	Mass, Momentum, and Energy	33
1.4.1	Relativistic Momentum	33
1.4.2	Relativistic Energy	34
1.4.3	Relativistic Mass	37
1.5	Problems	38

Part II Electricity and Magnetism

2	Electric Forces and Fields	43
2.1	Properties of Electric Charges	43
2.2	Insulators and Conductors	44
2.2.1	Charging by Conduction	45
2.2.2	Charging by Induction	47
2.3	Coulomb's Law	48
2.4	The Electric Field	50
2.4.1	Electric Field Lines	51
2.4.2	What happens when we have two charges together?	52
2.5	Conductors in Electrostatic Equilibrium	53
2.6	Faraday Cages	55
2.7	The van de Graaff Generator	56
2.8	Gauss' Law	58
2.8.1	Electric Flux	58
2.8.2	Gauss' Law as a Conservation Law	60
2.8.3	Example: The Field Around a Spherical Charge Distribution	60
2.8.4	Example: The Field Above a Flat Conductor	62
2.8.5	Example: The Field Inside and Outside a Hollow Spherical Conductor	63

2.8.6	Example: The Field Due to a Line of Charge	65
2.9	Miscellanea	66
2.10	Problems	67
3	Electrical Energy and Capacitance	71
3.1	Electrical Potential Energy	71
3.2	Electric Potential	74
3.2.1	Electric Potential and Potential Energy due to Point Charges	75
3.2.2	Energy of a System of Charges	76
3.3	Potentials and charged conductors	81
3.4	Equipotential Surfaces	82
3.5	Potential Difference Sources as Circuit Elements	83
3.6	Capacitance	83
3.6.1	Parallel-Plate Capacitors	84
3.6.2	Energy stored in capacitors	85
3.6.3	Capacitors as Circuit Elements	87
3.6.4	Combinations of Capacitors	88
3.6.5	Capacitors with (non-conducting) stuff inside	92
3.7	Dielectrics in Electric Fields	94
3.8	Problems	98
4	Current and Resistance	103
4.1	Electric Current	103
4.2	Getting Current to Flow	104
4.3	Drift Velocity and Current	106
4.4	Resistance and Ohm's Law	107
4.4.1	Drift Velocity and Collisions	108
4.4.2	Current, Electric Field, and Voltage	110
4.4.3	Resistance	111
4.4.4	Resistors as Circuit Elements	112
4.4.5	Resistivity of Materials	114
4.4.6	Variation of Resistance with Temperature	115
4.5	Electrical Energy and Power	116
4.6	Problems	117
5	Direct-Current Circuits	119
5.1	Sourcing Voltage	119
5.2	Sourcing Current	122
5.3	Combinations of Resistors	124
5.3.1	Resistors in Series	124
5.3.2	Resistors in Parallel	125
5.3.3	Example: a Complex Resistor Combination	127
5.4	Current and Voltage Measurements in Circuits	129
5.4.1	Measuring Voltage	129
5.4.2	Measuring Current	130
5.5	Kirchhoff's Rules and Complex dc Circuits	132
5.5.1	Example: analyzing a simple parallel or series circuit	133
5.5.2	Example: analyzing a complex circuit	134
5.6	RC Circuits	136
5.7	Miscellaneous	138
5.8	Problems	139

6	Magnetism	145
6.1	Magnetic Fields and Forces	145
6.1.1	The Magnetic Force	147
6.1.2	Magnetism as a Consequence of Relativity	148
6.1.3	Magnetic Field of a Long, Straight Wire	151
6.1.4	Handedness	153
6.2	Ampère's Law	154
6.2.1	Ampère's and Gauss' Laws	155
6.3	The Magnetic Field in Various Situations	156
6.3.1	Motion of a Charged Particle in a Magnetic Field	156
6.3.2	Magnetic Force on a Current-Carrying Conductor	159
6.3.3	Magnetic Force Between Two Parallel Conductors	160
6.3.4	Torque on a Current Loop	162
6.3.5	Magnetic Fields of Current Loops and Solenoids	163
6.4	Permanent Magnetic Materials	165
6.4.1	Non-permanent magnetic materials	167
6.4.2	Electromagnets	167
6.4.3	Permeability and Magnets on Your Fridge	168
6.5	Problems	170
7	Induced Voltages and Inductance	171
7.1	Induced Voltages and Magnetic Flux	171
7.2	Faraday's Law of Induction	172
7.3	Inductance	173
7.3.1	Mutual Inductance	173
7.3.2	Self Inductance	174
7.4	Transformers	178
7.5	Voltage Induced by the Motion of a Conductor in a Field	179
7.5.1	Eddy current brakes	181
7.6	Generators	182
7.7	A summary of sorts	183
8	ac Circuits and Electromagnetic waves	185
8.1	Resistors in an ac Circuit	185
8.2	Capacitors in ac Circuits	187
8.3	Inductors in ac Circuits	189
8.4	Filters	190
8.5	Electromagnetic Waves	192
8.5.1	Electromagnetic fields of accelerating charges	192
8.5.2	Production of Electromagnetic Waves by an Antenna	192
8.5.3	Properties of Electromagnetic Waves	193
8.5.4	Energy transferred by EM waves	194
8.5.5	The EM spectra	196

Part III Optics

9	Reflection and Refraction of Light	199
9.1	The Nature of Light	199
9.1.1	Wave Packets and Wave-Particle Duality	200
9.2	Reflection of Light	201
9.2.1	The Ray Approximation	201
9.2.2	The Law of Reflection	201
9.3	Refraction of Light	202
9.3.1	Snell's Law	204
9.3.2	Dispersion and Prisms	205
9.3.3	Rainbows	207

9.4	Total Internal Reflection	208
9.4.1	Fiber optics	209
9.4.2	Multi-Touch screens	210
10	Mirrors	213
10.1	Flat Mirrors	213
10.1.1	Image formation	213
10.1.2	Ray Diagrams	215
10.1.3	Conventions for Ray Diagrams	216
10.1.4	Handedness	217
10.2	Spherical Mirrors	217
10.2.1	Concave Mirrors	217
10.2.2	Convex Spherical Mirrors	221
10.3	Ray Diagrams for Mirrors	222
10.4	Parabolic Mirrors	223
11	Lenses	227
11.1	Spherical refracting surfaces	227
11.1.1	Flat Refracting Surfaces	230
11.2	Spherical Lenses	230
11.3	Types of spherical lenses	234
	Solutions to Problems	237
	Bibliography	276
	References	276

Part I

Relativity

Chapter 1

Relativity

The difference between truth and fiction is that fiction has to make sense. – Mark Twain

Abstract Nearly all of the mechanical phenomena we observe around us every day have to do with objects moving at speeds rather small compared to the speed of light. The Newtonian mechanics you learned in previous courses handled these cases extraordinarily well. As it turns out, however, Newtonian mechanics breaks down completely when an object's speed is no longer negligible compared to the speed of light. Not only does Newtonian mechanics fail in this situation, it fails *spectacularly*, leading to a variety of paradoxical situations. The resolution to these paradoxes is given by the theory of relativity, one of the most successful and accurate theories in all of physics, which we will introduce in this chapter. Nature is not always kind, however, and the consequences of relativity seem on their face to flout common sense and our view of the world around us. We are used to the notion that our position changes with time when we are in motion, but relativity implies that *passage of time itself* changes when we are in motion. Nevertheless, we shall see that relativity is an *inescapable* consequence of a few simple principles and experimental facts. Moreover, as it turns out, this new description of nature is critical for properly understanding electricity and magnetism, optics, and nuclear physics ... most of the rest of this course!

1.1 Frames of Reference

Describing motion properly usually requires us to choose a coordinate system, and an origin from which to measure position. Why this is so is more clear when we consider the difference between distance and displacement. For example, we can say that a person moves through a *displacement* of 10 meters, $\Delta x = 10\text{m}$, in a particular direction, *e.g.*, to the right. This does not describe the *position* of the person at all, only the *change* in that person's position over some time interval.

Describing position itself requires us to choose first a coordinate system (such as cartesian, spherical, *etc.*), and also an origin for this coordinate system to define our “zero” position. The essential difference is that displacement is independent of the coordinate system we choose, but position is not. Without choosing a coordinate system, we can only say that the person has run 10m in a certain time interval, moving from x_i to x_f .

As a concrete example, consider Fig. 1.1a. This illustrates a person moving 10m to the right, which perfectly describes a *displacement* Δx . We will choose an $x - y$ cartesian coordinate system, which we will call O , with its origin at the person's starting point. In this system, we can describe the initial and final positions P_i^O and P_f^O in this coordinate system as $P_i^O = (0, 0)$ and $P_f^O = (x_f, 0) = (\Delta x, 0)$. This is shown in Fig. 1.1b. The displacement is the same as it was without a coordinate system. In this chapter we will use the convention that superscripts refer to the coordinate system in which the quantity in question was measured.

What happens if we if we instead choose a different coordinate system O' , Fig. 1.1c, identical to O except that its origin is shifted downward by y_i and to the left by x_i ? Now the initial and final positions of the person are $P_i^{O'} = (x_i, y_i)$ and $P_f^{O'} = (x_f, y_f)$. Still, the displacement Δx is the same, as you can easily verify. No matter whether we observe the person from the O or O' system, we would describe the same *displacement*, even though the actual positions are completely different.

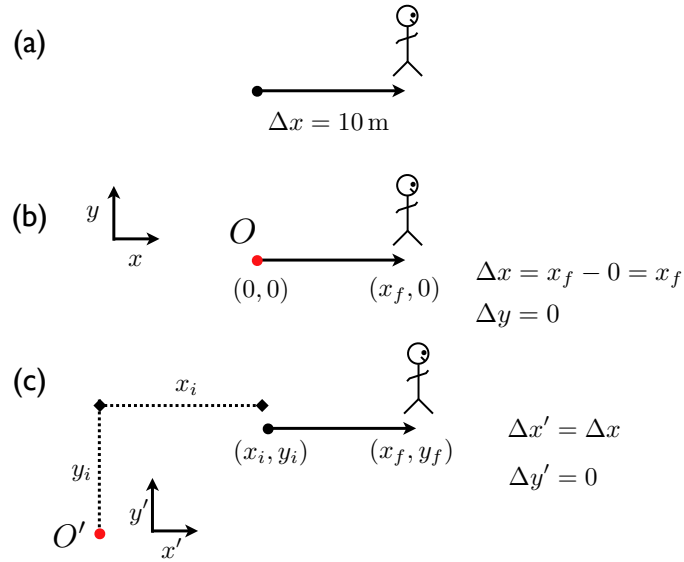


Fig. 1.1 Displacement is independent of the coordinate system we choose, but position is not. **(a)** Without choosing a coordinate system, we can only say that the person has run 10m in a certain time interval, moving from x_i to x_f . **(b)** If we choose an $x-y$ coordinate system O centered with its origin on the person's starting point x_i , we can describe the initial and final positions as $P_i^O = (x_i, 0)$ and $P_f^O = (x_f, 0)$. The displacement is the same. **(c)** If we choose a new coordinate system O' , identical to O except shifted downward by y_i and to the left by x_i , now the initial and final positions are $P_i^{O'} = (x_i, y_i)$ and $P_f^{O'} = (x_f, y_i)$. Still, the displacement is the same.

In special relativity, this simple situation no longer holds - observers in different coordinate systems do *not* necessarily describe even the same displacement, much less the same position. Fortunately, the corrections of special relativity to the Newtonian mechanics you have already learned are only appreciable at very high velocities (non-negligible compared to the speed of light), and for most every day situations our usual intuition is still valid.

In any case, particularly those cases where relativistic effects are important, it is crucially important that we specify in which coordinate system quantities have been measured. We will continue to do this with a superscript of some sort to specify the coordinate system, and a subscript of some sort to further describe what is being measured *within* that system. When we only have two frames, like the example above, we will often just use a prime ($'$) to tell them apart. In the previous example, this means we would use P_f' instead of $P_f^{O'}$, and just P_i instead of P_i^O . It seems pedantic now, but careful bookkeeping is the only thing saving us from terrible confusion later!

Coordinate system notation examples:

$x_{\text{final}}^O = x_f^O$ final x position of an object measured in the O coordinate system

$v_{\text{car}}^{O'} \equiv v'_{\text{car}}$ velocity of a car measured in the O' coordinate system

$P_f^{O'} \equiv P'_f = (x'_f, y'_i)$ final position of an object measured in the O' coordinate system

Finally, a word on terminology. In relativity, it is common to use “reference frame” in place of “coordinate system,” to make explicit the fact that our coordinate system and origin are the point of reference from which we measure physical quantities. We will use both phrases interchangeably from here on out.

1.2 Moving Frames of Reference

What about one observer measuring in a coordinate system *moving* at constant velocity relative to another? For example, take Fig. 1.2. A girl holding balloons is standing on the ground, and a bully on a skateboard throws a dart at her balloons. The bully is moving at a velocity v_{bully} relative to the girl, and he throws the dart at a velocity v_{dart} relative to himself. What is the dart's speed relative to the girl?

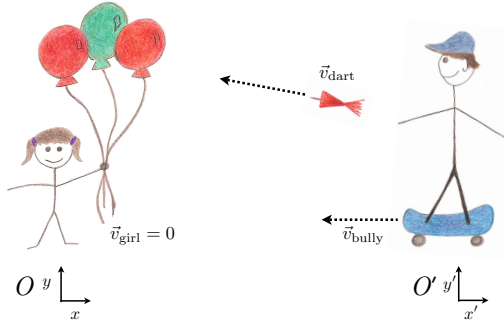


Fig. 1.2 A girl holding balloons is standing on the ground, at rest in reference frame O ($v_{\text{girl}}^O = 0$). Meanwhile a bully on a skateboard throws a dart at her balloons. The bully is moving at a velocity v_{bully}^O relative to the girl's reference frame, and he throws the dart at a velocity $v_{\text{dart}}^{O'}$ relative to himself (the O' frame). What is the dart's speed as measured by the girl? Drawings by C. LeClair

First of all, we have to be more explicit about specifying which quantity is measured in which frame. The velocity of the bully on the skateboard is measured relative to the girl standing on the ground, in the O system, so we write v_{bully}^O . When we talk about the dart, however, things are a bit less clear. The bully on the skateboard would say that the velocity of the dart is $v_{\text{dart}}^{O'}$, since he would measure its velocity relative to *himself* in the O' frame. The girl would measure the velocity of the dart relative to *herself* in the O frame, v_{dart}^O . Clearly, $v_{\text{dart}}^{O'} \neq v_{\text{dart}}^O$ – in principle, the two cannot agree on what the velocity of the dart is! Of course, that is a bit of an exaggeration. In this simple everyday case, relative motion is fairly easy to understand, and we can intuitively see exactly what is happening. Our intuition will start to fail us shortly, however, so it is best we proceed carefully.

Explicitly labeling the velocity with the reference frame in which it is measured helps keep everything precise, and helps us find a way out of this conundrum. It may seem like baggage now, but ambiguity would cost us dearly later. Just to summarize, here is how we will keep the velocities straight:

$$\begin{aligned} v_{\text{bully}}^O &= \text{velocity of bully measured from the ground} \equiv v_{\text{bully}} \\ v_{\text{dart}}^{O'} &= \text{velocity of dart measured from the skateboard} \equiv v'_{\text{dart}} \\ v_{\text{dart}}^O &= \text{velocity of the dart measured by the girl} \equiv v_{\text{dart}} \end{aligned}$$

Whenever we are only dealing with two different coordinate systems, we will trim down the notation a bit. We will just call one system the “primed” system and add a $'$ superscript to all quantities, and leave the other one as the “unprimed” system, dropping the ‘ O ’. Which one we call “primed” and which one is “unprimed” makes no difference, it is after all just notation.

What does the girl on the ground, in the O system really observe? Intuitively, we expect this her to see the dart moving at a velocity v_{dart} which is that of the dart relative to the skateboard *plus* that of skateboard relative to the ground:

$$v_{\text{dart}} = v'_{\text{dart}} + v_{\text{bully}} \quad (1.1)$$

velocity of the dart seen by the girl = velocity of dart relative to skateboard + velocity of skateboard relative to girl

The bully, in the O' system (who threw the dart in the first place), just sees v'_{dart} . Just to be concrete, let's say that the bully on the skateboard moves with $v_{\text{bully}} = 3 \text{ m/s}$, and he throws the dart with $v'_{\text{dart}} = 2 \text{ m/s}$. Then the girl sees the dart coming at her balloons at 5 m/s .

1.2.1 Lack of a Preferred Reference Frame

Even in the simple example above, *velocity depends on your frame of reference*. This simple example is completely arbitrary in a sense, though, and implies much more about relative motion. If these two observers can't agree on the velocity of the dart, as measured in their own reference frames, who is to say what the absolute reference frame should be? After all, isn't the ground itself moving due to the rotation of the earth about the sun? And isn't the sun moving relative to the center of the galaxy? *Nothing is absolutely at rest, we cannot pick any special frame of reference to define absolute unique velocities.*

Still, we might think be tempted to think that there is some sort of reference frame we are forgetting, one that is truly at rest. For instance, what about empty space itself? Can we define absolute coordinates and absolute motion relative to specific points in space? This is a tempting thought, particularly if we make an analogy with sound waves.

As you know from Mechanics, sound is really nothing more than (longitudinal) oscillations of matter, a sort of density wave in a material. We will find out in later Chapters that light is also a wave. If they are both waves, perhaps the nature of sound can help explain the nature of light? Sound can be propagated through matter, or even through air, but it requires a medium to be transmitted – no sound is transmitted in a vacuum. Could we view light as the vibrations of space itself, or of some all-pervasive “fluid” filling all of space? Certainly light waves also need a medium in which to propagate, so the reasoning goes. This all-pervasive fluid would provide a “background” frame of reference, allowing us to measure absolute velocity, somewhat like measuring the velocity of a boat by how fast water moves past its side.

Indeed, this was a very attractive viewpoint through the early 20th century, and “luminiferous æther” was the term used to describe the all-pervasive medium for the propagation of light. It fact, is a *testable* idea – this is a crucial point which makes the idea a true scientific theory. How do we test it? If space itself has a background medium within which light propagates, then we should be able to measure the velocity of the earth through this medium as it revolves around the sun. The earth moving through the æther fluid would experience some “drag,” again just like a boat moving through water.

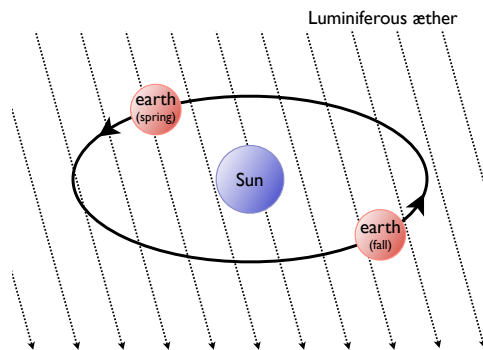


Fig. 1.3 If there were a “luminiferous æther” which light propagates on pervading all space, the earth’s revolution around the sun would experience a drag force depending on the season – the “æther wind.” At some times, this æther wind would augment the speed of light, and at others it would diminish it. Analyzing the speed of light in different directions at different times of year should allow one to extract the æther velocity. Instead, experiments proved that there is no æther.

Unfortunately, this idea just isn't right. It has been disproven countless times by experiments, and replaced by the far more successful theory of relativity. Light waves are not like sound waves. There is no æther, there is no preferred frame of reference, and all motion is relative. Why this must be the case, and how it arises is what we need to figure out next

1.2.2 Relative Motion

Fine. There is no preferred reference frame or coordinate system, and all motion is relative. So what? The example of Fig. 1.2 was plainly understandable. It is disturbingly easy to come up with examples which are *not* so plainly understandable, however, which is one motivation for the theory of relativity in the first place. Consider the two rockets in empty space traveling toward each other in Fig. 1.4, separated by a distance Δx . The pilot of rocket 1 might say he or she is traveling at a speed v_1 in his or her own reference frame (O), and the pilot of rocket 2 may claim he or she is traveling at a speed v_2' in their own O' . Without specifying what point they are measuring their velocity relative to, can we say who is moving at what speed?

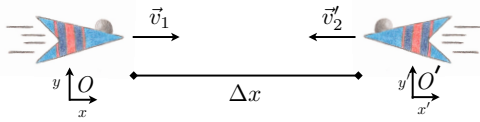


Fig. 1.4 Two identical rockets, separated by a distance Δx , are moving toward each other in empty space. Rocket 1 sees rocket 2 cover a distance $(v_1 + v_2)t$ in a time t , as if rocket 2 is heading toward it. On the other hand, rocket 2 sees rocket 1 cover the *same* distance $(v_1 + v_2)t$ in a time t , as if rocket 1 is heading toward it! Both of them cannot be correct. Without an external reference frame, it is impossible to say who is moving, and at what speed. Drawings by C. LeClair

We have to imagine that we are deep in empty space, with nothing around either rocket to provide a landmark or point of reference. The occupants of rocket 1 would feel as though they are sitting still, and observe rocket 2 coming toward them, covering a distance $(v_1 + v_2)\Delta t$ in a time interval Δt . The occupants of rocket 2, on the other hand, would think *they* are sitting still, and would observe rocket 1 coming toward *them*, also covering a distance $(v_1 + v_2)\Delta t$ in a time interval Δt .

Without any external reference point, or an absolute frame of reference, *not only can we not say with what speed each rocket is moving, we can't even say who is moving!* If we decide that rocket 1 is our reference frame, then it is sitting still, and rocket 2 is moving toward it. But we could just as easily pick rocket 2 as our reference frame. Specifying who is moving, and with what speed, is meaningless without a proper origin or frame of reference.

Has anything really changed physically? No. An analogy of sorts is to think about driving along side other cars on the highway, keeping pace with them. You might report your speed as 60 mi/hr. Relative to what? Clearly, in this case it is implied that the ground beneath you provides a reference frame, and you are talking about your velocity relative to the earth. You wouldn't say you are traveling at 60 mi/hr relative to the other cars (we hope) – your speed relative to the other cars is zero if you are staying along side them. Indeed, if you look out your window, the cars next to you appear to be sitting still. This is only true at constant velocity – we can easily detect accelerated motion, or an accelerated frame of reference due to the force experienced. Handling accelerated motion properly is the realm of *general* relativity, somewhat beyond the scope of our discussion.

In the end, one of the fundamental principles of special relativity is that a relative constant velocity does not matter, so far as the laws of physics are concerned. The laws of physics apply the same way to all objects in uniform (non-accelerated) motion, no matter how we measure the velocity. We *cannot* devise an experiment to measure uniform motion absolutely, only relative to a specific chosen frame of reference. More succinctly:

Principle of relativity:

All laws of nature are the same in all uniformly moving (non-accelerating) frames of reference.
No frame is preferred or special.

As another simple example, Fig. 1.5, consider Joe and Moe running at different (constant) speeds in the same direction, initially separated by a distance d_o . Without specifying any particular common

frame of reference, we must be able to describe their relative motion, or how the separation between Joe and Moe changes with time, even though we can't speak of their absolute velocities in any sense.

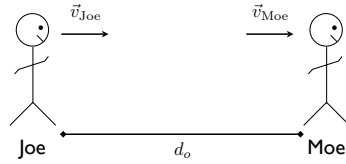


Fig. 1.5 Joe and Moe running at different speeds in the same direction. Both Joe and Moe measure the same *relative* velocity with respect to each other.

Let's say we arbitrarily choose Joe's position at $t = 0$ as our reference point. It is easy then to write down what Joe and Moe's positions are at any later time interval Δt :

$$x_{\text{Joe}} = v_{\text{Joe}} \Delta t \qquad x_{\text{Moe}} = d_0 + v_{\text{Moe}} \Delta t \qquad (1.2)$$

We can straightforwardly write down the separation between them (their relative displacement) as well:

$$\Delta x_{\text{Moe-Joe}} = x_{\text{Moe}} - x_{\text{Joe}} = d_0 + v_{\text{Moe}} \Delta t - v_{\text{Joe}} \Delta t = d_0 + (v_{\text{Moe}} - v_{\text{Joe}}) \Delta t \qquad (1.3)$$

Sure enough, their relative displacement only depends on their starting separation and their *relative* velocity, $v_{\text{Moe}} - v_{\text{Joe}}$. Further, both Joe and Moe would agree with this, since we could arbitrarily choose *Moe's* position at $t = 0$ as our reference point, and *we would end up with the same answer*. Since there is nothing special about either position, we can choose *any* point whatsoever as a reference, and wind up with the same result. We end up with the same physics no matter what reference point we choose, which one we choose is all a matter of convenience in the end.

Choosing a coordinate system:

1. Choose an origin. This may coincide with a special point or object given in the problem - for instance, right at an observer's position, or halfway between two observers. Make it convenient!
2. Choose a set of axes, such as rectangular or polar. The simplest are usually rectangular or *Cartesian* x - y - z , though your choice should fit the symmetry of the problem given - if your problem has circular symmetry, rectangular coordinates may make life difficult.
3. Align the axes. Again, make it convenient - for instance, align your x axis along a line connecting two special points in the problem. Sometimes a thoughtful but less obvious choice may save you a lot of math!
4. Choose which directions are positive and negative. This choice is arbitrary, in the end, so choose the least confusing convention.

This seems simple enough, but if we think about this a bit longer, more problems arise. Who measures the initial separation d_0 , Joe or Moe? Who keeps track of the elapsed time Δt ? Does it matter at all, can the measurement of distance or time be affected by relative motion? Of course, the answer is an awkward 'yes' or we would not dwell on this point. If we delve deeper on the problem of relative motion, we come to the inescapable conclusion that not only is velocity a relative concept, our notions of distance and time are relative as well, and depend on the relative motion of the observer. In order to properly understand these deeper ramifications, however, we need to perform a few more thought experiments.

1.2.3 Invariance of the Speed of Light

Already, relativity has forced us to accept some rather non-intuitive facts. This is only the beginning! A more fundamental and far-reaching principle of relativity is that *the speed of light is a constant, independent of the observer*. No matter how we measure it, no matter what our motion is relative to the source of the light, we will always measure its velocity to be the same value, c . Light does not obey the principle of relative motion!

Speed of Light in a Vacuum:

$$c \equiv 299792458 \text{ m/s} \approx 3 \times 10^8 \text{ m/s}$$

The numerical value of c is a fixed, exact value.

See, for example, http://en.wikipedia.org/wiki/Speed_of_light.

There is a relatively simple way to experimentally demonstrate that this seems to be true, depicted in Fig. 1.6. The earth itself is in constant motion in its orbit around the sun, moving at $\sim 3 \times 10^4 \text{ m/s}$ measured relative to distant stars (this in itself is a measurable quantity). Imagine now that we carefully set up three lasers, each oriented in a different direction relative to earth's orbital velocity – one parallel (A), one antiparallel (B), and one at a right angle (C). We will further set up each laser to emit short pulses of light, and carefully measure the time between pulses. In this way, we can determine the speed of the light coming out of each laser.

Based on simple Newtonian mechanics and velocity addition, we would expect to measure a slightly different velocity for each laser. In case A, we would expect the Earth's velocity to *add* to that of light, $v_A = v_{\text{light}} + v_{\text{orbit}}$, while in case B, it should *subtract*, $v_B = v_{\text{light}} - v_{\text{orbit}}$. In case C, we have to add vectors, $\vec{v}_C = \vec{v}_{\text{light}} + \vec{v}_{\text{orbit}}$, but the idea is the same.

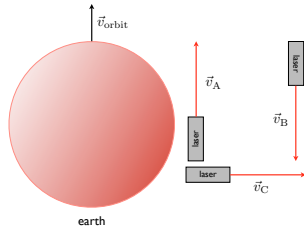


Fig. 1.6 If the velocity of light obeys Newtonian mechanics, then measuring the speed of light from a laser pointed in different directions compared to Earth's orbital velocity should yield different results. In case A, we would expect the Earth's velocity to *add* to that of light, $v_A = v_{\text{light}} + v_{\text{orbit}}$, while in case B, it should *subtract*, $v_B = v_{\text{light}} - v_{\text{orbit}}$. In case C, we have to add vectors, but the idea is the same. The effect should be small ($\sim 0.01\%$), but easily measurable. No effect has ever been observed, the speed of light *always* has the same value $c \approx 3 \times 10^8 \text{ m/s}$.

The effect should be small ($\sim 0.01\%$, given earth's mean orbital velocity of $3 \times 10^4 \text{ m/s}$ relative to the sun), but easily measurable. No effect is observed, the speed of light is *always* the same value c . This experiment has been performed with increasingly fantastic precision over the last 100 years[1], and no matter what direction we shine the light, we always measure the same speed! (The current best limit[1] on the constancy of the speed of light is about 1 part in 10^{16} .) One straightforward result of this experiment is that the idea of an æther is clearly not right, as we discussed above. There are much more far-reaching consequences, which we must consider carefully. First, let us re-iterate this idea more formally:

The speed of light is invariant

The speed of light in free space is *independent* of the motion of the source or observer. It is an invariant constant.

This is not just idle speculation or theory, it has been confirmed again and again by careful experiments. These experiments have established, for instance, that the speed of light¹ does not depend on the wavelength of light, on the motion of the light source, or the motion of the observer. As examples, lack of a wavelength dependence can be strongly ruled out by astronomical observations of gamma ray bursts (to better than 1 part in 10^{15}), while binary pulsars can rule out any dependence on source motion. The lack of a dependence on observer motion was disproved along with the æther (Sec. 1.2.1), which also proved that light requires no medium for propagation.

As an example of this, we turn again to Joe and Moe (Fig. 1.7). Joe is in a rocket (O'), traveling at 90% of the speed of light ($v=0.9c$), while Moe is on the ground (O) with a flashlight. Moe shines the flashlight parallel to Joe's trajectory in the rocket. On first sight, we would think that Moe would measure the speed of the light leaving the flashlight as c , while Joe would measure $v=c-0.9c=0.1c$.

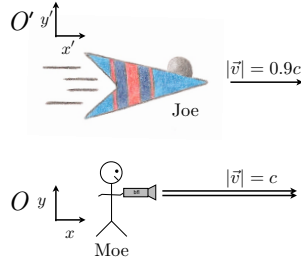


Fig. 1.7 Joe is traveling on a rocket at $|\vec{v}| = 0.9c$, while Moe on the ground shines a flashlight parallel to Joe's path. Both Joe and Moe observe the light from the flashlight to travel at $|\vec{v}| = c$ – contrary to our intuition from Newtonian Mechanics.

Both Joe and Moe measure *the same speed of light c* , despite their relative motion! What if we gave Joe the flashlight inside the rocket? No difference – both Joe and Moe measure the speed of the light to be c . Think back to our example of relative motion in Fig. 1.2. It doesn't seem to make sense that light behaves differently, but that is how it is. By the end of this chapter, though, we will be armed with the proper tools to analyze this situation correctly from both viewpoints and understand *how* both Joe and Moe manage to measure the same speed of light.

1.2.4 Principles of special relativity

From our discussions so far, relativity when non-accelerating (inertial) reference frames are considered has two basic principles which underpin the entire theory:

Principles of special relativity

1. **Special principle of relativity:** Laws of physics look the same in all inertial (non-accelerating) reference frames. There are no preferred inertial frames of reference.
2. **Invariance of c :** The speed of light in a vacuum is a universal constant, c , independent of the motion of the source or observer.

This theory of relativity restricted to inertial reference frames is known as the *special theory of relativity*, while the more general theory of relativity which also handles accelerated reference frames is simply known as the *general theory of relativity*.

The second postulate of special relativity - the invariance of the speed of light - can actually be considered as a consequence of the first according to some mathematical formulations of special relativity. That is, the constancy of the speed of light is *required* in order to make the first postulate true. We will continue to hold it up as a second primary postulate of special relativity, however, as

¹ Throughout this chapter, we refer to the speed of light in a *vacuum*.

some of the more non-intuitive consequences of special relativity are (in our view) more readily apparent when one keeps this fact in mind.

The first principle of relativity essentially states that all physical laws should be exactly the same in any vehicle moving at constant velocity as they are in a vehicle at rest. As a consequence, at constant velocity we are incapable of determining absolute speed or direction of travel, we are only able to describe motion relative to some other object. This idea does not extend to accelerated reference frames, however. When acceleration is present, we feel fictitious forces that betray changes in velocity that would not be present if we were moving at constant velocity. All experiments to date agree with this first principle: physics is the same in all inertial frames, and no particular inertial frame is special.

The principle of relativity is by itself more general than it appears. The principle of relativity describes a symmetry in the laws of nature, that the laws must look the same to one observer as they do to another. In physics, any symmetry in nature also implies a *conservation law*, such as conservation of energy or conservation of momentum. If the symmetry is in time, such that two observers at different times must observe the same laws of nature, then it is energy that must be conserved. If two observers at different physical locations must observe the same laws of physics (*i.e.*, the laws of physics are independent of spatial translation), it is linear momentum that must be conserved. The relativity principles imply deep conservation laws about space and time that make testable predictions – predictions which must be in accordance with experimental observations in order to be taken seriously. Relativity is not just a principle physicists have proposed, it is a postulate that was *required* in order to describe nature as we see it. The consequences of these postulates will be examined presently.

1.3 Consequences of Relativity

We have our principles laid forth, and their rationale clearly provided by our series of thought experiments. All experimental results to date are on the side of these two principles. The invariance of the speed of light and the principles of relativity force us to modify our very notions of perception and reality. We are not just fiddling with a few equations to handle special high-velocity cases, we must reevaluate some of our deepest intuitions and physical models. Many books have been written about the implications that relativity has had on philosophy in fact ... however, we will stick to physics.

1.3.1 Lack of Simultaneity

The speed of light is more than just a constant, it is a sort of ‘cosmic speed limit’ – no object can travel faster than the speed of light, and no information can be transmitted faster than the speed of light. If either were possible, causality would be violated: in some reference frame, information could be received before it had been sent, so the ordering of cause-effect relationships would be reversed. It is a bit much to go into, but the point is this: the speed of light is really a *speed limit*, because if it were not, either cause and effect would not have their usual meaning, or sending information backward in time would be possible. Neither is an easily-stomached possibility. A more readily grasped consequence of all of this is that we must give up on the notion of two events being simultaneous in any absolute sense – whether events are viewed as simultaneous depends on ones reference frame! It should seem odd that a seemingly simple principle like the speed of light being constant would muck things up so much, but in fact we can demonstrate that this *must* be true with a simple thought experiment.

Imagine that Joe is flying in a spaceship at $v = 0.9c$ (we will call his reference frame O'), and Moe is observing him on the ground (in frame O), as shown in Fig. 1.8. Joe, sitting precisely in the middle of the ship, turns on a light at time $t = 0$ also in the middle of the spaceship. A small amount of time Δt later, Joe’s superhuman eyes observe the rays of light reach the front and the back of the

spaceship simultaneously. So far this makes sense – if the light is exactly in the middle of the ship, light rays from the bulb should reach the front and back at the same time.

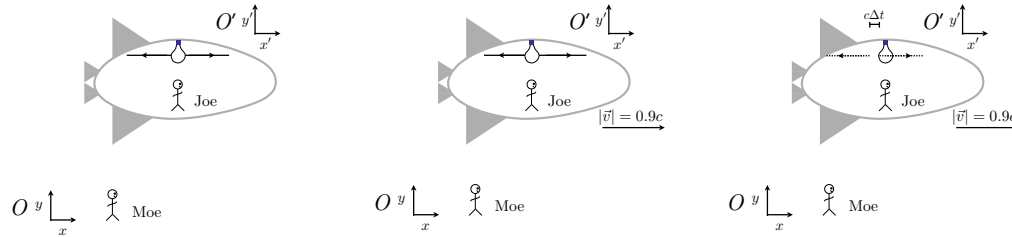


Fig. 1.8 **left:** Joe is traveling in a (transparent) rocket ship, and turns on a light bulb in the exact center of the rocket. **middle** A short time Δt later, in his frame O' Joe sees the light rays hit both sides of the ship at the same time. **right:** Moe on the ground observes Joe in his rocket moving at $v = 0.9c$. From his frame O , a time Δt after the light leaves the bulb, the ship moves forward by an amount $c\Delta t$ but the light rays do not. Moe sees the light hit the back of the ship first – Moe and Joe cannot agree on the simultaneity of events.

Now, what will Moe on the ground see?² From his frame O , Moe sees the light emitted from the bulb at $t = 0$. The ship and the light bulb are both moving relative to Moe at $v = 0.9c$, but we have to be careful. First, Moe observes the same speed of light as Joe, even though the bulb is moving. Once the light bulb is turned on, the first light leaves the bulb at $v = c$ and diverges radially outward from its point of creation. As this first light leaves the bulb, however, *the ship is still moving forward*. The front of the ship moves away from the point of the light's creation, while the back moves *toward* it.

Consequence of an invariant speed of light:

Events that are simultaneous in one reference frame are **not** simultaneous in another reference frame moving relative to it – and no particular frame is preferred. Simultaneity is not an absolute concept.

In some sense, once the light is created, it isn't really in either reference frame – it is traveling at $v = c$ no matter who observes it. The ship moved forward, but the point at which the light was created did not. We attempt to depict this in Fig. 1.8, where from Moe's point of view, after a time Δt the light rays emitted from the bulb seem to have emanated from a point somewhat behind where the bulb is at Δt – a distance $c\Delta t$ behind it. Thus, after some time, Moe sees the light hit *the back* of the ship first! Joe and Moe seem to observe different events, and they can not agree on whether the light hits the front and the back of the ship simultaneously. Events which are simultaneous in Joe's reference frame are **not** in Moe's reference frame, moving relative to him. Think about how this plays out from Joe and Moe's reference frames carefully. It is strange and non-intuitive, but if we accept the speed of light as invariant, the conclusion is inevitable, and causality is preserved.

1.3.2 Time Dilation

Now we have already seen that the constancy of the speed of light has some rather unintuitive and bizarre consequences. In fact, our concept of the passage of time itself must be “corrected.” Just as the notion of two events being simultaneous or not depend on one's frame of reference, the relative passage of time also depends on the frame of reference in which the measurement of time is made. Again, to illustrate this idea, we will perform a thought experiment.

First, we need a way to measure the passage of time. The constancy of the speed of light fortunately provides us with a conceptually straightforward – if not experimentally simple – manner in

² You might think nothing, as we neglected to mention that Joe's ship is transparent.

which to do this. We will measure the passage of time by bouncing light pulses between two parallel mirrors, carefully placed a distance d apart. Since we know the speed of light is an immutable constant, so long as the space between the mirrors remains fixed at d , the round-trip time Δt for a pulse of light to start at one mirror, bounce off the second, and return to the first will be a constant. The light pulse travels the distance d between the mirrors, and back again, at velocity c , so the time interval for a round trip is just $\Delta t = 2d/c$.

Now, let's imagine Joe is performing this experiment in a boxcar moving at velocity v relative to the ground, as shown in Fig. 1.9a. We will label Joe's own reference frame inside the boxcar as O' , such that the boxcar moves in the x' direction. Both Joe, the mirrors, and the light source are stationary relative to one another, and the mirrors and light source have been carefully positioned a distance d apart such that the light pulses propagate vertically in the y' direction. In Joe's reference frame, he can measure the passage of time by measuring the number of round trips that an individual light pulse makes between the two mirrors. For one round trip, Joe would measure a time interval

$$\Delta t'_{\text{Joe}} = \frac{2d}{c} \quad (1.4)$$

So far so good. Since Joe is not moving relative to the mirrors, nothing unusual happens – assuming he has superhuman vision, he just sees the light pulses bouncing back and forth between the mirrors, Fig. 1.9a, straight up and down, and counts the number of round trips. Moe monitors this situation from the ground, in his own reference frame O . Thankfully, the boxcar is transparent, and Moe is able to see the light pulses and mirrors as well as the boxcar, which is moving at a velocity v from his point of view.

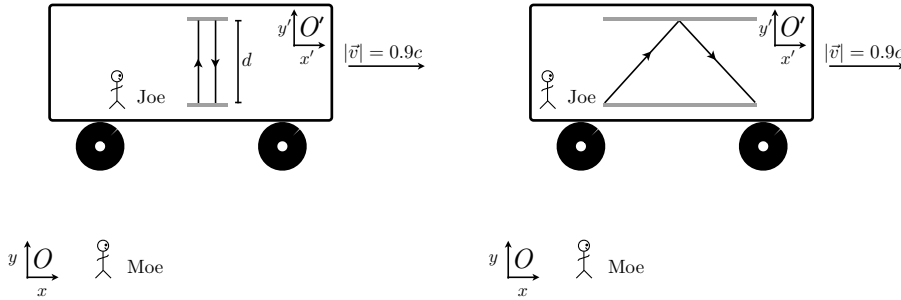


Fig. 1.9 left: Joe is traveling in a (transparent) boxcar, and he bounces laser beams between two mirrors inside the boxcar. Since the distance between the mirrors is known, and the speed of light is constant, he can measure time in this way. Joe measures the round trip time it takes the light to bounce from the bottom mirror, to the top, and back again. **right:** Moe observes the mirrors from the ground. From his frame O , the boxcar and mirrors are moving *but the light is not*. He therefore sees the light bouncing off of the mirrors at an angle. Using geometry and the constant speed of light, Moe also measures a round trip time interval, but since the path he observes for the light is different, *he measures a different time interval than Joe*.

What does Moe see? From his point of view, a light pulse is created at the bottom mirror while the whole assembly moves in the x direction – mirrors, light pulse, and all! Just like in the example of the light being flicked on in a space ship, the boxcar and mirrors have moved, but the point at which the light was created has not – Moe appears to see the light traveling at an angle. A light pulse is created at the bottom mirror, and it travels upward horizontally to reach the top mirror some time later, a bit further along the x axis. Rather than seeing the pulses going straight up and down, from Moe's point of view, they zig-zag sideways along the x axis, as shown in Fig. 1.9b.

So what? We know the speed of light is a constant, so both Joe and Moe must see the light pulses moving at a velocity c , even though they appear to be moving in along a different trajectory. If Moe also uses the light pulses' round trips to measure the passage of time, what time interval does he measure? The speed of light is constant, but the apparent distance covered by the light pulses is larger in Moe's case. Not only has the light traveled in the y direction a distance $2d$, over the course of one round trip it has also moved horizontally due to the motion of the boxcar. If the light has

apparently traveled farther from Moe's point of view, and the speed of light is constant, then the apparent passage of time from Moe's point of view must be greater!

Just how long does Moe observe the pulse round trip to be? Let us examine one half of a round trip, the passage of the light from the bottom mirror to the top. In that interval, from either reference frame, the light travels a vertical distance of d . From Joe's reference frame O' , the light does not travel horizontally, so the entire distance covered is just d , and he measures the time interval $\frac{1}{2}\Delta t'_{\text{Joe}} = d/c$. From Moe's reference frame, however, the car has also travelled horizontally at a velocity v in his time interval $\frac{1}{2}\Delta t_{\text{Moe}}$. From Moe's point of view, the car has moved forward by $\frac{1}{2}v\Delta t_{\text{Moe}}$. Thus, Moe would see the light cover a horizontal distance of $\frac{1}{2}v\Delta t_{\text{Moe}}$ and a vertical distance d , as shown in Fig. 1.10.

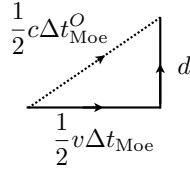


Fig. 1.10 Velocity addition for light pulses leading to time dilation. Within the boxcar (frame O' , Joe observes the light pulses traveling purely vertically, covering a distance d . On the ground (frame O), Moe sees the light cover the same vertical distance, but also sees them move horizontally due the motion of the boxcar at velocity v in his reference frame. The total distance the light pulse travels, according to Moe, is then the Pythagorean sum of the horizontal and vertical distances.

According to Moe, total distance that the light pulse covers in one half of a round trip is then the Pythagorean sum of the horizontal and vertical distances.

$$(\text{distance observed by Moe})^2 = d^2 + \left(\frac{1}{2}v\Delta t_{\text{Moe}}\right)^2 \quad (1.5)$$

According to special relativity, he must also observe the speed of light to be c just as Joe does. Moe would say that after one half round trip, the light has covered the above distance at a speed c , and would equate this with a time interval in his own reference frame Δt_{Moe} . Put another way, he would say that the distance covered by the light in one half round trip is just $\frac{1}{2}c\Delta t_{\text{Moe}}$. If we also note from Joe's observations that $d = \frac{1}{2}c\Delta t'_{\text{Joe}}$:

$$\left(\frac{1}{2}c\Delta t_{\text{Moe}}\right)^2 = d^2 + \left(\frac{1}{2}v\Delta t_{\text{Moe}}\right)^2 = \left(\frac{1}{2}c\Delta t'_{\text{Joe}}\right)^2 + \left(\frac{1}{2}v\Delta t_{\text{Moe}}\right)^2 \quad (1.6)$$

Now we see that if the speed of light is indeed constant, *there is no way that the time intervals measured by Joe and Moe can be the same!* The pulse seems to take longer to make the trip from Moe's perspective, since it also has to travel sideways, not just up and down. Solely due to the constant and invariant speed of light, Joe and Moe must measure different time intervals, and Moe's must be the longer of the two. We can solve the equation above to find out just what time interval Moe measures:

$$\left(\frac{1}{2}c\Delta t_{\text{Moe}}\right)^2 = \left(\frac{1}{2}c\Delta t'_{\text{Joe}}\right)^2 + \left(\frac{1}{2}v\Delta t_{\text{Moe}}\right)^2 \quad (1.7)$$

$$c^2 (\Delta t_{\text{Moe}})^2 = c^2 (\Delta t'_{\text{Joe}})^2 + v^2 (\Delta t_{\text{Moe}})^2 \quad (1.8)$$

$$(\Delta t_{\text{Moe}})^2 (c^2 - v^2) = c^2 (\Delta t'_{\text{Joe}})^2 \quad (1.9)$$

$$(\Delta t_{\text{Moe}})^2 = \frac{c^2}{c^2 - v^2} (\Delta t'_{\text{Joe}})^2 = \frac{1}{1 - \frac{v^2}{c^2}} (\Delta t'_{\text{Joe}})^2 \quad (1.10)$$

$$\Delta t_{\text{Moe}} = \Delta t'_{\text{Joe}} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \equiv \Delta t'_{\text{Joe}} \gamma \quad (1.11)$$

Here we defined a dimensionless quantity $\gamma = 1/\sqrt{1 - v^2/c^2}$ to simplify things a bit, we'll return to that shortly. So long as $v < c$, the time interval that Moe measures is always *larger* than the one Joe measures, by an amount which increases as the boxcar's velocity increases. This is a general result in fact: **the time interval measured by an observer in motion is always longer than that measured by a stationary observer**. Typically, we say that the moving observer measures a *dilated* time interval, hence this phenomena is often referred to as *time dilation*. The time dilation phenomena is symmetric – if Moe also had a clock on the ground, Joe would say that Moe's clock runs slow by precisely the same amount. It is only the relative motion that matters.

Time dilation

Two events take place at the same location. The time interval Δt between the events as measured by an observer moving with respect to the events is always *larger* than that measured by an observer who is stationary with respect to the events. The 'proper' time Δt_p is that measured by the stationary observer.

$$\Delta t'_{\text{moving}} = \gamma \Delta t_{\text{stationary}} = \gamma \Delta t_p \quad \text{where} \quad \gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (1.12)$$

In other words, time is stretched out for a moving observer compared to one at rest.

In the example above, it is Moe who is in a reference frame moving relative to our light 'clock' and Joe is the stationary observer. Therefore, Joe measures the 'proper' time interval, while Moe measures the dilated time interval. Incidentally, for discussions involving relativity, we basically assume that there is always a clock sitting at every possible point in space, constantly measuring time intervals, even though this is clearly absurd. What we really mean is the elapsed time that a clock at a certain position *would* read, if we had one there. For the purpose of illustration, it is just simpler to presume that everyone carries a fantastically accurate clock at all times.

Caveat for time dilation

The analysis above used to derive the time dilation formula relies on both observers measuring the same events taking place at the same physical location at the same time, such as two observers measuring the same light pulses. When timing between spatially separated events or dealing with questions of simultaneity, we must follow the formulas developed in Sect. 1.3.4.

The quantity dimensionless quantity γ is the ratio of the time intervals measured by the observers moving (Moe) and stationary (Joe) relative to the events being timed. This quantity, defined by Eq. 1.13, comes up often in relativity, and it is called the *Lorentz factor*. Since c is the absolute upper limit for the velocity of anything, γ is always greater than 1. So long as the relative velocity of the moving observer is fairly small relative to c , the correction factor is negligible, and we need not worry about relativity (e.g., at a velocity of $0.2c$, the correction is still only about 2%). In some sense, the quantity γ is sort of a gauge for the importance of relativistic effects – if $\gamma \approx 1$, relativity can usually be neglected, while if γ is much above 1, we must include relativistic effects like time dilation. Figure 1.11 provides a plot and table of γ versus v/c for reference. Note that as v approaches c , γ increases extremely rapidly.

Lorentz factor γ :

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \geq 1 \quad (1.13)$$

γ is dimensionless, and $\gamma \geq 1$ for $v \leq c$. γ approaches 1 for low velocities, and increases rapidly as v approaches c . If $\gamma \approx 1$, one can usually neglect the effects of relativity.

In the case above, for velocities much less than c , when $\gamma \approx 1$, Eq. 1.12 tells us that both Joe and Moe measure approximately the same time interval, just as our everyday intuition tells us. In fact, for most velocities you might encounter in your everyday life, the correction factor γ is only different from 1 by a miniscule amount, and the effects of time dilation are negligible. They are not, however, *unmeasurable* or *unimportant*, as we will demonstrate in subsequent sections – time dilation has been experimentally verified to an extraordinarily high degree of precision, and does have some everyday consequences.

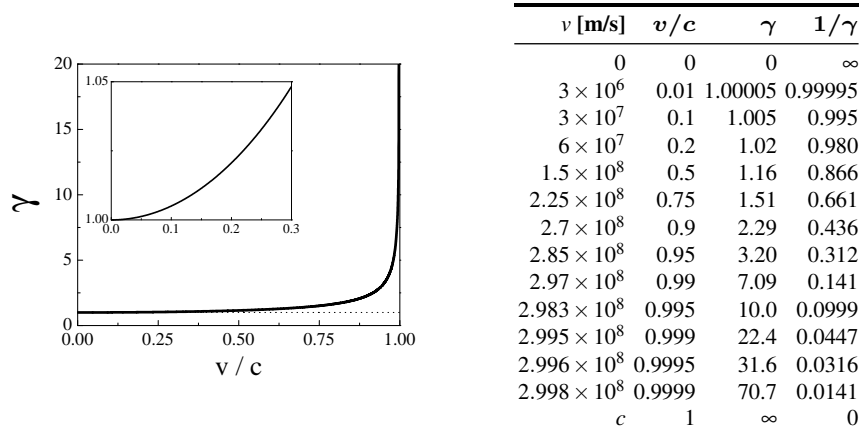


Fig. 1.11 The Lorentz factor γ and its inverse for various velocities in table and graph form. The inset to the graph shows an expanded view for low velocities.

1.3.2.1 Example: The Global Positioning System (GPS)

Before we discuss the stranger implications of time dilation, it is worth discussing at least one practical example in which the consequences of time dilation are important: the global positioning system. As you probably know, the Global Positioning System (GPS) is a network of satellites in medium earth orbit that transmit extremely precise microwave signals that can be used by a receiver to determine location, velocity, and timing. Each GPS satellite repeatedly transmits a message containing the current time, as measured by an onboard atomic clock, as well as other parameters necessary to calculate its exact position. Since the microwave signals from the satellites travel at the speed of light (microwaves are just a form of light, Sect. 8.5), knowing time difference between the moment the message was sent and the moment it was received allows an observer to determine their distance from the satellite. A ground-based receiver collects the signals from at least four distinct GPS satellites and uses them to determine its four space and time coordinates - $(x, y, z$ and $t)$.

How does relativity come into play? The 31 GPS satellites currently in orbit are in a medium earth orbit at an altitude of approximately 20,200km, which give them a velocity relative to the earth's surface of 3870 m/s.^{[2]³} This means that the actual atomic clocks responsible for GPS timing on the satellites are moving at nearly 4000 m/s relative to the receivers on the ground calculating position. Therefore, based on our discussion above, we would expect that the satellite-based GPS clocks would measure longer time intervals than those on the earth – the GPS clocks should run slow, a problem for a system whose entire principle is based on precise timing.

How big is this effect? We already know enough to calculate the timing difference. Let us assume that (somehow) at $t = 0$ we manage to synchronize a GPS clock with a ground-based one. From that moment, we will measure the elapsed time as measured by both clocks until the earth-based

³ You may remember from studying gravitation the orbital speed can be found from Newton's general law of gravitation and centripetal force, $v = \sqrt{GM/r}$, where G is the universal gravitational constant, M is the mass of the earth, and r is the radius of the orbit, as measured from the earth's center.

clock reads exactly 24 hours. We will call the earth-based clock's reference frame O , and that on the GPS satellite O' , and label the time intervals correspondingly. Since we are on the ground in the earth's reference frame, obviously we consider the earth-based clock to be the stationary one, measuring the proper time, and the GPS clock is moving relative to us. Applying Eq. 1.12, the elapsed time measured by the GPS clock and an earth-bound clock are related by a factor γ :

$$\Delta t'_{\text{GPS}} = \gamma \Delta t_{\text{Earth}} \quad (1.14)$$

The difference between the two clocks is then straightforward to calculate, given the relative velocity of the satellite of $v = 3870 \text{ m/s} \approx 1.3 \times 10^{-5} c$:

$$\text{time difference} = \Delta t_{\text{Earth}} - \Delta t_{\text{GPS}} = \Delta t_{\text{Earth}} - \gamma \Delta t_{\text{Earth}} \quad (1.15)$$

$$= \Delta t_{\text{Earth}} (1 - \gamma) = \Delta t_{\text{Earth}} \left[1 - \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \right] \quad (1.16)$$

$$= \left[24 \frac{\text{h}}{\text{day}} \cdot 60 \frac{\text{min}}{\text{h}} \cdot 60 \frac{\text{s}}{\text{min}} \right] \left[1 - \frac{1}{\sqrt{1 - (1.3 \times 10^{-5})^2}} \right] \quad (1.17)$$

$$\approx [86400 \text{ s/day}] [-8.32 \times 10^{-11}] \approx -7.2 \times 10^{-6} \text{ s/day} = -7.2 \mu\text{s/day} \quad (1.18)$$

A grand total of about $7 \mu\text{s}$ slow over an entire day (about $0.3 \mu\text{s}$ per hour), only about 80 parts per trillion (8×10^{-11}) per day! This may not seem like a lot, until one again considers that the GPS signals are traveling at the speed of light, and even a small error in timing can translate into a relatively large error in position. Remember, it is the travel time of light signals that determines distance in GPS. If time dilation were not accounted for, a receiver using that signal to determine distance would have an error given by the time difference multiplied by the speed of light. If we presume that, conservatively, position measurements are taken only once per hour:

$$\text{position difference in one hour} = \text{time difference in one hour} \times c \quad (1.19)$$

$$= [-3.0 \times 10^{-7} \text{ s/h}] [3.0 \times 10^8 \text{ m/s}] \quad (1.20)$$

$$\approx 90 \text{ m/h} \quad (1.21)$$

In the end, GPS must be far more accurate than this, and the effects of special relativity and time dilation must be accounted for, along with those of general relativity[2] (accounting for acceleration). Both effects together amount to a discrepancy of about $+38 \mu\text{s}$ per day. Since the orbital velocity of the satellites is well-known and essentially constant, the solution is simple: the frequency standards for the atomic clocks on the satellites are precisely adjusted to run slower and make up the difference. Though time dilation seems a rather ridiculous notion at first, it has real-world consequences we are familiar with, if unknowingly so.

Time dilation on a 747

The cruising speed of a 747 is about 250 m/s . After a 5 hour flight at cruising speed, by how much would your clock differ from a ground-based clock? How about after a year?

Using the same analysis as above, your clock would differ by about $6 \times 10^{-9} \text{ s}$ (6 ns) after five hours, and still only $10 \mu\text{s}$ after one year. Definitely not enough to notice, but enough to measure - current atomic clocks are accurate to $\sim 10^{-10} \text{ s/day}$ ($\sim 0.1 \text{ ns/day}$). In fact, in 1971 physicists performed precisely this sort of experiment to test the predictions of time dilation in relativity, and found excellent agreement.[3]

1.3.2.2 Example: The Twin ‘Paradox’

Now that we have a realistic calculation under our belt, let us consider a more extreme example. We will take identical twins, Joe and Moe, and send Moe on a rocket into deep space while Joe stays home. At the start of Moe’s trip, both are 25 years old. Moe boards his rocket, and travels at $v=0.95c$ to a distant star, and back again at the same speed. According to Joe’s clock on earth, this trip takes 40 years, and Joe is 65 years old when Moe returns. Moe, on the other hand, has experienced time dilation, since relative to the earth’s reference frame and Joe’s clock he has been moving at $0.95c$. Moe’s clock, therefore, runs more slowly, registers a smaller delay:

$$\Delta t_{\text{Joe}} = 40 \text{ yr} \quad (1.22)$$

$$\Delta t'_{\text{Moe}} = \gamma \cdot 40 \text{ yr} = \frac{40 \text{ yr}}{\sqrt{1 - \left(\frac{0.95c}{c}\right)^2}} \approx 12.5 \text{ yr} \quad (1.23)$$

It would seem, then that while Joe is 65 years old when Moe returns, having aged 40 years, Moe is 37.5 years old, having aged only 12.5 years! On the other hand, one of the principles of relativity is that there is no preferred frame of reference, it should be equally valid to use the clock on Moe’s rocket ship as the proper time. From Moe’s point of view, the earth is moving away from him at $0.95c$. In his reference frame, Joe’s earth-bound clock should run slow, and *Moe* should be older than Joe!

This is the so-called Twin ‘Paradox’ of special relativity. In fact, it is not a paradox, but a misapplication of the notion of time dilation. The principles of special relativity we have been discussing are only valid for *non-accelerating* reference frames. In order for Moe to move from the earth’s reference frame to the moving reference frame of the rocket ship at $0.95c$ and back again, he had to have accelerated during the initial and final portions of the trip, plus at the very least to turn around. The reference frame of the earth is for all intents and purposes not accelerating, but the reference frame on the ship *is*, and our calculation of the time dilation factor is not complete.

While the earth-bound clocks to run slow from the ship’s point of view *so long as the velocity of the spaceship is constant*, during the accelerated portions of the trip the earth-bound clocks actually run *fast* and gain time compared to the rocket’s clocks. An analysis including accelerated motion is beyond the scope of this text, but the gains of the earth-bound clock during the accelerated portion of the trip more than make up for the losses during the constant velocity portion of the trip, and no matter *who* keeps track, Joe will actually be younger than Moe from any reference frame. In short, there is no ‘paradox’ so long as the notions of relativity are applied carefully within their limits.

Inertial reference frames:

The principles of special relativity we have been discussing are only valid in *inertial* or non-accelerating reference frames. When accelerated motion occurs, a more complex analysis must be used.

1.3.3 Length Contraction

If the passage of time itself is altered by relative motion, what else must also be different? If the elapsed time interval depends on the relative motion of the clock and observer, then at constant velocity one would also begin to suspect that distance measurements must also be affected. After all, so far we have mostly talked about time in terms of objects or pulses of light traversing specific distances at constant velocity. Naturally, in order to explore this idea we need another thought experiment. Once again, it needs to involve a spaceship.

This time, the experiment is simple: a spaceship departs from earth toward a distant star, Fig. 1.12. In accordance with our discussion above, we stipulate that we *only consider the portion of the*

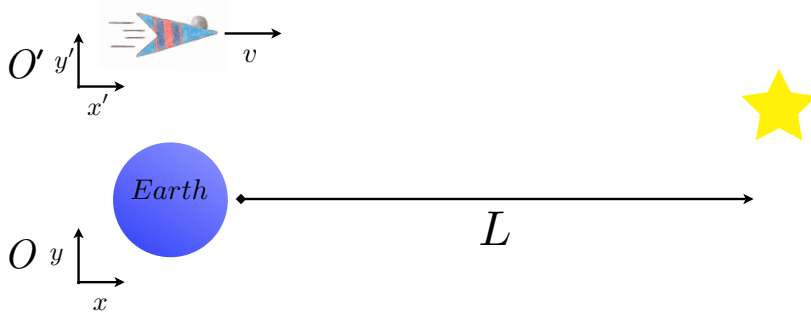


Fig. 1.12 Length contraction and travel to a distant star. A spaceship (frame O') sets out from earth (frame O) at a velocity v toward a distant star. Do the observers in the spaceship and the earth-bound observers agree on the distance to the star?

ship's journey where it is traveling at constant velocity, and there is no acceleration to worry about. According to observations on the earth, the star is a distance L away, and the spaceship is traveling at a velocity v . From the earth's reference frame O , the amount of time the trip should take Δt_E is easy to calculate:

$$\Delta t_E = \frac{L}{v} \quad (1.24)$$

Fair enough. On the spaceship, however, the passage of time is slowed by a factor γ due to time dilation, and from their point of view, the trip takes less time. Since our spaceship is not accelerating in this example (it doesn't even have to turn around), we can readily apply Eq. 1.12. From the spaceship occupant's point of view, the earth is moving relative to them, so the time interval should be *divided* by γ to reflect their shorter elapsed time interval.

$$\Delta t'_{\text{ship}} = \frac{\Delta t_E}{\gamma} \quad (1.25)$$

Keep in mind, by *clock*, we mean the passage of time itself, this includes biological processes. We already know what Δt_E must be from Eq. 1.24, so we can plug that in to Eq. 1.25 above:

$$\Delta t'_{\text{ship}} = \frac{L}{v\gamma} \quad (1.26)$$

Do I divide or multiply by γ ?

The Lorentz factor γ is always greater or equal to 1, $\gamma \geq 1$. If you are unsure about whether to divide or multiply by γ , think qualitatively about which quantity should be larger or smaller. In the example above, Eq. 1.25, we know the spaceship's time interval should be smaller than that measured on earth, so we know we have to *divide* the earth's time interval by γ .

If the occupants of the ship also measure their velocity relative to the earth (we will pretend they even communicate with earth to make sure all observers agree on the relative velocity, $v' = v$), then they will presume that upon arrival at the distant star, the distance covered must be their velocity times their measured time interval. From the ship occupant's point of view, then, the distance to the star measured in their reference frame, L' is

$$L' = v\Delta t'_{\text{ship}} = \frac{v\Delta t_E}{\gamma} = \frac{L}{\gamma} \neq L \quad (1.27)$$

If you ask the people on the ship, the distance to the star is shorter, because their apparent time interval is! As we might have guessed, the relativity of time measurement also manifests itself in measurements of length, a phenomena known as *length contraction* or *Lorentz contraction*.

Length Contraction The length of an object (or the distance to an object) as measured by an observer in motion is *shorter* than that measured by an observer at rest by a factor $1/\gamma$. The proper length, L_p , is measured at rest with respect to the object.

$$L'_{\text{moving}} = \frac{L_{\text{stationary}}}{\gamma} = \frac{L_p}{\gamma} \quad (1.28)$$

That is, objects and distances appear shorter by $1/\gamma$ if you are moving relative to them.

This analysis isn't just for distances, but any spatial dimension in the direction of motion. The length of an object is measured to be shorter when it is moving relative to the observer than when it is at rest - objects and distances appear shorter if you are moving relative to them. For example, a baseball moving past you at very high velocity would be shortened only along one axis parallel to the direction of motion, and would appear as an ellipsoid, not as a smaller sphere. It would be "squashed" along the direction of the baseball's motion only, as shown in Fig. 1.13. The length contraction appears *only along the direction in which there is relative motion*. In this case, that means the sphere looks contracted only in one direction, so it is squashed instead of just smaller.

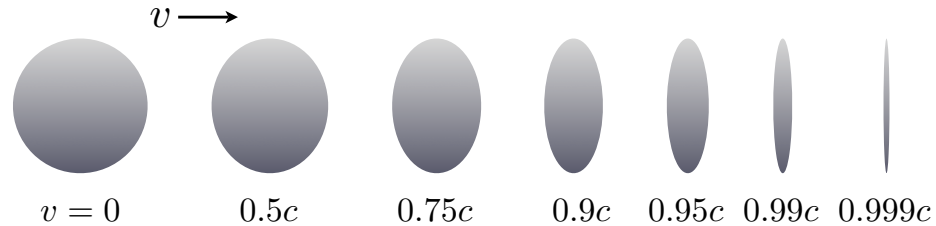


Fig. 1.13 Length contraction of a sphere traveling at various speeds, viewed side-on. The length contraction occurs only along the direction of motion. Hence, to a stationary observer, the moving sphere appears 'flattened' along the direction of motion into an ellipsoid.

Just like time dilation, the length contraction effect is negligibly small at everyday velocities: if v is small compared to c , then $\gamma \approx 1$ and the two lengths are essentially equal. Unlike time dilation, there is as yet no everyday application of time dilation, and no simple and straightforward experimental proof. We have no practical way of measuring the length of an object at extremely high velocities with sufficient precision at the moment. Collisions of elementary particles at very high velocities in particle accelerators provides some strong but indirect evidence for length contraction, and in some sense, since length contraction follows directly from time dilation, the experimental verifications of time dilation all but verify length contraction.

A summary of sorts:

1. objects and distances in relative motion appear shorter by $1/\gamma$
2. the length contraction is only along the direction of motion
3. the objects do not actually get shorter in their own reference frame, it is only apparent to the moving observer

1.3.4 Time and position in different reference frames

Now that we have a good grasp of time dilation and length contraction, we can start to answer the more general question of how we translate between time and position of events seen by observers in different reference frames. For example, consider Fig. 1.14. A girl in frame O is stationary relative to a star at point P , known to be a distance x away, which suddenly undergoes a supernova explosion. At precisely this instant, a boy travels past her in on a skateboard at constant relative velocity v along the x axis (frame O'). For convenience, we will assume that at the moment the explosion occurs, he is exactly the same distance away as the girl. When and where does the supernova occur, according to their own observations? How can we relate the distances and times measured by the girl to that measured by the boy, and *vice versa*?

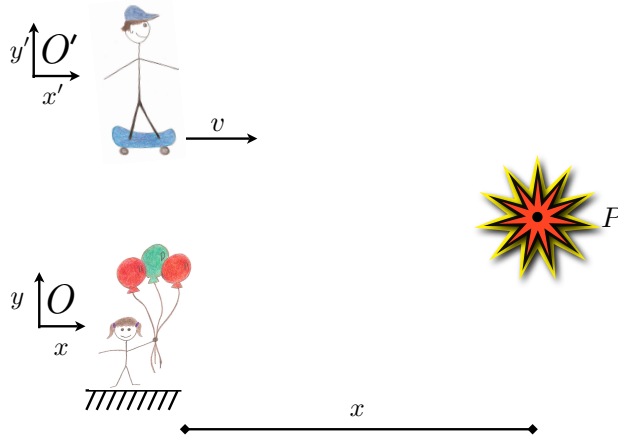


Fig. 1.14 A stationary and moving observer watch a supernova explosion. A girl in frame O is stationary relative to the supernova, a distance x away. A boy on a skateboard in the O' frame is traveling at v relative to frame O . How long does it take before the first light of the supernova reaches each of them?

All we need to do is apply what we know of relativity thusfar, and compare what each observer would measure in their own frame with what the *other* would measure. In the girl's case, the situation is fairly straightforward. She is a distance x from the star, and the first light from the explosion travels that distance at a velocity c . Therefore, according to her observations, the first light from the supernova arrives after:

$$t_{\text{arrival}} = \frac{x}{c} \quad (1.29)$$

What about the boy on the skateboard, in frame O' ? Since he is moving relative to the star, the distance to the star appears length contracted from his point of view. At the instant of the supernova, he measures a distance shorter by a factor γ compared to that measured by the girl. Furthermore, from his point of view in his own reference frame, he is sitting still, and *the supernova is moving toward him* at velocity v . Therefore, from his point of view the supernova is getting closer to him. After t' seconds by his clock, the supernova is a distance vt' closer. Putting these two bits together, the distance x' the boy would measure to the supernova is:

$$x' = \frac{x}{\gamma} - vt' \quad (1.30)$$

So the distance to the supernova he claims is the original distance, length contracted due to his motion relative to the supernova, minus the rate at which he gets closer to the supernova.

What would the girl say about all this? The distance between the boy and the supernova, from her point of view, would have to be contracted to x'/γ since the boy is in motion relative to her. Additionally, from her point of view, since the boy is moving away from her at v , the distance between the two is *increasing* by vt after t seconds. We can express her perceived distance to the supernova as the sum of two distances: the distance from her to the boy, and the distance from the boy to the supernova:

$$x = vt + \frac{x'}{\gamma} \quad (1.31)$$

Now we have consistent expressions relating the distance measured by one observer to that measured by the other. If we rearrange Eqs. 1.30 and 1.31 a bit, and put primed quantities on one side and unprimed on the other, we arrive the transformations between positions measured by moving observers in their usual form:

Transformation of distance between reference frames:

$$x' = \gamma(x - vt) \quad (1.32)$$

$$x = \gamma(x' + vt') \quad (1.33)$$

Here (x, t) is the position and time of an event as measured by an observer in O stationary to it. A second observer in O' , moving at velocity v , measures the same event to be at position and time (x', t') .

Again, at small velocities compared to c , $\gamma \approx 1$, and we recover a familiar result from mechanics: $x' \approx x - vt$. These equations include the effects of length contraction and time dilation we have already developed, as well as including the relative motion between the observers. If we use Eqs. 1.30 and 1.31 together, we can also arrive at a more direct expression to transform the measurement times as well. To start, we'll take Eq. 1.32 as written, and substitute it into Eq. 1.33:

$$x = \gamma(x' + vt') \quad (1.34)$$

$$= \gamma(\gamma(x - vt) + vt') \quad (1.35)$$

$$= \gamma^2 x - \gamma^2 vt + \gamma vt' \quad (1.36)$$

So far its a bit messy, but it will get better. Now let's solve that for t' . A handy relationship we will make use of is $(1 - \gamma^2)/\gamma^2 = -v^2/c^2$, which you should verify for yourself.

$$\gamma vt' = (1 - \gamma^2)x + \gamma^2 vt \quad (1.37)$$

$$\Rightarrow t' = \gamma t + \frac{(1 - \gamma^2)x}{\gamma v} = \gamma \left[t + \frac{1 - \gamma^2}{\gamma^2} \left(\frac{x}{v} \right) \right] = \gamma \left[t - \frac{vx}{c^2} \right] \quad (1.38)$$

And there we have it, the transformation between time measured in different reference frames. A similar procedure gives us the reverse transformation for t in terms of x' and t' .

Time measurements in different non-accelerating reference frames:

$$t' = \gamma \left(t - \frac{vx}{c^2} \right) \quad (1.39)$$

$$t = \gamma \left(t' + \frac{vx'}{c^2} \right) \quad (1.40)$$

Here (x, t) is the position and time of an event as measured by an observer in O stationary to it. A second observer in O' , moving at velocity v , measures the same event to be at position and time (x', t') .

The first term in this equation is just the time it takes light to travel across the distance x from point P , corrected for the effects of time dilation we now expect. The second term is new, and it

represents an additional *offset* between the clock on the ground and the one in the car, not just one running slower than the other. What it means is that events seen by the girl in frame O do *not* happen at the same time as viewed by the boy in O' !

This is perhaps more clear to see if we make two different measurements, and try to find the elapsed time between two events. If our girl in frame O sees one event take place at position x_1 and time t_1 , labeled as (x_1, t_1) , and a second event at x_2 and t_2 , labeled as (x_2, t_2) , then she would say that the two events were spatially separated by $\Delta x = x_2 - x_1$, and the time interval between them was $\Delta t = t_2 - t_1$. If we follow the transformation to find the corresponding times that the boy observes, t'_1 and t'_2 , we can also calculate the boy's perceived time interval between the events, $\Delta t'$:

Elapsed times between events in non-accelerating reference frames:

$$\Delta t' = t'_2 - t'_1 = \gamma \left(\Delta t - \frac{v \Delta x}{c^2} \right) \quad (1.41)$$

If observer in O stationary relative to the events (x_1, t_1) and (x_2, t_2) measures a time difference between them of $\Delta t = t_2 - t_1$ and a spatial separation $\Delta x = x_2 - x_1$, an observer in O' measures a time interval for the same events $\Delta t'$. Events simultaneous in one frame ($\Delta t = 0$) are only simultaneous in the other ($\Delta t' = 0$) when there is no spatial separation between the two events ($\Delta x = 0$).

For two events to be simultaneous, there has to be no time delay between them. For the girl to say the events are simultaneous requires that she measure $\Delta t = 0$, while for the boy to say the same requires $\Delta t' = 0$. We cannot satisfy both of these conditions based on Eq. 1.41 unless there is no relative velocity between observers ($v = 0$), or the events being measured are not spatially separated ($\Delta x = 0$). This means *two observers in relative motion will only find the same events simultaneous if the events are not spatially separated! Put simply, events are only simultaneous in both reference frames if they happen at the same spot.* At a given velocity, the larger the separation between the two events, the greater the degree of non-simultaneity. Similarly, for a given separation, the larger the velocity, the greater the discrepancy between the two frames. This is sometimes called “failure of simultaneity at a distance.” Not to drill the point home too much, but once again for velocities small compared to c we recover our normal Newtonian result. If $v \ll c$, $\gamma \approx 1$, and we can also neglect the second term in parentheses in Eq. 1.41, leaving us with $\Delta t' \approx \Delta t$.

In the end, this is our *general* formula for time dilation, including events which are spatially separated. If we plough still deeper into the consequences of special relativity and simultaneity, we will find that our principles of relativity have indeed preserved causality - cause always precedes effect - it is just that what one means by “precede” depends on which observer you ask. What relativity says is that cause must precede its effect according to all observers in inertial frames, which equivalently prevents both faster than light travel or communication and influencing the past.

1.3.4.1 Summary of sorts: the Lorentz Transformations

We are now ready to make a summary of the relativistic transformations of time and space. Let us consider two reference frames, O and O' , moving at a **constant** velocity v relative to one another. For simplicity, we will consider the motion to be along the x and x' axes in each reference frame, so the problem is still one-dimensional. The observer in frame O measures an event to occur at time t and position (x, y, z) . The event is *at rest with respect to the O frame*. Meanwhile, the observer in frame O' measures the *same event* to take place at time t' and position (x', y', z') . Based on what we have learned so far, we can write down the general relations between space and time coordinates in each frame, known as the *Lorentz transformations*:

Lorentz transformations between coordinate systems:

$$x' = \gamma(x - vt) \quad \text{or} \quad x = \gamma(x' + vt') \quad (1.42)$$

$$y' = y \quad (1.43)$$

$$z' = z \quad (1.44)$$

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) \quad \text{or} \quad t = \gamma\left(t' + \frac{vx'}{c^2}\right) \quad (1.45)$$

Here (x, y, z, t) is the position and time of an event as measured by an observer in O stationary to it. A second observer in O' , moving at velocity v along the x axis, measures the same event to be at position and time (x', y', z', t') .

Here we have provided both the ‘forward’ and ‘reverse’ forms of the transformations for convenience. Again, the distance is only contracted along the direction of motion, the x and x' directions – the y and z coordinates are thus unaffected. When the velocity is small compared to c ($v \ll c$), the first equation gives us our normal Newtonian result, the position in one frame relative to the other is just offset by their relative velocity times the time interval, and the time is the same. These compact equations encompass all we know of relativity so far - length contraction, time dilation, and lack of simultaneity.

Relativity for observers in relative motion at constant velocity:

1. Moving observers see lengths contracted along the direction of motion.
2. Moving observers’ clocks ‘run slow’, less time passes for them.
3. Events simultaneous in one frame are not simultaneous in another unless they occur at the same position
4. All observers measure the same speed of light c

1.3.5 Addition of Velocities in Relativity

The invariance of the speed of light has another interesting consequence, namely, that one can no longer simply add velocities together to compute relative velocities in different reference frames in the way we did at the beginning of this chapter. Think about one of our original questions regarding relative motion, Fig. 1.2, in which a bully threw a dart off of a moving skateboard at a little girl’s balloon. In that case, we said that the girl observed the dart to move at a velocity which was the *sum* of the velocities of the skateboard relative to the girl and the dart relative to the skateboard. When velocities are an appreciable fraction of the speed of light, this simple velocity addition breaks down.

In the end, it *has to*, or the speed of light could not be an absolute cosmic speed limit. Think about this: if you are driving in your car at 60 mi/hr down the freeway and turn on your headlights, do the light beams travel at c , or c plus 60 mi/hr? We know already that the answer must be c , but that is not at all consistent with our usual ideas of relative motion. If we can’t just add the velocities together, what do we do? Is there a way to combine relative velocities such that the speed of light remains a constant and an upper limit? There is a relatively simple mathematical way to accomplish this. Once again, we will derive the result in the context of yet another thought experiment and try to show you how to use it.

The present thought experiment is just a variation the dart thrown from the skateboard, and is shown in Fig. 1.15. An observer on the ground (frame O) sees a person on a cart (frame O') moving at velocity v_a , as measured in the ground-based reference frame O . The person on the cart throws a ball at a velocity v'_b relative to the cart, which is measured as v_b in the ground-based frame.

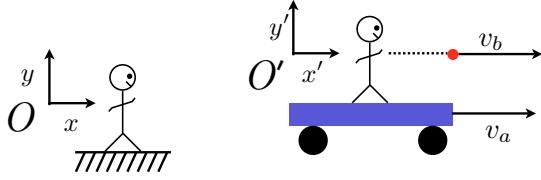


Fig. 1.15 Relativistic addition of velocities. An observer on the ground (frame O) sees a person on a cart (frame O') traveling at velocity v_a throw a ball off of the car at a velocity v_b relative to the ground. How do we relate the velocities as measured in the different reference frames?

The ground-based observer measures v_a and v_b , while the observer on the cart measures the cart's velocity as v'_a and the ball's velocity as v'_b . How do we relate the velocities measured in the different frames O and O' , without violating the principles of relativity we have investigated so far?

We can't simply add and subtract the velocities like we want to, our thought experiment of Sect. 1.2.3 involving a flashlight and a rocket ruled this out already, since this does not keep the speed of light invariant. So how *do* we properly add the velocities? Velocity is just displacement per unit time. If we calculate the displacement and time in one reference frame, then transform *both* to the other reference frame, we can divide them to *correctly* find velocity.

Let's start with the velocity of the ball as measured by the observer on the cart, v'_b . The displacement of the ball relative to the cart at some time t' after it was thrown, also measured in the cart's frame O' , is just $x'_b = v'_b t'$. This is just how far ahead of the car the ball is after some time t' . We can substitute this into Eq. 1.42 to find out what displacement the observer on the ground in O should measure, remembering that v_a is the relative velocity of the observers:

$$x_b = \gamma(x'_b + v_a t') = \gamma(v'_b t' + v_a t') \quad (1.46)$$

But now we have x , the displacement of the ball seen from O , in terms of t' , the time measured in O' . If we want to find the velocity of the ball as measured by an observer in O , we *have to divide the distance measured in O by the time measured in O* . We can't divide one person's position by another person's time, we have to transform *both*. So we should use Eq. 1.45 to find out what t is from t' too:

$$t = \gamma\left(t' + \frac{v_a x'}{c^2}\right) = \gamma\left(t' + \frac{v_a v'_b t'}{c^2}\right) \quad (1.47)$$

Now we have the displacement of the ball x and the time t as measured by the observer on the ground in O . The velocity in O is just the ratio of x to t :

$$v_b = \frac{x}{t} \quad (1.48)$$

$$= \frac{\gamma(v'_b t' + v_a t')}{\gamma\left(t' + \frac{v_a v'_b t'}{c^2}\right)} \quad (1.49)$$

$$= \frac{v'_b + v_a}{1 + \frac{v_a v'_b}{c^2}} \quad (1.50)$$

For the last step, we divided out $\gamma t'$ from everything, by the way. So, this is the proper way to compute relative velocity of the ball observed from the ground, consistent with our framework of relativity.

$$\text{velocity of ball observed from the ground} = v_b = \frac{v_a + v'_b}{1 + \frac{v_a v'_b}{c^2}} \quad (1.51)$$

In the limiting case that the velocities are very small compared to c , then it is easy to see that the expression above reduces to $v_b = v_a + v'_b$ – the velocity of the ball measured from the ground is the velocity of the car relative to the ground plus the velocity of the ball relative to the car. But, this is *only* true when the velocities are small compared to c .⁴ Similarly, we could solve this equation for v'_b instead and relate the velocity of the ball as measured from the car to the velocities measured from the ground:

$$\text{velocity of ball observed from the car} = v'_b = \frac{v_b - v_a}{1 - \frac{v_a v_b}{c^2}} \quad (1.52)$$

The equation above allows us to calculate the velocity of the ball as observed from the car if we only had ground-based measurements. Again, for low velocities, we recover the expected result $v'_b = v_b - v_a$. What about the velocity of the car? We don't need to transform it, since it is already the *relative* velocity between the frames O and O' , and hence between the ground-based observer and the car. **We only need the velocity addition formula when a third party is involved.** Out of the three relevant velocities, we only ever need to know two of them.

So this is it. This simple formula is all that is needed to properly add velocities and obey the principles of relativity we have put forward. Below, we put this in a slightly more general formula.

Relativistic velocity addition:

We have an observer in a frame O , and a second observer in another frame O' who are moving relative to each other at a velocity v . Both observers measure the velocity of another object in their own frames (v_{obj} and v'_{obj}). We can relate the velocities measured in the different frames as follows:

$$v_{\text{obj}} = \frac{v + v'_{\text{obj}}}{1 + \frac{v v'_{\text{obj}}}{c^2}} \quad v'_{\text{obj}} = \frac{v_{\text{obj}} - v}{1 - \frac{v v_{\text{obj}}}{c^2}} \quad (1.53)$$

Again, v_{obj} is the object's velocity as measured from the O reference frame, and v'_{obj} is its velocity as measured from the O' reference frame.

Velocities greater than c ?

The velocity addition formula shows that one cannot accelerate something past the speed of light. No matter what subluminal velocities you add together, the result is *always* less than c . Try it! Our relativistic equations for momentum and energy will further support this.

Remember, c isn't just the speed of light, it is a limiting speed for *everything*!

1.3.5.1 Example: throwing a ball out of a car

Just to be clear, let us make our previous example more concrete. Let's say we have Joe in reference frame O , sitting on the ground, while Moe is in a car (frame O') moving at $v_{\text{car}} = \frac{3}{4}c$. Moe throws a ball *very hard* out of the car window, such that he measures its velocity to be $v'_{\text{ball}} = \frac{1}{2}c$ in his reference frame. What would Joe say that the velocity of the ball is, relative to his reference frame on the ground?

⁴ Or, more precisely, when the *product* of the velocities is small compared to c^2 .

Basically, Joe wants to know the velocity of the ball relative to the ground, not relative to the car. What we need to do is relativistically combine the velocity of the car relative to the ground and the velocity of the ball relative to the car. Classically, we would just add them together:

$$v_{\text{ball}} = v_{\text{car}} + v'_{\text{ball}} = \frac{3}{4}c + \frac{1}{2}c = \frac{5}{4}c = 1.25c \quad \text{INCORRECT!}$$

Clearly this is an absurdity - the ball cannot be traveling faster than the speed of light in *anyone's* reference frame. We need to use the proper relativistic velocity addition formula, Eq. 1.53. We know the velocity of the ball relative to the car in frame O' , v'_{ball} and the velocity of the car relative to the ground in the O frame, v_{car} , so we just substitute and simplify:

$$v_{\text{ball}} = \frac{v_{\text{car}} + v'_{\text{ball}}}{1 + \frac{v_{\text{car}}v'_{\text{ball}}}{c^2}} \quad (1.54)$$

$$= \frac{\frac{3}{4}c + \frac{1}{2}c}{1 + \frac{(\frac{3}{4}c)(\frac{1}{2}c)}{c^2}} \quad (1.55)$$

$$= \frac{\frac{5}{4}c}{1 + \frac{3}{8}} = \frac{10}{11}c \approx 0.91c \quad (1.56)$$

So, in relativity, three quarters plus one half is only about 0.9! But this is the result we are looking for - no matter what velocities $v < c$ we add together, we always get an answer less than c . Put another way, no matter what reference frame we consider, the velocity of an object will *always* observed to be less than c . So our relativistic velocity addition works so far. But what about applying it to light, which is actually traveling right at c . Does everything still come out ok?

1.3.5.2 Example: shining a flashlight out of a rocket

What if, instead of throwing a ball out of the window, Moe uses a flashlight to send out a light pulse? In that case, we have to find that the velocity of light is c no matter which frame we use. Remember our problem in Sect. 1.2.3? We had Joe traveling on a rocket at $0.99c$, while Moe on the ground shines a flashlight parallel to Joe's path, shown again in Fig. 1.16. Our claim at the time was that both Moe and Joe should measure the same speed of light. Does our new velocity addition formula work for this case?

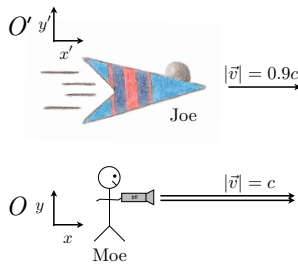


Fig. 1.16 Joe is traveling on a rocket at $|\vec{v}| = 0.99c$, while Moe on the ground shines a flashlight parallel to Joe's path. Both Joe and Moe observe the light from the flashlight to travel at $|\vec{v}| = c$, consistent with our relativistic velocity addition formula!

In this case, Joe is on a rocket (frame O') moving at $v_{\text{rocket}} = 0.99c$ relative to Moe on the ground. Moe knows that in his frame O , the light from the flashlight travels away from him at velocity $v_{\text{light}} = c$. What is the velocity of light observed by Joe in the rocket, v'_{light} , if we use the velocity addition formula? All we have to do is subtract the speed of light as measured by Moe from Joe's speed on the rocket ship, according to the second equation in 1.53:

$$v'_{\text{light}} = \frac{v_{\text{light}} - v_{\text{rocket}}}{1 - \frac{v_{\text{rocket}} v_{\text{light}}}{c^2}} \quad (1.57)$$

$$= \frac{c - 0.99c}{1 - \frac{(0.99c)(c)}{c^2}} \quad (1.58)$$

$$= \frac{0.01c}{1 - 0.99} = c \quad (1.59)$$

Lo and behold, the thing works! Our velocity addition formula correctly calculates that both Joe and Moe have to measure the same speed of light, since the speed of light is the same when observed from any reference frame. We shouldn't be too surprised, however: the velocity addition formula was *constructed* to behave in exactly this way. How about if Joe holds the flashlight while in the rocket, what is the speed of light as measured by Moe on the ground? Now we have to add the velocities of the light coming out of the rocket and the velocity of the rocket itself, according the first equation in 1.53. Still no problem:

$$v_{\text{light}} = \frac{v_{\text{rocket}} + v'_{\text{light}}}{1 + \frac{v_{\text{rocket}} v'_{\text{light}}}{c^2}} \quad (1.60)$$

$$= \frac{0.99c + c}{1 + \frac{(0.99c)(c)}{c^2}} \quad (1.61)$$

$$= \frac{1.99c}{1 + 0.99} = c \quad (1.62)$$

In the end, we have succeeded in constructing a framework of mechanics that keeps the speed of light invariant in all reference frames, and answers (nearly) all the questions raised at the beginning of the chapter.

Is everything relative then?

Not quite!

- All observers will agree on an objects *rest* length
- All observers will agree on the proper time
- All observers will agree on an objects rest mass
- The speed of light is an upper limit to physically attainable speeds

1.3.6 Space-time Intervals

What we have established thus far is a framework to describe physical events taking place. When we consider an event taking place, we must carefully consider both the position and time of both event and observer. The fact that light travels at a finite speed of c means that the influence of an event at one point in space can only be observed at another location after a delay which corresponds to the time it takes light to cover the distance separating event and observer. Essentially, this means we must treat spatial and temporal separations on equal footing, or consider both space and time to be linked as part of a larger structure we call *spacetime*, which is nothing more than a quantity that describes both the position coordinate and *time coordinate* of a particular event. Once again, we can proceed most simply by way of example.

Consider two observers in their own reference frames O and O' , which are in relative motion at constant velocity v . At a time $t = t' = 0$, the origins of O and O' coincide, and at exactly that moment, a light pulse is emitted from the common origin. Our question now is, how do the two observers

describe that light pulse as moving out from the origin? The observer in O would say that after time Δt the light pulse is at a position (x, y, z) and has covered a distance

$$r = \sqrt{x^2 + y^2 + z^2} = c\Delta t \quad (1.63)$$

The observer in O' , on the other hand, would say that the light pulse is at position (x', y', z') after time $\Delta t'$, having covered a distance

$$r' = \sqrt{x'^2 + y'^2 + z'^2} = c\Delta t' \quad (1.64)$$

This is nothing new, it is merely restating our conclusion that both elapsed time and distance covered are relative quantities. However, what we notice from the above is that *both observers would agree on the difference between distance covered and the time interval*. Specifically, we can construct what is called the *spacetime interval* s which combines the distance covered with the time interval, and results in a quantity that all observers can agree on:

$$s^2 = x'^2 + y'^2 + z'^2 - c^2\Delta t'^2 = x^2 + y^2 + z^2 - c^2\Delta t^2 = 0 \quad (1.65)$$

In this particular case, considering the motion of a light pulse, the spacetime interval is zero, because the spatial distance between the two events (the emission of the light pulse and its subsequent observation) is exactly balanced by the time between the two events. This is always true for the motion of light, the spacetime interval is always zero. Generally speaking, the spacetime interval describes both the spatial and temporal separation between events, and its definition is simple:

Spacetime intervals:

The interval s between two events is defined as

$$s^2 = \Delta r^2 - c^2\Delta t^2 = x^2 + y^2 + z^2 - c^2\Delta t^2 \quad (1.66)$$

Here c is the speed of light, Δt the differences in time coordinates between events, and Δr is the spatial separation between the two events. The spacetime interval is independent of any observer, i.e., all observers can agree.

In essence, the spacetime interval is the quantity that all observers can agree on, and replaces our everyday separate measurements of space and time intervals. In normal geometry, we would say that lengths are left invariant by rotations or translations – a meter stick is still a meter long if we twirl it around or move it across the room. We know this is not true in relativity, owing to length contraction and time dilation, but constructing the spacetime interval allows us to regain an analogous quantity, one which is conserved not just under translation and rotation, but between all reference frames. Even though two observers may not agree on the spatial separation between events or their time interval, but they can always agree on the spacetime interval. As with our discussions of simultaneity, the essential point is that a quantity that one observer can measure with only a meter stick must be measured with both meter sticks and clocks by another. A more subtle point to note is that in the expressions above time enters the equation in the same way as space does, save a sign change. This is yet another indication that in relativity time and space are to be treated on equal footing, and are of equal importance.

In order to get a feeling for what the spacetime interval s is and how it is useful, it is instructive to consider the three logical cases based on the definition of s : $s^2 < 0$, $s^2 > 0$, and $s^2 = 0$. All three cases have distinct physical meaning. For negative spacetime intervals between events, a frame exists in which the two events occur at different times, but the same place. Positive spacetime intervals mean that it is always possible to find a frame in which the events are simultaneous, but occur at different places. If the spacetime interval is exactly zero, the events can be connected perfectly by a pulse of light.

1.3.6.1 Time-like intervals, $s^2 < 0$ or $c^2\Delta t^2 > \Delta r^2$

The first case, $s^2 < 0$, is termed a “time-like” interval, because it corresponds to the situation where more time passes between events than required for light to traverse the distance between the events. Put more simply, if two events are separated by a time-like interval, enough time passes between the events that there could be a cause-effect relationship between the two. Of course, that does not mean there must be a causal relationship between the two, it only means that it is *possible*, since enough time has passed relative to the separation distance for at least a light pulse to have been present at both events. Two events which are separated by a time-like interval can be said to have occurred in each other’s past or future meaningfully. Conversely, the fact that $s^2 < 0$ there is *no reference frame* in which the two events can be said to have occurred at the same time. (We could, however, find a reference frame in which both events appear at the same location.)

Another way of thinking about time-like intervals is to consider the motion of material objects. As we will justify further below, any particle with mass must travel at speeds less than c , only massless particles (such as photons, light particles) can travel with a speed of c . If a particle (or other object) is traveling with $v < c$, any two events involving the particle must be separated by a time-like interval – if the particle traveling at $v < c$ has enough time to reach both events, then light traveling at $v = c$ has more than enough time, and thus s^2 must be negative. In other words, matter travels along time-like curves, since they always cover less distance (vt) than light would in the same time period (ct).

The term “time-like” comes from the fact that the separation between the events is not so far that causal relationships are impossible. In light of this, the most useful measure of a time-like spacetime interval is just the *proper time* interval, $\Delta\tau$:

$$\Delta\tau = \Delta t_p = \sqrt{\Delta t^2 - \frac{\Delta r^2}{c^2}} \quad \text{proper time} \quad (1.67)$$

Formally, the proper time interval would be measured by an observer traveling between the two events at constant velocity. The proper time interval is “proper time” just as we considered previously, it is the time between two events as measured by a single clock at the same place that the events take place.

Just like the spacetime interval, proper time intervals intrinsically involve both spatial and temporal separation, consistent with our earlier derivation of the failure of simultaneity. The farther away two events are, the longer the proper time interval between them, and for events taking place at the same location, proper time is just the time measured at the event location. Thus, the proper time interval is nothing new, it only takes on a slightly different form when we think in terms of spacetime intervals: we say that the proper time is the invariant spacetime interval along a time-like path.

1.3.6.2 Space-like intervals, $s^2 > 0$ or $c^2\Delta t^2 < \Delta r^2$

Time-like intervals, a negative squared spacetime interval, correspond to events being sufficiently close together that they could potentially have a cause-effect relationship. The physical meaning of the opposite case, $s^2 > 0$ or $c^2\Delta t^2 < \Delta r^2$, is then clear: it must correspond to situations where the two events are so far apart that, given the elapsed time between them, not even a light pulse could be present at both. We call the spacetime intervals in these situations *space-like intervals*. Two events separated by a space-like interval cannot have a causal relationship, since not even light can cross the distance between the two events in the time between them. Generally, this means we cannot speak of the two events as being in each other’s past or future in a meaningful way. However, the positive squared spacetime interval means that we *can* always find a reference frame in which the two events are observed at the same time (though we cannot find a reference frame in which the events appear at the same location).

When we have events separated by a time-like interval, we can generally speak of a *proper distance*. Just as before, the proper distance is the distance between two events, measured in an inertial reference frame in which the events are simultaneous. The proper length of a spacetime path is just the invariant spacetime interval along a space-like path,

$$\Delta\sigma = L_p = \sqrt{\Delta r^2 - c^2 \Delta t^2} \quad \text{proper distance} \quad (1.68)$$

Nothing, not even light, can travel on a space-like interval separating two events, because even light does not travel fast enough to be present at both events. One thing to note: if the interval between events is time-like, their ordering is absolute, and there is no reference frame that can change the perceived order of the two events. If the separation is space-like, we have relativity of simultaneity, and which event is perceived first depends on the system from which they are observed.

1.3.6.3 Light-like intervals, $s^2 = 0$ or $c^2 \Delta t^2 = \Delta r^2$

This is the case we originally considered, where the spatial separation of two events is exactly balanced by the time between the two events. This means that a light pulse emitted at one event makes it perfectly to the second event, for example. On the other hand, *only* light can follow light-like intervals, since traveling along these intervals with $s^2 = 0$ implies travel at the speed of light.

1.3.6.4 Visualizing spacetime: Minkowski diagrams

In mechanics, one of the first things we usually did to grasp a new situation was to draw a schematic figure of sometime. In the case of kinematics, this often meant sketching the path that a particle followed as a function of time and space, or plotting position versus time. For example, we might draw a parabola for the position of a thrown ball as a function of time. In this case, the vertical and horizontal coordinates give us the ball's position, the slope of the curve at a given point gives us the particle's velocity. Implicitly, most of our diagrams merged space and time coordinates of complicated motion into a single diagram.

Spacetime intervals can be more readily grasped in a similar way, though for (probably) historical reasons the analogous construction is somewhat awkward at first. Spacetime intervals are graphed with the position of a particle on the horizontal axis and time on the vertical axis (making it inherently a discussion of one-dimensional motion, typically). The time coordinate is measured out in units of ct , or in other words the distance light travels in a given unit of time, with the space coordinate in the same units of distance. This choice of units has the conceptual advantage that traveling at the speed of light is simply represented by a 45° line. Anything traveling with $v < c$ has a larger slope, less distance covered than light in the same unit of time. Given the “backward” axes we work with in these diagrams, known as *Minkowski diagrams*, the velocity of a particle is actually the reciprocal of the slope, not the slope, so lines of higher slope correspond to slower objects. The trajectory of a particle on one of these diagrams is called a *world line*.

Perhaps this is easier to grasp with an example. Below in Figure 1.17 we plot a Minkowski diagram which includes the motion of a photon (light particle), a rocket, and a particle at rest.

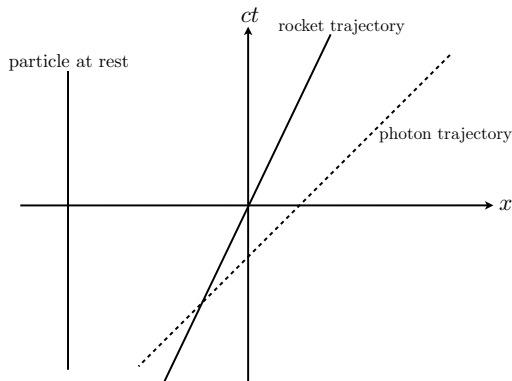


Fig. 1.17 A spacetime diagram showing the motion of a photon (a particle of light), a rocket, and a particle at rest. The particle at rest has constant position for all times, and is thus a vertical line. The photon travels at the speed of light, so $t = x/c$ or $ct = x$. The rocket travels at less than the speed of light, and thus covers less distance in the same time as the photon, so its slope is correspondingly larger. The slope of a worldline for an object is c/v , inversely proportional to its velocity.

The particle at rest has constant position for all times, and is thus a vertical line. The photon travels at the speed of light, so $t = x/c$ or $ct = x$. The rocket travels at less than the speed of light, and

thus covers less distance in the same time as the photon, so its slope is correspondingly larger. The slope of a worldline for an object is c/v , inversely proportional to its velocity. Suppose we want to describe your worldline. You start out from the origin at time $t=0$. Since you must travel at less than the speed of light, your world line must have a slope greater than one: your motion for increasing t will be restricted to the triangular region between the lines $ct=x$ and $ct=-x$. We can refer to this triangular region as your future, since it corresponds to the locus of all possible points you could reach. Similarly, the corresponding triangular region below the horizontal axis for previous times ($t < 0$) are points you could have been at in your past.

Points outside the triangular region would require you to exceed the speed of light, and are thus separated from you by a space-like interval: you can't even reach those points, nor could you have come from them. Points lying inside the triangular region are thus separated from you by a time-like region, since you could reach them by traveling at some velocity less than c . Essentially, your past and future can only be influenced by the spacetime intervals inside these two triangles defined by the two possible trajectories of light. Of course, this is all purely one-dimensional: extending this to two-dimensional motion, the lines would become *cones*, and are usually referred to as *light cones*. Only events within your light cone can influence your past or future, and no attainable speed can move you outside of your light cone. Your light cone is the region of spacetime available to you, in a way. The so-called forward light cone, extending above the horizontal axis, is your possible future, while the backward light cone must contain your past (or possible pasts, at least). We illustrate these basic ideas in Figure 1.18. Extending the description to three dimensions of space, we lose our power of visualization, since this would require four axes: three for space, and one for time. While the math works out perfectly fine in four dimensions, pictures do not . . .

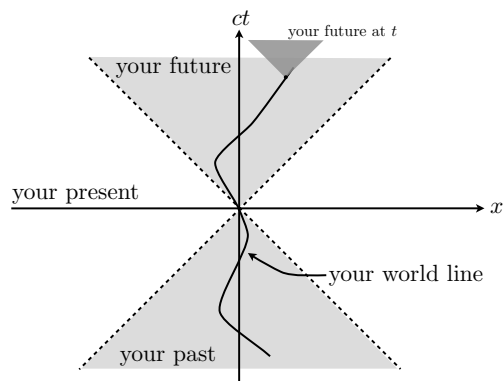


Fig. 1.18 Your past and future in a spacetime diagram.

What more can we do with this? Let's say we have two events at two different times and places. First, how should we notate this? Given that temporal and spatial coordinates are of equal importance in relativity, it is common to just pretend that time is simply another coordinate. For instance, in three-dimensional space, we might refer to the position of an object by its $x-y-z$ Cartesian coordinates, (x,y,z) . In relativity, we simply make time another coordinate, with two special distinctions. First, we multiply time by c to make it the distance light travels in the given time t , which keeps the units consistent and gives us a natural 'yardstick' for lengths. Second, the sign of the time coordinate is negative, consistent with our definition of the spacetime interval above. Thus, in relativity an event occurring at spatial coordinates (x,y,z) at time t would be written as $(-ct,x,y,z)$. This looks just like a normal position, or the definition of a vector, in *four* dimensions instead of the usual three, prompting the moniker *four vector*. Four vector sounds weird and mathematical, but it is really just lumping time and space coordinates together, with a factor of c to keep the units straight and a factor -1 to reflect the somewhat different nature of time compared to space. This is why you will often hear terms like "four-dimensional spacetime" and so on, which really just means three space dimensions plus one time coordinate. However, do not forget: time and space are clearly *not* the same thing, no matter what popular science accounts might tell you. The factor $-c$ is always there to remind you of as much.

Now, imagine we have two events taking place at two different spacetime coordinates. For convenience, we'll work in one dimension and say that the events have coordinates $(-ct_1, x_1)$ and $(-ct_2, x_2)$. Constructing spacetime interval s we discussed previously is nothing more than connecting these two points with a line. The slope of this line immediately tells you the nature of the spacetime interval: if the slope is greater than 1, the interval is time-like, and can be traversed by objects traveling at speeds less than c ; if the slope is smaller than 1, the interval is space-like, and the interval cannot be traversed by *anything* and no causal relationship can exist between these two events; if the slope is exactly 1, the interval is light-like, and can only be traversed by light.

This is illustrated in Fig. 1.19, showing three events A, B, and C. Event A precedes B and C, while C precedes B. For convenience, we place event A at the origin. Connecting events A and B gives a line whose slope has a magnitude greater than 1, so event B is within the light cone of event A. Therefore, it is possible that events A and B have a cause-effect relationship, and we would say they are connected by a time-like interval. This also means that it is possible to find a reference frame (by rotating our time-position coordinate system, for instance) in which A and B occur at the same place, but never one in which A and B occur at the same time.

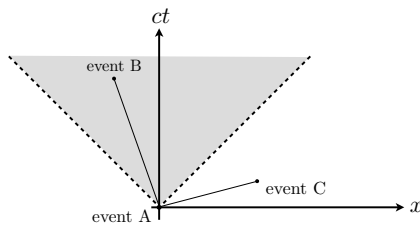


Fig. 1.19 Connecting different events on a spacetime diagram.

On the other hand, events A and C are connected by a line whose slope is smaller than 1, so C falls outside the light cone of A. There is no chance that A causes C or vice versa, but it *is* possible to find a reference frame in which A and C occur simultaneously. Incidentally, rotating our time-position coordinate system to find such a reference frame is nothing more than the Lorentz transformation we derived earlier! The Lorentz transformation corresponds “physically” to the allowed changes of coordinates that leave the spacetime interval unchanged, such as a rotation of the ct and x axes.

There is much more we can do with Minkowski diagrams, and a great deal that can be learned by recalling a bit of the geometry of hyperbolas. However, having covered the gist of the matter, we will leave this for another course, and move on relativistic energy and momentum.

1.4 Mass, Momentum, and Energy

So far, the simple principles of relativity have had enormous consequences. Our basic notions of time, position, and even simultaneity all needed to be modified. If position and time must be altered, then it stands to reason that *velocity* - the change of position with time - must also be altered. Sure enough, the velocity addition formula was also a required change. What next? If our notions of relative velocity need to be altered, then the next thing must surely be momentum and kinetic energy. As it turns out, even our concept of *mass* needs to be tweaked a bit.

1.4.1 Relativistic Momentum

First, let's consider momentum. Classically, we define momentum in terms of mass and velocity, $\vec{p} = m\vec{v}$. A basic principle of classical mechanics you have learned is that momentum must be conserved, no matter what. What about in relativity? In relativity, exactly what \vec{v} is depends on the reference frame in which it is measured. That means that our usual definition of momentum above depends on the reference frame as well. It gets worse. Using our simple $\vec{p} = m\vec{v}$, not only would the total amount

of momentum depend on the choice of reference frame, conservation of momentum in one frame would not necessarily be true in another. How can a fundamental conservation law depend on the frame of reference?

It cannot - this is one of our basic principles of relativity, *viz.*, the laws of physics are the same for all non-accelerating frames of reference. We *must* have conservation of momentum, independent of what frame in which the momentum is measured. How do we construct a new equation for momentum, one for which conservation of momentum is always valid, but at low velocities reduces to our familiar $\vec{p} = m\vec{v}$? The answer is that we need to be a bit more careful about our definition of velocity. Velocity is a change in position with time, but *what* position, and what time? From our discussion above, it is the spacetime interval which represents invariant “distance” which all observers can agree on, and correspondingly, the proper time. Do we simply divide these two quantities to get proper velocity? Not quite: remember that the contraction of distances was really just time dilation in another guise, and thus it is the passage of time we need to worry about. What we are interested in is how much distance is covered per unit of *proper* time.

Think of it this way: if you are traveling to a faraway city from your hometown, you can reliably determine the distance to the faraway city, since presumably the cities are not in relative motion (continental drift velocities being considered negligible). Your quantity of interest is then how much proper time elapses during your travel: the distance will remain the same, but the faster you travel, the shorter the trip seems to take, and the larger your apparent velocity if you simply noted the proper time at the start and end of the trip. A proper relativistic velocity is then *the distance you cover divided by proper time*, which just means an additional factor γ since there is no separation to worry about when you are the only observer ($\Delta r = 0$). If we call the proper velocity η , and the “ordinary” velocity of you relative to the faraway city v , your “ordinary” velocity is just the distance covered divided by the time you measure, $\Delta x / \Delta t$. Thus,

$$v_{\text{proper}} \equiv \eta = \frac{\Delta x}{\Delta t} = \frac{\Delta x}{\Delta t_p / \gamma} = \gamma \frac{\Delta x}{\Delta t} = \gamma v \quad (1.69)$$

The result is not surprising: we only need to transform velocity the same way we transformed position, and we have a definition of velocity which is consistent with our notions of relativity. Momentum can then be properly defined in the usual way, in terms of a body’s mass and its *proper* velocity:

Relativistic momentum:

$$\vec{p} = m\eta = \gamma m\vec{v} \quad (1.70)$$

A full derivation is a bit beyond the scope of our discussion, but defining momentum in this way makes it independent of the choice of reference frame, and restores conservation of momentum as a fundamental physical law. For low velocities ($v \ll c$), $\gamma \approx 1$, and this reduces to the familiar result. For velocities approaching c , the momentum grows much more quickly than we would expect. In fact, an object traveling at c would require *infinite* momentum (and therefore infinite kinetic energy), clearly an absurdity. This is one good reason why nothing with finite mass can ever travel at the speed of light! Only light itself, with no mass, can travel at the speed of light.

1.4.2 Relativistic Energy

The relativistic correction to momentum is straightforward. Given that kinetic energy depends on the momentum of an object (one can write $KE = p^2 / 2m$), one would expect a necessary revision for kinetic energy as well. This one is not so straightforward, however. First, we need to think about what we mean by energy in the first place.

In classical mechanics, for a single point mass in linear motion (*i.e.*, not rotating), the kinetic energy simply goes to zero when the body stops, $KE = \frac{1}{2}mv^2 = p^2 / 2m$. For an arbitrary body, however, the result is not so simple. If a composite object contains multiple, independently moving

bodies (such the individual atoms making up matter, for instance), the individual entities may interact among themselves and move about, and the object possesses *internal energy* E_i as well as the kinetic energy due to the motion of the whole mass. Overall, classically the kinetic energy of such a body is the sum of these two energies – the energy due to the motion of the object as a whole, and the energy due to the motion of the constituents of the object, $KE = \frac{1}{2}mv^2 + E_i$. Any moving body more complex than a single point mass has a contribution due to its internal energy.

In relativity, the kinetic energy does still depend on the motion of a body as a whole as well as its internal energy content. As with momentum, conservation of energy requires that *the energy of a body is independent of the choice of reference frame*, the *total energy* of a body cannot depend on the frame in which it is measured. The total energy – kinetic plus internal – must be the same in all reference frames. A derivation requires somewhat more math than we would like, but the result is simple:

Relativistic energy of a moving body:

$$E = \gamma mc^2 \quad (1.71)$$

This equation already tells us that the energy content of a body grows rapidly as v approaches c , and reaching the speed of light would require a body to have infinite energy. What is more interesting, however, is when the velocity of the body is *zero*, i.e., $\gamma = 1$. In this case, $E = mc^2$ – the body has finite energy even when not in motion! This is Einstein’s most famous equation, and it represents the fundamental equivalence of mass and energy. Any object has an *intrinsic, internal energy* associated with it by virtue of having mass. This constant energy is called the *rest energy*:

Rest Energy:

$$E_R = mc^2 \quad (1.72)$$

As Einstein himself put it, “Mass and energy are therefore essentially alike; they are only different expressions for the same thing.”[4] Matter is basically an extremely dense form of energy – is convertible into energy, and *vice versa*. In fact, the rest energy content of matter is enormous, owing to the enormity of c^2 – one gram of normal matter corresponds to about 9×10^{13} J, the same energy content as 21 kttons of TNT! It is the conversion of matter to energy that is responsible for the enormous energy output of nuclear reactions, such as those that power the sun, a subject we will return to.

The equivalence of matter and energy, or, if you like, the presence of an internal energy due solely to a body’s matter content, is an unexpected consequence of relativity. But we still have not determined the actual kinetic energy of a relativistic object! Again, the derivation is somewhat laborious, but the result is easy enough to understand. If we take the total energy of an object, Eq. 1.71, and subtract off the velocity-independent rest energy, Eq. 1.72, what we are left with is the part of a body’s energy that depends solely on velocity. This is the kinetic energy we are looking for, and it means the *total energy of a body is the sum of its rest and kinetic energies*:

Relativistic kinetic energy:

$$KE = (\gamma - 1)mc^2 \quad (1.73)$$

Total energy:

$$E_{\text{total}} = KE + E_R \quad (1.74)$$

Since $\gamma = 1$ when $v = 0$, the kinetic energy of a stationary body is zero, as we expect. At low velocities ($v \ll c$), one can show that this expression correctly reduces to $\frac{1}{2}mv^2$. As with the total

energy, for a body to actually acquire a velocity of c it would need an infinite kinetic energy, again, a primary reason why no object with mass can travel at the speed of light.

For completion, we should note that it is still possible to relate relativistic energy and momentum, just like it was possible to relate classical kinetic energy and momentum, though we will not derive the expressions here:

Relativistic energy-momentum equations:

$$E^2 - (pc)^2 = (mc^2)^2 \quad (1.75)$$

$$Ev = pc^2 \quad (1.76)$$

here p is the momentum of a body, m its mass, v its velocity, E its energy, and c is the speed of light. We can use this to write the relativistic kinetic energy and momentum equations in a different form:

$$KE = \sqrt{p^2c^2 + m^2c^4} - mc^2 \quad \text{and} \quad p = \sqrt{\frac{E^2}{c^2} - m^2c^4} \quad (1.77)$$

The first relationship bears a closer look: if a body's mass is not changing, then the right hand side of Eq. 1.75 is a constant, independent of reference frame. This is a more general relativistic combination of conservation of energy and momentum: the quantity $E^2 - p^2c^2$ is an invariant constant for all bodies, just as the spacetime interval and proper time are invariant constants for all observers. Energy and momentum are separately conserved, the idea that the combined quantity $E^2 - p^2c^2$ is *invariant* is slightly different: while mass, energy, and momentum are *conserved*, they are not *invariant*, they do indeed depend on velocity, and are therefore not *invariant*. However, the combination $E^2 - p^2c^2$ is an invariant constant for a given body, and does *not* depend on velocity! This is extremely useful, in fact, since it allows you to calculate E from p (or vice versa) without ever knowing v .

The energy content of a body still scales with its momentum, and for a body at rest ($p=0$), the energy content is purely the rest energy mc^2 . Once again we have an unexpected result, however: *objects with no mass must also have momentum, so long as they have energy*. For massless particles – such as the photons that make up a beam of light – we have the result $E = pc$, or $p = E/c$. This is truly another odd result of relativity, completely unexpected from classical physics! How can objects with no mass still have momentum? Since matter and energy are equivalent according to relativity, having energy is just as good as having mass, and still leads to a net momentum. This will become an important consideration when we begin to study optics and modern physics.

Momentum of massless objects:

$$p = \frac{E}{c} \quad (1.78)$$

If you combine Eqs. 1.76 and 1.78, you come to an even wilder conclusion. If the particle has zero mass, but *some* energy greater than zero, then we can write

$$v = \frac{pc^2}{E} = \frac{\frac{E}{c}c^2}{E} = c \quad (1.79)$$

A particle with zero mass always moves at the speed of light, and can never stop moving! It doesn't matter what the energy of the particle is, anything with finite energy but zero mass has to travel at the speed of light. The converse is true as well – anything moving at the speed of light must be massless. Just to drive the point home one last time: *the speed of light is an upper limit to physically attainable speeds for material bodies.*

1.4.3 Relativistic Mass

About the only thing left we have not modified with relativity is mass. Most modern interpretations of relativity consider mass to be an *invariant* quantity, properly measured when the body is at rest (or measured within its own reference frame). This rest mass of an object in its own reference frame is called the *invariant mass* or *rest mass*, and is an observer-independent quantity synonymous with our usual definition of “mass.”

These days, we say that while the *momentum* of a body must be the same in all reference frames, and hence must be transformed, the *mass* of a body is just a constant, and is measured in the body’s own reference frame. Rest mass is in some sense just counting the number of atoms in an object, something we really only do in the object’s reference frame anyway. If we are measuring an object from another reference frame, we will typically be measuring its *momentum*, or kinetic energy, not counting how many atoms it contains. Thus it is momentum and kinetic energy we transform to be invariant in all reference frames and mass we simply say is a property of an object measured in its own reference frame.

1.5 Problems

Solutions begin on page 237.

1.1. In the 1996 movie *Eraser*,^[5] a corrupt business Cyrez is manufacturing a handheld rail gun which fires aluminum bullets at nearly the speed of light. Let us be optimistic and assume the actual velocity is $0.75c$. We will also assume that the bullets are tiny, about the mass of a paper clip, or $m = 5 \times 10^{-4}$ kg.

- (a) What is the relativistic kinetic energy of such a bullet?
- (b) Let us further assume that Cyrez has managed to power the rail guns by matter-energy conversion. What amount of mass would have to be converted to energy to fire a single bullet? (For comparison, note that 1 kg of TNT has an equivalent energy content of about 4×10^9 J.)

1.2. An astronaut traveling at $v = 0.80c$ taps her foot 3.0 times per second. What is the frequency of taps determined by an observer on earth? (*Hint: be careful about the difference between time and frequency!*)

1.3. A spaceship moves away from earth at high speed. How do experimenters on earth measure a clock in the spaceship to be running? How do those in the spaceship measure a clock on earth to be running?

1.4. If you are moving in a spaceship at high speed relative to the earth, would you notice a difference in your pulse rate? In the pulse rate of the people back on earth?

1.5. The period of a pendulum is measured to be 3.00 in its own reference frame. What is the period as measured by an observer moving at a speed of $0.950c$ with respect to the pendulum?

1.6. The Stanford Linear Accelerator (SLAC) could accelerate electrons to velocities very close to the speed of light (up to about $0.9999999995c$ or so). If an electron travels the 3 km length of the accelerator at $v = 0.999c$, how long is the accelerator from the *electron's* reference frame?

1.7. A spacecraft with the shape of a sphere of diameter D moves past an observer on Earth with a speed $0.5c$. What shape does the observer measure for the spacecraft as it moves past?

1.8. Suppose you're an astronaut being paid according to the time you spend traveling in space. You take a long voyage traveling at a speed near that of light. Upon your return to earth, you're asked how you would like to be paid: according to the time elapsed by a clock on earth, or according to the ship's clock. Which do you choose to maximize your paycheck?

1.9. Show that the kinetic energy of a (non-relativistic) particle can be written as $KE = p^2/2m$, where p is the momentum of a particle of mass m .

1.10. A pion (π) at rest ($m_\pi = 273m_{e^-}$) decays to a muon (μ ; $m_\mu = 207m_{e^-}$) and an antineutrino ($\bar{\nu}$; $m_{\bar{\nu}} \approx 0$). This reaction is written as $\pi^- \rightarrow \mu^- + \bar{\nu}$. Find the kinetic energy of the muon and the energy of the antineutrino in electron volts. *Hint: relativistic momentum is conserved.*

1.11. An alarm clock is set to sound in 15 h. At $t = 0$ the clock is placed in a spaceship moving with a speed of $0.77c$ (relative to Earth). What distance, as determined by an Earth observer, does the spaceship travel before the alarm clock sounds?

1.12. The average lifetime of a pi (π) meson in its own frame of reference (*i.e.*, the proper lifetime) is 2.6×10^{-8} s

- (a) If the meson moves at $v = 0.98c$, what is its mean lifetime as measured by an observer on earth?
- (b) What is the average distance it travels before decay, measured by an observer on Earth?
- (c) What distance would it travel if time dilation did not occur?

1.13. You are packing for a trip to another star, and on your journey you will travel at $0.99c$. Can you sleep in a smaller cabin than usual, because you will be shorter when you lie down? Explain your answer.

1.14. A deep-space probe moves away from Earth with a speed of $0.88c$. An antenna on the probe requires 4.0 s, in probe time, to rotate through 1.0 rev. How much time is required for 1.0 rev according to an observer on Earth?

1.15. A friend in a spaceship travels past you at a high speed. He tells you that his ship is 24 m long and that the identical ship you are sitting in is 18 m long.

- (a) According to your observations, how long is your ship?
- (b) According to your observations, how long is his ship?
- (c) According to your observations, what is the speed of your friend's ship?

1.16. A Klingon space ship moves away from Earth at a speed of $0.700c$. The starship Enterprise pursues at a speed of $0.900c$ relative to Earth. Observers on Earth see the Enterprise overtaking the Klingon ship at a relative speed of $0.200c$. With what speed is the Enterprise overtaking the Klingon ship as seen by the crew of the Enterprise?

1.17. An observer sees two particles traveling in opposite directions, each with a speed of $0.99000c$. What is the speed of one particle with respect to the other?

1.18. How fast must a meter stick be moving if its length is measured to be only 0.50 m?

1.19. For what velocity does $\gamma = 1.05$? For speeds lower than this, time dilation and length contraction are effects are less than 5%.

1.20. At what speed is a clock moving if it is measured to run at half the rate as a clock at rest with respect to an observer?

1.21. A spaceship is measured to be 75.0 m long and 25.0 m wide while at rest. Another observer views the spaceship as it flies by at $0.95c$. What length and width does this observer measure?

1.22. Two spacecraft are moving in opposite directions. An observer on earth measures the speed of the first to be $0.80c$, and the speed of the second to be $0.9c$. What is the velocity of the second craft as observed by passengers of the first?

1.23. A car speeds past an observer on the ground at $0.9c$. A passenger in the car throws a ball out the car window at $0.7c$ relative to the car. What is the velocity of the ball with respect to the observer on the ground?

1.24. A muon formed high in the Earth's atmosphere travels at $v = 0.990c$ for 4.60 km before it decays into an electron, a neutrino, and an antineutrino ($\mu^- \rightarrow e^- + \nu + \bar{\nu}$). (a) How long does the muon live, as measured in its own reference frame? (b) How far does the Earth travel, as measured in the frame of the muon?

1.25. The nonrelativistic expression for the momentum of a particle $p = mv$ agrees with experiments when $v \ll c$. For what speed does the nonrelativistic equation give an error of (a) 1.0%? (b) 5.0%?

Part II
Electricity and Magnetism

Chapter 2

Electric Forces and Fields

The most incomprehensible thing about the world is that it is at all comprehensible. – Albert Einstein

Abstract Electricity has become ubiquitous in modern life, so much that we rarely think about life without it. Though ancient Greeks first began experimenting with electricity around 700 B.C., it was not until the 18TH and 19TH centuries that we began to clearly understand electricity and how to harness it. In this chapter, we will discuss electric charges and the electric force, quantified through Coulomb’s law, and introduce the electric field associated with charges. With these concepts, we will be able to explain many of the myriad electrostatic phenomena around us.

2.1 Properties of Electric Charges

Probably you have noticed that after running a plastic comb through your hair, the comb will attract bits of paper. Often this attraction is strong enough to suspend the paper from the comb, completely counteracting the force of gravity. Another simple experiment is to rub an inflated balloon against your shirt or hair, with the result that the balloon will then stick to the wall or ceiling.

Both of these situations arise because the materials involved have become electrically charged. The same thing happens when you get “shocked” after dragging your feet on the carpet – you have built up electric charge on your body. An object that is electrically charged has built up an imbalance of electric charge. What is electric charge though? Experiments have demonstrated a few basic facts about electric charges:

Some basic properties of charges:

1. There are two types of electric charge, **positive** and **negative**.
2. Like charges repel one another, unlike charges attract one another.
3. Charge comes in discrete units.
4. *Protons* are the positive charges, *electrons* are the negative charges.
5. Electrically neutral objects have an equal number of positive and negative charges.
6. Electrically neutral objects do not experience an electric force in the presence of electric charges.

Normal objects usually contain equal amounts of positive and negative charges – they are electrically neutral. Electric forces arise only when there is an imbalance in electric charge, and objects carry a net positive or negative charge. On the atomic scale, the carriers of positive charge are the protons. Along with neutrons, which have no electric charge, they comprise the nucleus of an atom (which is about 10^{-15} m across). Electrons are the carriers of negative charge. In a gram of normal matter, there are about 10^{23} protons and an equal number of electrons, so the net charge is zero.

Electrons are far lighter than protons, and are more easily accelerated by forces. In addition, they occupy the outer regions of atoms, and are more easily gained or lost. Objects that become charged to

Table 2.1 Properties of electrons, protons, and neutrons

Particle	Charge [C]	$[e]$	Mass [kg]
electron (e^-)	-1.60×10^{-19}	-1	9.11×10^{-31}
proton (p^+)	$+1.60 \times 10^{-19}$	+1	1.67×10^{-27}
neutron (n^0)	0	0	1.67×10^{-27}

so by gaining or losing electrons, not protons. Table 2.1 gives some properties of protons, electrons, and neutrons.

Charge can be transferred from one material to another. Many chemical reactions are, in essence, charge transfer from one species to another (see page 66 for some examples). Rubbing two materials together facilitates this process by increasing the area of contact between the materials – e.g., rubbing a balloon on your hair. Since it is a gain or loss of electrons that give a net charge, this means that when objects become charged, **negative charge is transferred from one object to another**.

Units of charge: The SI unit of charge is the **Coulomb, [C]**.

One unit of charge is $e = 1.6 \times 10^{-19} [C]$

Charge is never created or destroyed, only transferred from one object to another. **Objects become charged by gaining or losing electrons**, transferring them to other objects. **Charge is also quantized**, meaning it only comes in multiples of the fundamental unit of charge e .

An object can have a charge of $\pm e, \pm 2e, \pm 3e$, etc, but not $+0.27e$ or $-0.71e$.¹ Electrons have a negative charge of one unit ($-e$), and protons have a positive charge of one unit ($+e$). The SI unit of charge is the **coulomb [C]**, and e has the value $1.6 \times 10^{-19} C$. Since e is so tiny when measured in Coulombs, and since it is the basic fundamental unit of charge, we will sometimes simply measure a small amount of charge in “ e ’s” – how many individual unit charges are present.

Summarizing the properties of charge:

1. Charge is quantized in units of $|e| = 1.6 \times 10^{-19} C$
2. Electrons carry one unit of negative charge, $-e$
3. Protons carry one unit positive charge, $+e$
4. Objects become charged by gaining or losing electrons, not protons
5. Electric charge is always conserved

2.2 Insulators and Conductors

How do materials respond to becoming charged, and how do we charge up a material in the first place? What do we mean by “becoming charged” anyway? This will be more clear shortly, but for now, we will presume that “charging” simply means creating an imbalance of electric charges in a material. A net negative charge can be achieved by adding excess electrons to a material, and a net positive charge can be created by taking away some electrons from a material.

For our purposes, materials respond to becoming charged in one of two ways: the excess charge can move about freely and evenly distribute themselves, or the excess charge can stay localized to the region where it was created. Conductors and insulators are the two broad classes of materials, respectively, which fit these criteria - in **conductors**, excess charges move freely in response to an electric force. All other materials are insulators, and the charges do not move!

¹ Quarks are an exception we will cover at the end of the semester.

In fact, there is nothing particularly special about the excess charge. The excess charge will move in the material in the same way any other charges do - we can't tell the charges apart. In other words, conductors are materials in general where charges move freely, and insulators are materials in general where they do not. There does not need to be *excess* charge for this to be true, charges inside conductors are still in motion even if, over all, they cancel each other out.

Conductors:

1. *e.g.*, metals – silver, gold, aluminum, steel, *etc.*
2. charges are mobile, and move in response to an electric force
3. large number of charges
4. charge distributes evenly over surface

Insulators:

1. *e.g.*, glass, most ceramics, rubbers, and plastics
2. charges are immobile
3. charge deposited on insulators stays localized

Semiconductors:

1. *e.g.*, silicon, gallium arsenide, germanium
2. in between conductors and insulators
3. charges are highly mobile ...
4. ... but the number of charges is small, depends *e.g.*, on temperature and purity
5. conducting properties can be widely varied

Copper and aluminum are typical conductors. When conductors are charged in some small region, the charge readily distributes itself over the entire surface of the material. Thus, on a conductor charge is always equally distributed over its entire surface. Charge flows through a conductor readily, and if given a chance, out of it. This is an electric current, as we will see shortly. Glass and rubber are typical insulators. When insulators are charged (*e.g.*, by rubbing), only the rubbed areas are charged. There is no tendency for the charge to flow to other regions of the material - charge deposited on insulators will stay localized to a small region.

2.2.1 Charging by Conduction

Conduction is charging through physical contact, which moves e^- from one object to another. One example is charging a balloon by rubbing it on your hair. After doing this, the balloon easily sticks to a wall or picks up little bits of paper, and your hair stands a bit on end. What you have really done is transferred charges from the balloon to your hair, or *vice versa*. Each of your individual hairs becomes charged the same way (either all positive or all negative, depending on what you rubbed on your hair), and the individual strands repel each other. Their repulsion makes them want to maximize the distance between them, which is achieved by standing on end, radiating outward.

As another example, consider rubbing an insulating rod (*e.g.*, rubber, hard plastic glass) against a piece of silk. The act of rubbing these two insulating materials will physically force some charges to move from one object to the other. When charges are transferred to the insulating rod, they do not move – regions of localized charge are created in the rubbed regions. **No charge has been created or destroyed, we simply moved some charges from one place to another - one object ends up with a net positive charge, the other with a net negative charge, equal in magnitude.** One can verify that both objects are charged by trying to pick up bits of paper with them. This is also true when you rub a balloon on your hair - it is clear immediately that *both* the balloon and your hair have become charged! It couldn't be any other way, or we would have had to create charges out of thin air.

Figure 2.1 and its accompanying box illustrates the process of charging a metallic object by conduction. In this example, you take a rubber rod you have already charged (say, with a piece of silk or your hair), and use that to charge a third object.

Charging by conduction: follow Figure 2.1

1. Take a charged rod of rubber
2. Bring it near a metal conducting sphere.
3. The charged rod redistributes the charge on the sphere
4. Contact the sphere and the rod!
5. Charges want to neutralize, opposites attract.
6. Negative charge leaves the rod to neutralize the positive charges on the sphere
7. This leaves a net negative region on the sphere

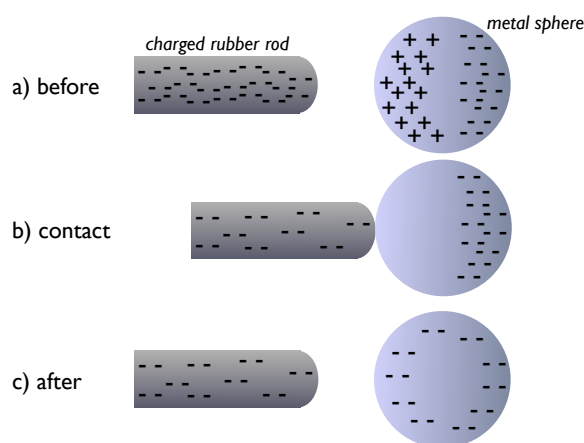


Fig. 2.1 Charging a metallic object by conduction. **a)** Just before contact, the negatively charged rod repels the conducting sphere's electrons. **b)** After contact, electrons from the rod flow onto the sphere, neutralizing the positive charges. **c)** When the rod is removed, the sphere is left with a negative charge.

2.2.1.1 Grounding

Sounds simple enough. Why can't we just take a piece of Copper pipe and rub it with a cloth? You can, if you are careful ... charges flow evenly through a conductor, and if possible, out of the conductor entirely. Only isolated conductors can be charged, electrically contacted conductors cannot. By 'electrically connected,' we mean the conductor we are trying to charge cannot have any sort of conducting path to the earth. The Earth can be considered an (essentially) infinite reservoir for electrons, either sourcing or sinking as many charges as we need. Since charges distribute themselves evenly over a conducting surface, if there were a path to the earth, the mobile charges would follow it to the earth, and keep doing so until none were left on the conductor.


Given a conducting path to the earth, charges from the conductor **will always keep flowing**. If charges can find a way to Earth, they *will* get there (*e.g.*, through pipes, wires, or you!). Another phrase for when *you* are the connection to Earth is "ground fault." This is when you accidentally make *yourself* the connection between a charged (or current-carrying) wire and the Earth ... with potentially disastrous results. A so-called "ground fault interrupter" (GFI) senses when this happens, and very quickly breaks the connection.²

As it turns out, *YOU* are a sufficient conductor to let the charges flow away! Any charges transferred to the Copper rod will flow straight through it, and through you down to the ground. You can make it work, if you wear some insulated rubber or plasticized gloves. The same trick works on a

² Electrical outlets with GFI should be present in your house, usually in bathrooms and kitchens. They typically have little buttons labeled 'test' and 'reset' or something like that.

rubber or plastic rod without gloves, because charges deposited on the insulating rods do not flow through the rod and out of it.

The “ground connection” or “ground point” is the place in an electric circuit which is *purposely* connected to the earth, either for safety reasons or just to provide a reference point. The ground point (or just “ground”) in a circuit or electrical diagram is usually shown like this:

Circuit diagram symbol for a ground point: 

2.2.2 Charging by Induction

Can we charge without contacting it at all? Yes! This is **induction charging**. Now we explicitly need a **ground point** or reference point for this to work though. An object connected to a conducting wire or pipe buried in the Earth is said to be **grounded**, the Earth itself is the *ground point*. As mentioned above, the Earth can be considered an infinite reservoir for electrons, sourcing or sinking an infinite number of charges. Using this idea, we can understand a non-contact charging process known as **induction**.

Figure 2.2 illustrates the process of charging a metallic object by induction. **Charging an object by induction requires no contact with the object inducing the charge.** First, we take an isolated conducting (metal) sphere. From our discussion above, it is crucial that it not be contacted to the ground in any way. Placing it on an insulating stand will do nicely. Next, we bring a negatively charged rod *near*, but not touching, the sphere. We can prepare a negatively charged glass rod by rubbing it with silk (charging the glass by conduction).

Charging by induction: follow Figure 2.2

1. Take a neutral conducting sphere
2. Bring a negatively charged rod *near* (but not touching) the sphere.
3. This creates a charge imbalance on the sphere, due to repulsion from the charged rod.
4. Ground the opposite side of the sphere – the charge imbalance forces some e^- to flow to ground!
5. Disconnect the ground wire – this leaves a net $+$ charge on sphere!
6. Remove the charged rod, the net charge has to stay on sphere, and it will distribute itself evenly over the surface.

When the charged rod is near the conducting sphere, the negative charges on the rod will repel the free negative charges (electrons) on the sphere, with the result that the half of the sphere nearest the rod will have a net negative charge (Fig. 2.2b). Now, if we take a conducting wire and connect the *far* end of the sphere to the ground (Fig. 2.2c), the excess negative charge on that side, repelled by the rod, will want to flow down the wire into the earth, effectively draining away a quantity of negative charge from the sphere. Once we have done that, the sphere now has a *net positive charge*.

Removing the ground connection, Fig. 2.2d, will instantaneously leave the side of the sphere near the rod positively charged, and the far side (nearly) uncharged, since we just drained away the negative charges. After a *very* short time, the conducting sphere reaches equilibrium, and we must have a uniform distribution of charge on the surface of the conductor. Thus, the excess positive charge has to be evenly distributed on the surface of the sphere. We are left with a charged conducting sphere!

A process similar to charging by induction in conductors takes place in insulators (such as neutral atoms or molecules in particular). The presence of a charged object can result in more positive charge on one side of an insulating body than the other, by realignment of the charges within the

2.1. Why must hospital personnel wear special conducting shoes while working around oxygen in an operating room? What might happen if they wore shoes with rubber soles?

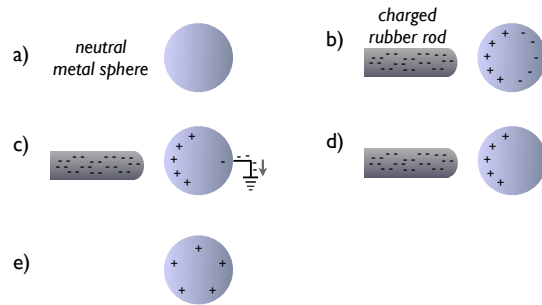


Fig. 2.2 Charging a metallic object by *induction*. **a)** A neutral metallic sphere with equal numbers of positive and negative charges. **b)** The charge on a neutral metal sphere is redistributed when a charged rod is brought near it. **c)** When the sphere is then grounded, some of the negative charges (electrons) leave it through the ground wire. **d)** When the ground connection is removed, excess positive charge is left on the sphere. **e)** When the charged rod is removed, the excess positive charge redistributes itself until the sphere's surface is uniformly charged.

individual molecules. This process is known as **polarization**, and we will cover it in more depth in the following chapter.

Our discussion of charging allows us to now better appreciate the distinction between conductors and insulators. The difference in the degree of conductivity between conductors and insulators is staggeringly enormous, a factor of 10^{20} . For instance, a charged Copper sphere connected to the ground loses its charge in a millionth of a second, while an otherwise identical glass sphere can hold its charge for years.

2.3 Coulomb's Law

When you charge two objects, such as a balloon and your hair, you invariably end up observing an attraction or repulsion between the charged objects. What is the character of this force? How does it depend on how much the objects have been charged, how far away they are, or anything else? If you continue to experiment with charged objects, you will find that the *force* due to electrically charged objects has the following properties:

An **electric force** has the following properties:

1. It is directed along a line joining the two particles.
2. It is inversely proportional to the square of the distance r_{12} separating them.
3. It is proportional to the product of the magnitudes of the charges, $|q_1|$ and $|q_2|$, of the two particles.
4. It is attractive if the charges are of opposite sign, and repulsive if the charges have the same sign

These properties led Coulomb³ to propose a neat mathematical form for the electric force between two charges:⁴

Coulomb's Law: the force between two charges q_1 and q_2 , separated by a distance r_{12} is given by:

$$\vec{F} = k_e \frac{q_1 q_2}{r_{12}^2} \hat{r}_{12} \quad (2.1)$$

³ Charles Augustin de Coulomb (1736–1806), a French physicist. He discovered an inverse relationship on the force between charges and the square of their separation, later named after him.

⁴ See Sect. ?? for a summary of units and notation conventions

where k_e is the “Coulomb constant,” and \hat{r}_{12} is a unit vector pointing along a line connecting the two charges.

Equation 2.1 is known as “**Coulomb's law**”. What Coulomb's law states is that the force between two charged objects \vec{F} , depends only on how big the charges are (q_1 and q_2), and how far apart they are (r_{12}). Keep in mind that force is a vector, and the dimensionless *unit vector* \hat{r}_{12} reminds us that the electric force is directed along a line connecting the two charges q_1 and q_2 .

Figure 2.3 schematically shows the electric force between two like and two unlike charges. The distance between the charges r_{12} is given in the SI unit of **meters, [m]**, and the charges q_1 and q_2 are measured in the SI unit of charge, the **Coulomb, [C]**. The charges q_1 and q_2 can be either positive or negative, which makes the resulting force \vec{F} repulsive when both charges have the same sign, and attractive when they are opposite – just like we expect. The Coulomb constant k_e gives the relative strength of the electric force, just as G gives the relative strength of the gravitational force, and has the SI value and units:

Coulomb's constant

$$k_e = 8.9875 \times 10^9 \frac{\text{N} \cdot \text{m}^2}{\text{C}^2} \quad (2.2)$$

In most calculations, k_e can be safely rounded to $\approx 9 \times 10^9$, which makes it a bit easier to remember. Also, k_e is much, much larger than G^5 , by about *twenty* orders of magnitude, meaning that if we treat Coulombs on equal footing with kilograms for a minute, the electric force is *far, far* stronger than the gravitational force. A pair of 1 Coulomb charges interacting *via* the electric force is the same as two masses of 10^{10} kilograms interacting *via* the gravitational force. Equivalently, one might say gravity is just exceptionally weak, so far as fundamental forces go.

Note that no matter what the two charges are, Newton's third law⁶ still holds, *viz.*, $\vec{F}_{21} = -\vec{F}_{12}$. The force on charge 1 due to charge 2 is equal and opposite the force on charge 2 due to charge 1, *always*. Even if one charge is a million times larger than the other, this must still be true. Mathematically, this is easy to see from Eq. 2.1 – the force between two charges depends on the *product* of the two charge values $q_1 q_2$, which means it is totally symmetric if we swap 1 for 2 or *vice versa*.

2.2. Show that the units of Coulomb's constant in Equation 2.2 give a force in Newtons when used in Equation 2.1.

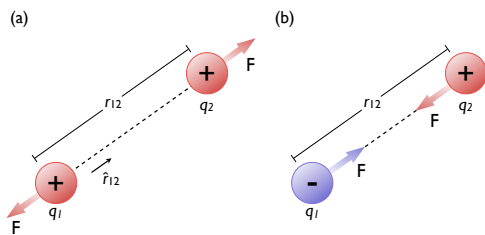


Fig. 2.3 Electrical force between point charges. **(a)** Two particles q_1 and q_2 which both have positive charges. The force is repulsive, as it would be for two negative charges, and directed along the dashed line connecting the two charges. The unit vector \hat{r}_{12} is indicated. **(b)** Two particles q_1 and q_2 with charges of opposite sign, separated by a distance r_{12} . The force is now attractive, as we expect.

When a number of separate charges act on a single charge, each exerts its own electric force. These electric forces can all be computed separately, one at a time, and then added as vectors. This is the powerful **superposition principle**, the same one you used with gravitation. This makes calculating the net force from many charges a lot simpler than you might think. In fact, gravitation and electrostatic forces have a number of similarities, with a few crucial differences, which we list below.

⁵ $G = 6.67 \times 10^{-11} \text{ m}^3/\text{kg} \cdot \text{s}^2$

⁶ When object A exerts a force on object B, B simultaneously exerts a force on A with the same magnitude, in the opposite direction.

2.3. At what distance below a proton would the upward force on an electron equal the electron's weight?

The electric force is *similar* to the gravitational force:

1. Both act at a distance without direct contact
2. Both act in a vacuum, without a medium, and propagate at a speed c
3. Both are inversely proportional to the distance squared, with the force directed along a line connecting the two bodies
4. The mathematical form is the same, if one interchanges k_e and G .
5. Both gravitational masses and electric charges obey the superposition principle
6. Both are conservative forces⁷

The electric force is *different* from the gravitational force:

1. Electric forces can be *attractive* or *repulsive*. Gravity is only attractive.
2. Gravitational forces are independent of the medium, while electric forces depend on the intervening medium
3. The electric force between charged elementary particles is far stronger than the gravitational force between the same particles.

One lingering question is how to relate the *microscopic* charge carriers, the electrons, to the *macroscopic* behavior of charged objects. When we charge a glass rod and pick up bits of paper, how many charges are we dealing with? Referring to Table 2.1, the charge on the proton (p^+) has a magnitude of $e = 1.6 \times 10^{-19} \text{ C}$, while an electron (e^-) has a charge of $-e = -1.6 \times 10^{-19} \text{ C}$.⁸ This means it takes $1/e \approx 6.3 \times 10^{18}$ protons or electrons to make up a total charge of $\pm 1 \text{ C}$ – so 1 C is a *seriously* large amount of charge. Typical net charges in electrostatic situations (*i.e.*, static electricity) are of the order of $1 \mu\text{C}$,⁹ which is still 10^{12} or so electrons – or about one electron for every dollar of our national debt, if that helps bring the magnitude in perspective!

Technically, Coulomb’s law applies in this particular mathematical form only for point charges, or spherical charge distributions (in which case r_{12} is the distance between the centers of the charge distributions, see Sect. 2.8.3). Coulomb’s law covers **electrostatic forces**, which are what we call forces between unmoving (stationary) charges. Really, though we only need to take care when we have charges moving at very high velocities, or when charges accelerate. Accelerating charges produce electromagnetic radiation – light – which we will cover in Chapter 8

2.4 The Electric Field

Both the gravitational force and the electrostatic force are capable of acting through space, without any physical contact or intervening medium (Sects. 1.2.1, 8.5). That is, electric and gravitational forces can act across an empty vacuum, with no matter to carry them.¹⁰ These types of forces are known as **field forces**. Corresponding to the electrostatic force, an **electric field** is said to exist in the region of space surrounding a charged object. **The electric field exerts an electric force on any other charged object within the field.**

The field concept partially eliminates the conundrum of “force at a distance”, since the force on a charged object is now said to be caused by the *electric field* at that point in space.

⁷ A conservative force is one which does no net work on a particle that travels along any *closed* path in an isolated system. For any path, not just a closed one, the work done by a conservative force depends only on the initial and final positions, not on the path taken. Gravity is conservative, friction is not, for example.

⁸ The symbol e will frequently be used to represent the charge of a proton or electron.

⁹ $1 \mu\text{C} = 10^{-6} \text{ C}$, see Table ?? in Sect. ??

¹⁰ We will find out later that *light* carries electric forces, in fact, and there is no need to invoke “action at a distance.”

The electric field \vec{E} produced by a charge q at the location of a small “test” charge q_0 is defined as the electric force \vec{F} exerted by q on q_0 , divided by the test charge q_0 .

$$\vec{E} = \frac{\vec{F}}{q_0} \quad \text{or,} \quad \vec{F} = q_0 \vec{E} \quad (2.3)$$

The SI unit for electric field is **Newtons per Coulomb [N/C]**. The direction of \vec{E} is the direction of the force that acts on a positive test charge q_0 placed in the field.

The test charge q_0 is hypothetical – what *would* the force be on a charge q_0 if we *did* place it at some distance r away? We say that an electric field exists at a point if a test charge at that point would be subject to an electric force there.

Using equations 2.1 and 2.3, we can write the *magnitude* of the electric field¹¹ due to a charge q as:

Magnitude of the electric field at a distance r from a point charge q :

$$|\vec{E}| = k_e \frac{|q|}{r^2} \quad (2.4)$$

The *direction* of the electric field is the same as the direction of the electric force, since the two are related by a scalar.

The electric field produced by a charge depends only on the magnitude of that charge which sets up the field, and how far away from that charge you are. It does not depend on the presence of a hypothetical test charge.

The principle of superposition also holds for electric fields, just as it did for the electric *force*. In order to calculate the electric field from a group of charges, one may calculate the field from each charge individually, and add (as vectors) the individual fields. Symmetry is also very important. For example, if a equal and opposite charges are placed on the x axis at $x = a$ and $x = -a$, the field at the origin is zero – the fields from the positive and negative charges cancel. On page 66 you can find basic instructions on how to approach and solve electric field problems.

Table 2.2 Approximate electric field values, in [N/C]

Source	$ \vec{E} $	Source	$ \vec{E} $
Fluorescent lighting tube	10^1	Atmosphere (fair weather)	10^2
Balloon rubbed on hair	10^3	Atmosphere (under thundercloud)	10^4
Photocopier	10^5	Spark in air	10^6
Across a transistor gate dielectric	10^9	Near electron in hydrogen atom	10^{11}

2.4.1 Electric Field Lines

A convenient way to visualize the electric field is to draw lines pointing in the direction of the electric field vector at any point – **Electric field lines**. Electric field lines have three key properties:

¹¹ When it is unambiguous, we will often write the magnitude of a vector, such as the electric field $|\vec{E}|$, as simply E for convenience. Similarly, $|\vec{B}|$ becomes B , and \vec{x} becomes x .

Key properties of electric field lines:

1. The electric field vector \vec{E} is tangent to the electric field line at any point.
2. The density of the lines (number per unit area) is proportional to the magnitude of \vec{E} .
3. Arrows on the lines point in the direction that a hypothetical *positive* test charge would move. Arrows are not always used.

So \vec{E} is large when the lines are close together, and small when they are far apart. Below are some example, to give you an idea.

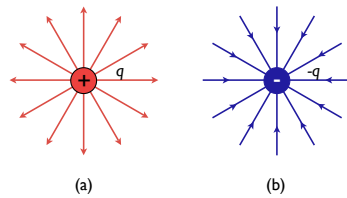


Fig. 2.4 The electric field lines for point charges. (a) For a positive point charge, the lines are directed radially outward. (b) For a negative point charge, the lines are directed radially inward. Note that the figures show only those field lines that lie in the plane of the page. (c) The dark areas are small pieces of thread suspended in oil, which align with the electric field produced by a small charged conductor at the center.

These 2-D drawings represent field lines for individual point charges. They only contain field lines in the plane of the paper – there are equivalent field lines pointing in all directions. A positive “test charge” placed in the field of the positive charge Fig. 2.4a field would be repelled, hence the lines point outward. On the other hand, for the negative charge in Fig. 2.4b, a positive test charge is attracted and the arrows point in. Note that the lines get more dense as they get closer to the charge, indicating that the field strength is increasing – just what we expect from Equation 2.4.

Rules for drawing field lines:

1. The lines for a group of point charges must **start on positive charges and end on negative charges**. If there is excess charge, some lines will begin or end infinitely far away (or at least off of your page).
2. The number of lines drawn leaving a positive charge or ending on a negative charge is proportional to the magnitude of the charge
3. Field lines cannot cross each other.

2.4.2 What happens when we have two charges together?**2.4.2.1 Two Opposite Charges**

Figure 2.5 shows nicely symmetric field lines for two charges of equal magnitude and opposite sign. Here we have omitted the arrows for simplicity, by now you should know how to add them in. This configuration is also known as an **electric dipole**. The number of lines beginning at the positive charge must equal the number of lines ending at the negative charge. Close to each charge, the lines are nearly radial, and the high density of lines between the charges indicates a large electric field in this region. Finally, note that the lines are symmetric about a line connecting the two charges, and to a line perpendicular to that one halfway between the charges.



Fig. 2.5 **left** Field lines for two equal and opposite charges, an “electric dipole.” The number of lines leaving the positive charge equals the number terminating at the negative charge **right** Field lines for two positive charges of equal magnitude. Can you rank the relative field strengths at points A, B, and C?¹³

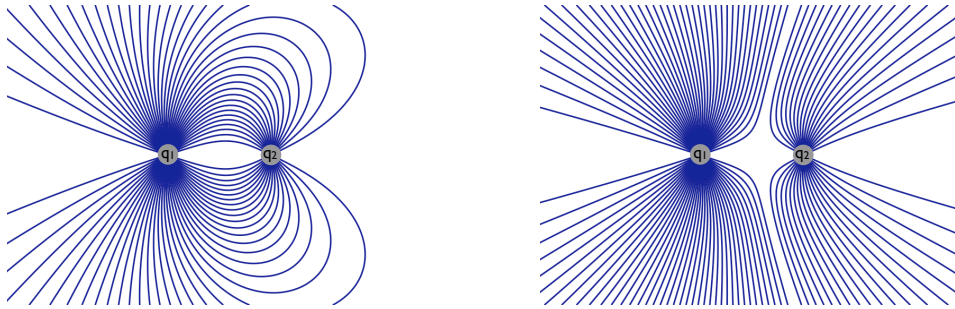


Fig. 2.6 **left** Field lines for opposite charges of different magnitude. Which is the larger charge? **right** Field lines for two charges of the same sign, but of different magnitude. Which is the larger charge?

2.4.2.2 Two like charges

Figure 2.5 also shows the field lines for two positive charges. Again the lines are nearly radial near the charges. The same number of lines leave each charge, since they are of the same magnitude.

Far away from either charge, the field looks nearly the same as it would for a single charge twice as big as either lone charge. In between the charges, the field lines “bulge,” representing the repulsive nature of the electric force between like charges. Again, note that the lines are symmetric about a line connecting the two charges, and to a line perpendicular to that one halfway between the charges. The symmetries of the electric field surrounding charge distributions can be very useful in solving electric field problems – for instance, we know without lifting a pencil that the field is precisely zero along the vertical line halfway between the two charges.

Finally, Fig. 2.6 shows the field lines for two charges of different magnitude in two situations. Can you tell which is the larger charge in each case? Can you tell which plot is for charges of the same sign, and which is for charges of opposite signs?

2.5 Conductors in Electrostatic Equilibrium

A good electric conductor like copper, even when electrically neutral, contains electrons which aren’t bound to any particular atom, and are free to move about. This is one reason why charge is distributed evenly over the surface of a conductor – the mobile electrons.

Though the individual “free” electrons in the conductor are constantly in motion, in an isolated conductor there is no *net* motion of charge. The random motions of all free electrons cancel out over all. When no net motion of charge occurs, this is called **electrostatic equilibrium**. An isolated conductor is one that is insulated from the ground, and has the following properties:

Properties of isolated conductors:

1. The electric field is zero everywhere inside the conductor.
2. Any excess charge on an isolated conductor must be entirely on its surface.
3. The electric field just outside a charged conductor is perpendicular its surface.
4. On irregularly shaped conductors, charge accumulates at sharp points, where the radius of curvature is smallest.

The first property is most easily understood by thinking about what would happen if it were **not** true, *reductio ad absurdum*. If there were fields inside a conductor, the free charges would move, and “bunch up” at the regions of higher and lower field (depending on whether they are positive or negative). This contradicts the very definition of a conductor – charges are supposed to be mobile, and spread out *evenly* through the conductor. Even if we did create a field inside a conductor, since the charges are mobile they would immediately start to flow to the region where the electric field is, gathering in sufficient number until they cancelled it out. Anyway, if this happened, we would no longer have electrostatic equilibrium in the first place, which is defined by *no net motion of charges*.

The second property is a result of the $1/r^2$ repulsion of like charges in Equation 2.4. If we had excess charge inside a conductor, the repulsive forces between these excess charges would push them as far apart as possible. Since the charges are mobile in a conductor, this happens readily. Every like charge wants to maximize its distance from every other like charge, so excess charge quickly migrates to the surface.

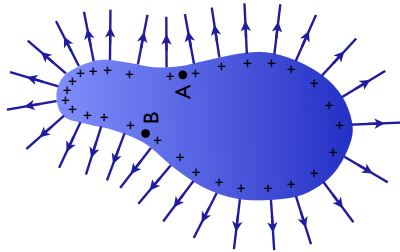


Fig. 2.7 An arbitrarily shaped conductor carrying a positive charge. When the conductor is in electrostatic equilibrium, all of the charge resides at the surface, $\vec{E} = 0$ inside the conductor, and the direction of \vec{E} just outside the conductor is perpendicular to the surface. Note from the spacing of the positive signs that the surface charge density is nonuniform due to the varying degree of curvature along the surface.

This is only true because Coulomb’s law (Equation 2.4) is an inverse square law! If it were some other power law, like $1/r^{2+\delta}$, even for *very* tiny δ , excess charges would exist inside the conductor, which we could observe. One of many special facts about inverse square laws, which has been used to test Coulomb’s law with fantastic precision.

The third property we also understand by thinking about what would happen if it were not true. If the field was not perpendicular to the conductor’s surface, it would have to have a component parallel to the surface. If that were true, free charges on the surface of the conductor would feel this field, and therefore a force (Eq. 2.3) along the surface. Under this force, they would subsequently flow along the surface, and once again, there is a net flow of charge, so we are by definition not in electrostatic equilibrium.

The fourth property is perhaps easiest to understand geometrically, as a consequence of the third property. The requirement that field lines be perpendicular to the surface forces them to “bunch up” wherever the radius of curvature is small, at “sharp” points, see Fig. 2.7. The presence of a sharp point with a high radius of curvature enhances the electric field in that region, and as a result, the mobile surface charges will instantly flow to this region of high curvature. They will do this until the electric field along the surface is cancelled. The sharper the point, the more charges need to flow into the region to ensure that the parallel component of the surface electric field is totally cancelled. This does result in an uneven surface charge density for irregularly shaped conductors, but also an electric field which is uniform and perfectly normal to the surface everywhere.

These rules might be easier to grasp pictorially.¹⁴ Figure 2.8 shows the field lines between oppositely charged conducting plates – an example of a device known as a *capacitor*, which we will study in Ch. 3. Note that the field in the region between the plates is very uniform, due to the requirement that it be perpendicular to the surface of the conductors. Near the edges of each plate, the field “fringes”, and starts to curve slightly outward. Further from the edges of the plates, the field starts to resemble that of a dipole (Fig. 2.5, turned 90°). This is no accident – the excess charges on the very edges of the plates *do* essentially form a dipole, so viewed from far away, the edges of this parallel plate structure look like a long row of dipoles stacked together. Microscopically, this is almost exactly what is happening!

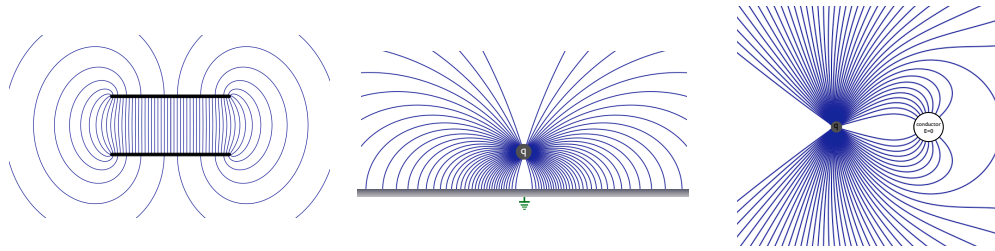


Fig. 2.8 (a) Field lines between two oppositely charged plates, (b) a point charge above a grounded conducting plane, and (c) a point charge near a conducting sphere. Field lines must be perpendicular to the surface of a conductor at every point, and their density increases near “sharp” points. Note also that there are no field lines inside the sphere, as the field inside a conductor must be zero.

Figure 2.8 also shows the field lines due to a point charge suspended above a grounded conducting plate. In this case, we again see that field lines *always* intersect the conducting surface at right angles. Again, this resembles Fig. 2.5 – this looks like half of the dipole field, as if there were a mirror halfway between the two charges. This is exactly what is happening – since the field lines have to intersect the plate at right angles, the point charge a distance d from the conducting plate behaves in the same way as if there were an equal and opposite charge a distance $2d$ away. Really, *a conductor is a mirror for electric field lines*! One can use this as a problem-solving trick, known as the “method of images.” This is a bit beyond the scope of the current text, but a neat time-saving trick to be aware of. What this also means qualitatively is that *when a charge is present near a conductor, the charge induces an equal and opposite charge spread out on the surface of the conductor*. In this case, a charge q above the conducting plate induces an *overall* charge $-q$ over the whole surface of the conducting plate.

Finally, Fig. 2.8 shows a point charge near a hollow conducting sphere. Note that *everywhere*, the field is perpendicular to the conducting sphere, and the field is zero inside the conductor. Oddly, this figure looks a bit like what we would expect if the conducting sphere were replaced by another charge, opposite in sign but smaller than the existing point charge. Can you see why this might be? As a hint, think about conductors being *mirrors* for field lines.

2.6 Faraday Cages

A “Faraday Cage” is an enclosed region formed by conducting material – essentially a hollow conductor. Since the electric field inside a conductor is zero, *anything* we enclose inside a hollow conductor will be *completely* shielded from any static electric fields.¹⁵ You can see Faraday Cages all around you, if you look carefully – electrical conduits inside the walls are metal boxes, the inside of your cell phone is surrounded by metal foil, and your computer hides inside a metal (or metal-lined) box.

2.4. All four properties are exemplified in Fig. 2.8, can you spot where?

¹⁴ Appendix ?? may provide an interesting read for the mathematically inclined.

¹⁵ Again, Appendix. ?? may be insightful.

Faraday cages are named for Michael Faraday (Fig. ??), who built one in 1836 and explained its operation.[6] Charges in the enclosing conducting shell repel one another, and will always reside on the outside surface of the cage (as discussed above). Any external (static) electrical field will cause the charges on the surface to rearrange until they completely cancel the field's effects in the cage's interior. No matter how large the field outside the cage, the field inside is *precisely zero*, so long as there is no charge inside the box. It seems incredible that the charges on the conductor's surface know just how to arrange themselves to exactly cancel the external field, but this is really what happens.

The most important application of Faraday cages is for this sort of electromagnetic shielding. One example is a shielded coaxial cable (*e.g.*, RCA cables for your stereo, or the coax connecting your cable or satellite box), which has a wire mesh shield surrounding an inner core conductor. The mesh shielding keeps any signal from the core from escaping, and perhaps more importantly, prevents spurious signals from reaching the core.

A more subtle example of a Faraday cage is probably sitting in your kitchen. The door of a microwave oven has a screen built into the glass of the window, with small holes in it. As we will find out in Sect. 8.5.5, this too is a Faraday cage, even though there are holes in the screen. Why does it still work, even though there are holes? How do electric fields relate to microwaves? Before too long, you will know!

2.7 The van de Graaff Generator

In 1929 Robert J. van de Graaff (1901-1967), a Tuscaloosa native and UA graduate (BS '22, MS '23), designed and built an electrostatic generator that has been extensively used in nuclear physics research. Dr. van de Graaff can be considered the inventor of the first accelerator providing intense particle beams of precisely controllable energy, and one of the pioneers of particle physics.[7]¹⁶

The principles of its operation can be understood using the properties of electric fields and charges you have (hopefully) just learned. Figure 2.9 shows the basic construction of Dr. van de Graaff's device, and Fig. 2.10 shows illustrations from Dr. van de Graaff's original patent on the "Electrostatic Generator" from 1931. A motor-driven pulley moves a belt past positively-charged metallic needles at position A. Negative charges are attracted to the needles from the belt, which leaves the left side of the belt with a net positive charge. The moving belt transfers these positive charges up toward the conducting dome.

The positive charges attract electrons on to the belt as it moves past a second set of needles at point B, which increases the excess positive charge on the dome. Because the electric field inside the conducting metal dome is negligible (it would be precisely zero if there were not holes in the dome), the positive charge on it can be easily increased – near zero electric field means near zero repulsive force to add more charge. The result is that extremely large amounts of positive charge can be deposited on the dome.

This charge accumulation cannot occur indefinitely. Eventually, the electric field due to the charges becomes large enough to ionize the surrounding air, increasing the air's conductivity. When sufficiently ionized, the air is nicely conducting, and the charges may rapidly flow off of the dome through the air – a "spark" jumps off of the dome to the nearest ground point. A spectacular example of this can be seen in Figure 2.11.

Since the "sparks" are really charge flowing off of the dome, this eventually limits the highest electric fields obtainable. The easy solution to increase the voltage is to make the domes bigger (decrease their radius of curvature), and put them higher off the ground (the farther a "spark" has to go, the more electric field it takes to create one).

One of the largest Van de Graaff generators in the world, built by Dr. Van de Graaff himself, is now on permanent display at Boston's Museum of Science (it is the one shown in Figure 2.11). It uses 15 foot aluminum spheres standing on columns many feet tall, and can reach 2 million volts. The Van de Graaff generator is operated several times per day in the museum's "Theater of Electricity."

¹⁶ You might think that is how the Tuscaloosa airport got its name. You would be wrong.[8]

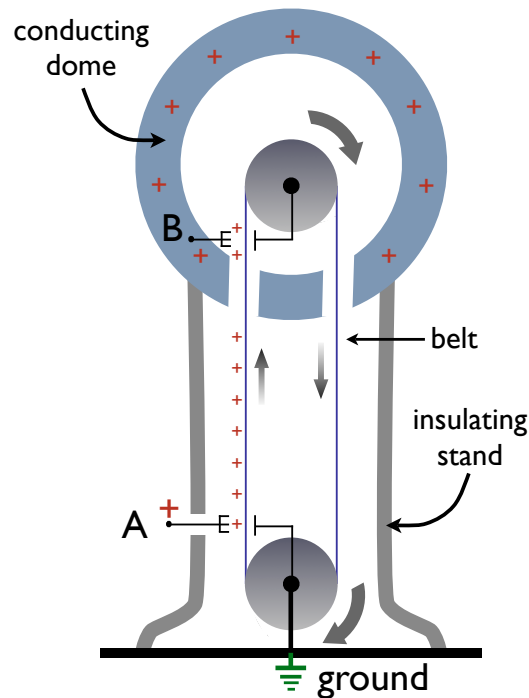


Fig. 2.9 A diagram of a van de Graaff generator. Charge is transferred to the dome by means of a rotating belt. The charge is deposited on the belt at point A and transferred to the dome at point B.

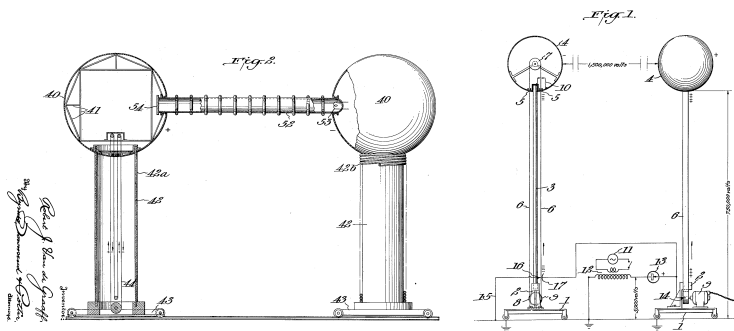


Fig. 2.10 Images from van de Graaff's original patent on the "Electrostatic Generator," filed 16 December, 1931.[9]

Van de Graaff information and pictures can be found through the Museum of Science:

<http://www.mos.org/sln/toe/toe.html>

More pictures of the largest van de Graaff generator, including its construction and historical pictures, can be found through the MIT Institute Archives:

<http://libraries.mit.edu/archives/exhibits/van-de-graaff/>

An interesting article from the Tuscaloosa News about Robert van de Graaff:

<http://bama.ua.edu/~jharrell/PH106-S06/vandegraaff.htm>

You can also visit his boyhood home at 1305 Greensboro Ave in Tuscaloosa.



Fig. 2.11 The world's largest van de Graaff generator originally built at Round Hill, near South Dartmouth, Massachusetts in 1933. [10] In the early 1950's, the giant Van de Graaff generator was donated to the Museum of Science in Cambridge, Massachusetts, where it is now operated at least twice daily for demonstrations. Do you know why the person in this picture is in no danger? Re-read Section 2.6 ... Photo credits: T.L. Carroll[11]

2.8 Gauss' Law

Gauss's law is a very sneaky technique (based on some basic theorems of vector calculus) for calculating the average electric field over a closed surface. What do we mean by a closed surface? A closed surface has an inside and an outside, it is one that encloses a volume and has no holes in it. A sphere and a cube are simple examples. If, due to symmetry, the electric field is constant everywhere on a closed surface, the exact electric field can be found – in most cases, much more easily than *via* Coulomb's law.

2.8.1 Electric Flux

How do we use this sneaky law? First, we need the concept of **electric flux**, denoted by Φ_E . Electric flux is a measure of how much the electric field vectors penetrate a given surface. If the electric field vectors are tangent to the surface at all points, they don't penetrate at all and the flux is zero. *Basically, we count the number of field lines penetrating the surface per unit area* – lines entering the inside of the surface are positive, those leaving to the outside are negative.

An analogy of electric flux is fluid flux, which is just the volume of liquid flowing through an area per second. The electric flux due to an electric field \vec{E} constant in magnitude in direction through a surface of area A is $\Phi = |\vec{E}|A \cos \theta_{EA}$, where θ_{EA} is the angle that \vec{E} makes with the surface *normal*.

Definition of electric flux through a surface

$$\Phi_E = |\vec{E}|A \cos \theta_{EA} \quad (2.5)$$

where θ_{EA} is the angle between the normal and the electric field.

Consider the surface in Figure 2.12a. The electric field is uniform in magnitude and direction. Field lines penetrate the surface of area A uniformly, and are perpendicular to the surface at every point ($\theta = 0^\circ$). The flux through this surface is just $\Phi = |\vec{E}|A$.

Now consider the surface A in Figure 2.12b. The uniform electric field penetrates the area A that is at an angle θ to the field, so now the flux is $\Phi_E = |\vec{E}|A \cos \theta$. For the surface A' , the field lines are perpendicular, but the area is reduced by the same amount, so the flux is the same through A and A' .

Just like electric forces and fields, flux also obeys the superposition principle. If we have a number of charges inside a closed surface, the total flux through that surface is just the sum of the fluxes from each individual charge.

Now: on to Gauss's law. What Gauss's law actually relates is the **electric flux** through a closed surface to the total electric charge contained inside that surface – **the electric flux through a closed surface is proportional to the charge contained inside the surface**. To see how this works, con-

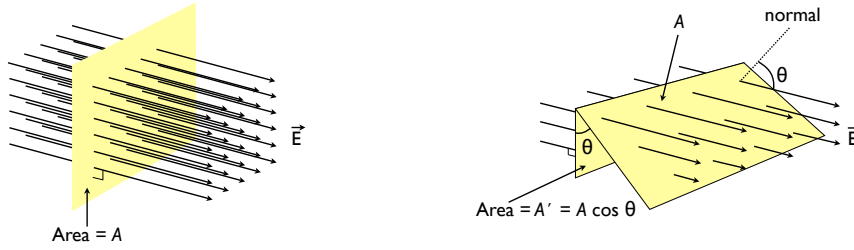


Fig. 2.12 (a) Field lines representing a uniform electric field E penetrating a plane of area A perpendicular to the field. The electric flux Φ_E through this area is equal to $|\vec{E}|A$. (b) Field lines representing a uniform electric field penetrating an area A that is at an angle θ to the field. Because the number of lines that go through the area A' is the same as the number that go through A , the flux through A' is given by $\Phi_E = |\vec{E}|A \cos \theta$.

sider the point charge in Figure 2.13a. The innermost surface is just a sphere, whose radius we will call r . The strength of the electric field everywhere on this sphere is

$$|\vec{E}| = k_e \frac{q}{r^2} \quad (2.6)$$

since every point on the sphere's surface is a distance r from the charge. We also know that \vec{E} is perpendicular to the surface everywhere, thanks to the radial symmetry. Finally, we know that the surface area of a sphere is $A = 4\pi r^2$, so the electric flux is

$$\Phi_E = |\vec{E}|A = k_e \frac{q}{r^2} (4\pi r^2) = 4\pi k_e q \quad (2.7)$$

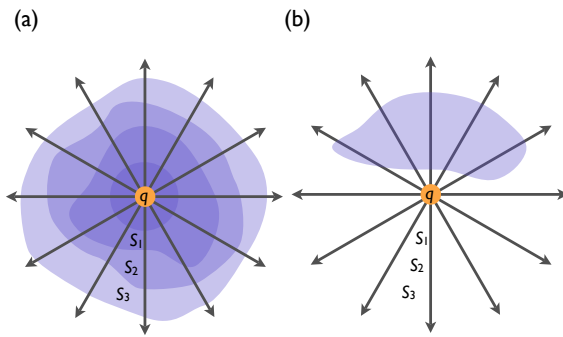


Fig. 2.13 (a) Closed surfaces of various shapes surrounding a point charge q . The net electric flux is the same through all surfaces. (b) If the point charge is *outside* the surface, the net flux is zero through that surface since the same number of field lines enter and leave. If no charge is *enclosed* by the surface, there is no net flux.

If the point charge is *outside* the surface, Fig. 2.13b, the net flux is zero through that surface since the same number of field lines enter and leave. If no charge is *enclosed* by the surface, there is no net flux.

Now the power in Gauss's law is that if we take *any* arbitrarily more complicated surface, so long as it surrounds the point charge q and doesn't have holes in it, we will *always get the same flux!* What this means is that we always choose very convenient surfaces, ones for which the electric field is just a constant over the whole surface. For convenience, we define a new constant $\epsilon_0 = 1/4\pi k_e$, known as the "permittivity of free space:"

Permittivity of free space:

$$\epsilon_0 = \frac{1}{4\pi k_e} = 8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N} \cdot \text{m}^2} \quad (2.8)$$

Recall k_e is Coulomb's constant from Equation 2.2. (This means of course that we can put all of our other equations, like Eq. 2.1, in terms of ϵ_0 instead of k_e , since $k_e = 1/4\pi\epsilon_0$. You will often see them this way.) This gives Gauss's law a nice simple form:

Gauss's law: the electric flux Φ_E through any *closed* surface is equal to the net charge inside the surface, Q_{inside} , divided by ϵ_0 :

$$\Phi_E = \frac{Q_{\text{inside}}}{\epsilon_0} \quad (2.9)$$

We will not *derive* Gauss' law here, but simply state it as fact, and show you a few examples of how to use it.

2.8.2 Gauss' Law as a Conservation Law

Fundamentally, Gauss' law is a manifestations of the *divergence theorem* (a.k.a. Green's theorem or the Gauss-Ostrogradsky theorem). Essentially, it states that *the sum of all sources minus the sum of all sinks gives the net flow out of a region*. The same law applies to fluids. If a fluid is flowing, and we want to know how much fluid flows out of a certain region, then we need to add up the sources inside the region and subtract the sinks. The divergence theorem is basically a conservation law - the volumetric total of all sources minus sinks equals the flow across a volume's boundary.

In the case of electric fields, this gives Gauss' law (Eq. 2.9) – the electric flux through any closed surface must relate to a *net charge* inside the volume bounded by that surface. The net magnitude of the vector components of the electric field pointing outward from a surface must be equal to the net magnitude of the vector components pointing inward, *plus* the amount of free charge inside. This is a manifestation of the fact that electric field lines do have to originate from somewhere – charges.

The difference between the flow of field lines into a surface and the flow out of a surface is just how many charges are within the surface, that is all that Gauss' law says. This is fundamentally due to the fact that for *all* inverse square laws, like Coulomb's law or Newton's law of gravitation, the strength of the field falls off as $1/r^2$, but the area of an enclosing surface *increases* as r^2 . The two dependencies cancel out, and we are left with the result that the flux is only related to difference between the number of enclosed sources and sinks.

Though Gauss's law is very powerful, it is usually used in specially symmetric cases (spheres, cylinders, planes) where it is easy to draw a surface of constant electric field around the charges of interest (like a sphere around a point charge). We will work through a few of these examples presently.

2.5. Why would we not want to choose a cube as our surface enclosing the point charge?

2.8.3 Example: The Field Around a Spherical Charge Distribution

We can use Gauss' law to calculate the electric field of *any* spherically symmetric distribution of charge, and as a bonus, discover an important fact. A spherically symmetric distribution of charge just means that the number of charges per unit volume (the charge *density*) depends only on the radius from a central point. That doesn't mean that the density doesn't vary with radius, just that it

doesn't vary with *angle*. An example of such a distribution is shown in Fig. 2.14a – in this case, the density decreases with radius up to a distance R , beyond which it is zero.

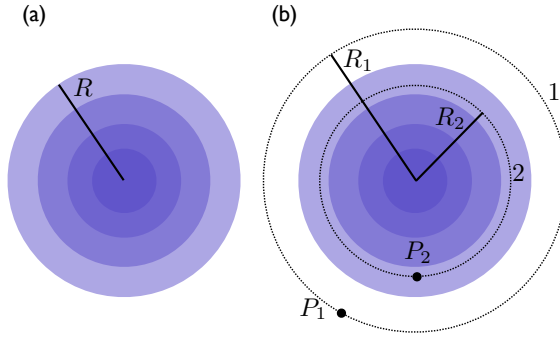


Fig. 2.14 (a) A spherically symmetric charge distribution. The density of charge depends only on the distance from the center point. (b) Two Gaussian surfaces to determine the field at an arbitrary point outside (P_1) and inside (P_2) the distribution.

What is the electric field at some arbitrary point P_1 outside the distribution, or at some arbitrary point P_2 inside it (Fig. 2.14b)? Do we have to calculate the field from every tiny bit of charge in the distribution and sum them all together? No, this is the point of Gauss' law – if you have a problem with special symmetries, they can usually be exploited to save a lot of labor.

The charge distribution is, by definition, spherically symmetric. As you may have noticed, the electric field must take on the same symmetry as the charge distribution.¹⁷ That means that the electric field in this case will be spherically symmetric, and will be directed radially from the central point. No other direction is special or unique in this problem, only the radial direction. That means that if we draw a spherical surface of radius $R_1 > R$ completely surrounding the sphere, surface 1 in Fig. 2.14b, the electric field will be constant everywhere on that surface. We can easily calculate the flux through this surface, and hence the electric field:

$$\Phi_E = EA = \frac{Q_{\text{inside1}}}{\epsilon_0} \quad R_1 > R \quad (2.10)$$

$$= E \times 4\pi R_1^2 = \frac{Q_{\text{inside1}}}{\epsilon_0} \quad R_1 > R \quad (2.11)$$

$$\Rightarrow E = \frac{Q_{\text{inside1}}}{4\pi\epsilon_0 R_1^2} = \frac{k_e Q_{\text{inside1}}}{R_1^2} \quad R_1 > R \quad (2.12)$$

What we now see is that this is the same thing as the field from a point charge – *the field outside a spherically symmetric charge distribution behaves exactly as if all of its charge is concentrated at the center*. This is, in fact, a particular property of $1/r^2$ laws, and you should recall that this principle is true in the gravitational case for spherically symmetric mass distributions. The earth attracts other bodies as if its mass were concentrated at a point in the center. So long as we are dealing with spherically symmetric distributions, it is not even an approximation to deal with infinitesimal point charges!

One thing to keep in mind: this is *not* something like the center of mass. A perfect cube does *not* behave as if it had all its mass concentrated at its center. This all really comes from the nature of $1/r^2$ forces and the divergence theorem.

What about surface 2, radius R_2 , drawn inside the charge distribution? From the analysis above, all that matters is how much charge is contained inside the surface. Everything outside the surface contributes an equal contribution, but in all different directions, and the whole thing cancels. What is outside the surface may just as well not exist, so far as the electric field is concerned. Finding the field at point P_2 is then just a matter of figuring out how much charge is inside the second surface.

¹⁷ Appendix ?? may help you think about that.

Depending on the distribution, that may not be so easy ... but it would have been a *lot* worse without Gauss' law.

We have actually developed a more important result than we set out to. Using only Gauss' law, we have *derived* that the field from spherically symmetric charge distributions is equivalent to that of a point charge, and follows a $1/r^2$ law. *Actually, we have derived Coulomb's law from Gauss' law. In fact, the two are equivalent.* We could have started from Gauss' law in the first place and arrived at Coulomb's law, instead of *assuming* Coulomb's law to be true and *then* introducing Gauss' law. Gauss' law is in fact far more general in an important way, as we have noted above, since it gives the equivalence relationship for *any* flux (*e.g.*, liquids, electric fields, gravitational fields) flowing out of any closed surface and the enclosed sources and sinks of the flux (*e.g.*, electric charges, masses). We will see in Ch. 6.2.1 that there is also a Gauss' law for magnetism, just as there is a Gauss' law for gravity, *viz.*:

$$\Phi_g = 4\pi GM \quad (2.13)$$

where Φ_g is the flux from the gravitational field through a closed surface, G is the universal gravitational constant, and M is the mass enclosed by the surface. Just as we proved that any spherically symmetric charge distribution behaves as a point charge and follows an inverse square law, one can prove that any spherically symmetric mass distribution is equivalent to a point mass, and follows the familiar inverse square law for gravitation.

2.6. A charge of $100\mu\text{C}$ is at the center of a cube of side 0.8 m. **(a)** Find the total flux through each face of the cube. **(b)** Find the flux through the whole surface of the cube. **(c)** Would your answers to the first two parts change if the charge were not at the center of the cube?

2.8.4 Example: The Field Above a Flat Conductor

If we can come up with a clever surface on which to apply Gauss' law, we can solve some otherwise nasty problems. Figure 2.15 shows a large ("infinite") conducting plate, whose surface is charged. What is the field at the surface of this plate due to the charges? We know it is uniform and constant, but that is about it.

Since it is a conductor, the charge distribution on the surface, and hence electric field, are uniform. Since we do not want to restrict ourselves to a plate of any particular size, but rather, solve a *general* problem, we will say that the plate has a certain charge per unit area σ_E , defined as the total charge of the plate divided by its surface area. That way, we can later find the field near *any* plate.

What sort of surface should we take to find the flux? A plain box is a good choice, as it turns out, due to the symmetry of the problem. We will take a box with a top and bottom whose area are A . The area of the sides are not important, as it turns out, but we can call them B just to be complete.

Why would we choose a box in this case, when we just said it is a bad choice for a point charge? We know that the field is perpendicular to the surface of a conductor everywhere, so in this case the field is going to be purely perpendicular to the plate. Therefore, it is only important that we draw a Gaussian surface such that every part of the surface is either perfectly parallel or perfectly perpendicular to the plate. A cylinder would work perfectly fine too, which should be clear from the rest of the discussion.

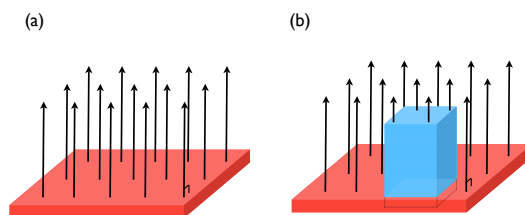


Fig. 2.15 **(a)** A large, flat charged conducting slab. The charge distribution on the surface, and hence electric field, are uniform. **(b)** A cylinder is our surface for Gauss' law. Along the sides of the cylinder, the flux is zero since the field lines are parallel – the flux is non-zero only through the end caps.

Along the surface making up the sides of the box, the flux is zero since the field lines are parallel to it everywhere. On the top end cap, the flux is perfectly normal. The bottom end cap is completely inside the conductor, so we know the field has to be zero there! If we call the magnitude of electric field above the plate E , we can readily calculate the flux. Because the plate is supposed to be very large in extent, the field can be assumed to be completely uniform so long as the distances above the plate we consider are small compared to the size of the plate.

The total charge enclosed by this cylinder is just the cross-sectional area of the plate enclosed by the box times the charge per unit area σ_E :

$$Q_{\text{inside}} = \sigma_E A \quad (2.14)$$

Applying Gauss' law is now straightforward, we just have to find the flux through the top end cap:

$$\Phi_E = EA = \frac{Q_{\text{inside}}}{\epsilon_0} = \frac{\sigma_E A}{\epsilon_0} \quad (2.15)$$

$$\Rightarrow E = \frac{\sigma_E}{\epsilon_0} \quad (2.16)$$

No problem! The electric field is indeed constant, as it has to be, and *independent of the distance from the plate*. This makes sense too, since the plate is supposed to be very, very large. Strictly, this is true only for an infinite plate, but it is close so long as we consider distances above the plate which are very small compared to the size of the plate. Finally, it should be clear now that it didn't matter what sort of shape we used at all, so long as it has a flat end parallel to the plate, and sides perpendicular to it.

2.8.5 Example: The Field Inside and Outside a Hollow Spherical Conductor

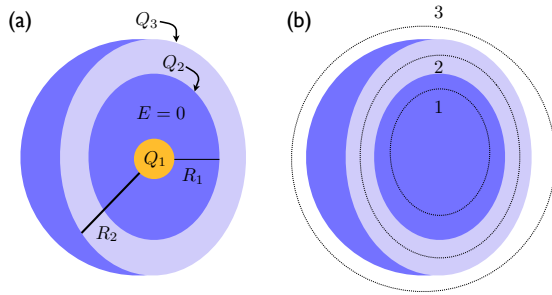


Fig. 2.16 (a) A point charge Q_1 is inside a thin spherical conducting shell with inner radius R_1 and outer radius R_2 . The presence of the point charge induces an equal but opposite charge on the inner surface of the conductor to satisfy Gauss' law. (b) Three Gaussian surfaces to find the field inside and outside the conductor.

Figure 2.16 shows a point charge Q_1 is inside a thin spherical conducting shell with inner radius R_1 and outer radius R_2 . How can we find the electric field inside the shell and outside the shell? Easy, we just have to apply Gauss' law a couple of times.

For any spherical surface *inside* the sphere, say a sphere of radius $r < R_1$ like surface 1 in Fig. 2.16, only the point charge is inside the volume enclosed by the sphere. If we center the sphere exactly on the point charge, since the field of a point charge is spherically symmetric the field is constant everywhere on the sphere's surface. Gauss' law then gives us:

$$\Phi_E = EA = \frac{Q_{\text{inside}}}{\epsilon_0} \quad r < R_1 \quad (2.17)$$

$$= E \times 4\pi r^2 = \frac{Q_1}{\epsilon_0} \quad r < R_1 \quad (2.18)$$

Now we just need to solve for E , and make use of the fact that $\epsilon_0 = \frac{1}{4\pi k_e}$ (Eq. 2.8):

$$E = \frac{Q_1}{4\pi\epsilon_0 r^2} = \frac{k_e Q_1}{r^2} \quad r < R_1 \quad (2.19)$$

Of course, this makes perfect sense – the field inside is just that of the point charge, as if the conductor were not there at all! As we saw above, electric fields are like gravitational fields in this way – inside a spherical shell, both the gravitational and electrical forces cancel in all directions by symmetry.

Next, we consider surface 2, a surface inside the conductor itself. We know already that everywhere *inside* the conductor, *i.e.*, $R_1 < r < R_2$, we must have $E = 0$. Done! That seemed too easy, didn't it? It was – we missed one little point.

In the end, we also want to find the field *outside* the shell entirely, and for this we have to consider surface 1. Now we have to be careful, and think about what we have missed. For surface 2, drawn inside the conductor, we said $E = 0$ as it has to be for a conductor. This is true. But how can that be, with a point charge sitting right inside? Actually, it can't: what happens is that the point charge Q_1 induces a *equal but opposite* charge $Q_2 = -Q_1$ on the inside surface of the conductor. Think of it this way – if this did *not* happen, then the total charge enclosed by surface 2 would not be zero, and by Gauss' law the field inside the conductor could not be zero. The induced charge Q_2 ensures that the total charge enclosed by surface 2 is zero, and thus the field inside the conductor is zero as it has to be. Then we would have a contradiction on our hands, which is not OK. This also is another aspect of conductors looking like mirrors for field lines. Physically, the charge Q_1 attracts opposite mobile charges in the conductor, giving a net negative charge on the inner surface.

Now, what about surface 3? Before we placed the charge Q_1 inside the conductor, it was electrically neutral. This still has to be true after we place the charge – overall, the conductor must have no net charge. Well, if there is a charge $Q_2 = -Q_1$ on the inner surface, and overall it is neutral, then there must be a charge $Q_3 = Q_1$ induced on the *outer* surface to cancel the induced charge on the inner surface. The net negative charge on the inner surface attracted by the point charge Q_1 leaves a deficit of negative charge on the outer surface, for a net positive surface charge. Now we can run Gauss' law for surface 3:

$$\Phi_E = EA = \frac{Q_{\text{inside}}}{\epsilon_0} \quad r > R_2 \quad (2.20)$$

$$E \times 4\pi r^2 = \frac{Q_1 + Q_2 + Q_3}{\epsilon_0} \quad r > R_2 \quad (2.21)$$

$$4\pi r^2 E = \frac{Q_1 - Q_1 + Q_1}{\epsilon_0} \quad r > R_2 \quad (2.22)$$

$$\implies E = \frac{Q_1}{4\pi\epsilon_0 r^2} \quad r > R_2 \quad (2.23)$$

Lo and behold, the field outside the sphere looks just like that of the original point charge, same as inside the sphere (remembering that $\epsilon_0 = \frac{1}{4\pi k_e}$, Eq. 2.8). Again, what happens physically is that the point charge pulls the mobile charges from the conductor to its inner surface, leaving the inner surface with an equal and opposite charge. This means that the outer surface must be deficient in those same charges, and thus has an equal and like charge to Q_1 .

Now we can combine our results, and we have the electric field in all three regions:

$$E = \frac{k_e Q_1}{r^2} \quad r > R_2 \quad (2.24)$$

$$E = 0 \quad R_1 < r < R_2 \quad (2.25)$$

$$E = \frac{k_e Q_1}{r^2} \quad r < R_1 \quad (2.26)$$

2.8.6 Example: The Field Due to a Line of Charge

As one last example, we will use Gauss' law to find the electric field due to an infinite line of charge, or equivalently, a conducting wire with a net surface charge, as shown in Fig. 2.17a. What does the electric field look like? If the line of charge is infinite (or at least very long compared to the distance we are away from it), all of the transverse components of the field will cancel each other, and by symmetry, the field must be radially symmetric about the wire. That is, the field must point perpendicularly away from the wire axis.

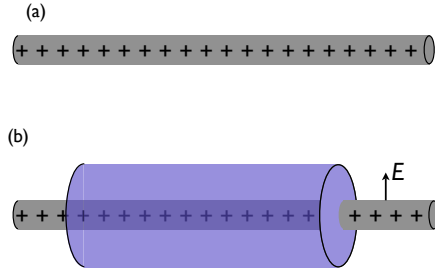


Fig. 2.17 (a) An “infinite” line charge, with λ charges per unit length. (b) A cylindrical Gaussian surface. On the caps of the cylinder, the field is parallel, and the flux is zero.

With the symmetry of the wire being cylindrical, it makes most sense to use a cylinder drawn concentrically around the wire as our Gaussian surface, Fig. 2.17b. We will choose a cylinder of radius r , and length l . The field is parallel to the end caps of the cylinder, so they contribute no flux at all. Being radially symmetric, the field is perfectly *perpendicular* to the round surface of the cylinder, and we can easily calculate the flux and find the electric field. First, we remember that the surface area of a cylinder (without the end caps) is $2\pi rl$. Second, the cylinder of length l encloses a length l of the wire, which must contain λl charges since λ is the charge per unit length. Putting that all together:

$$\Phi_E = E \cdot 2\pi rl = \frac{Q_{\text{encl}}}{\epsilon_0} = \frac{\lambda l}{\epsilon_0} \quad (2.27)$$

$$\Rightarrow E = \frac{\lambda}{2\pi r \epsilon_0} = \frac{2k_e \lambda}{r} \quad (2.28)$$

In this case, the field falls off as $1/r$, far slower than a point charge, but *not independent of distance* like we found for the sheet of charge. It *is* independent of the length of the cylinder we chose, as it *must* be: the wire is supposed to be infinite, and the value of l was chosen arbitrarily! File this result away. We will need it again in Chapter 6!

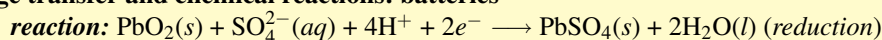
2.9 Miscellanea

Solving electric field problems

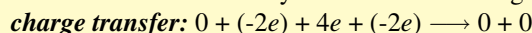
1. **Convert all units** to SI – charges in Coulombs, distances in meters.
2. **Draw** a diagram of the charges in the problem.
3. **Identify** the charge of interest, and what you want to know about it.
4. **Choose** your coördinate system and origin – pick the most convenient one based on the symmetry of the problem. Usually, this is an x – y Cartesian system, with the origin at some special point (*e.g.*, on one charge or between two charges)
5. **Apply Coulomb's law** For each charge Q , find the electric force on the charge of interest q . The vector direction of the force is along the line of the two charges, directed away from Q if it has the same sign as q and toward Q if it has the opposite sign as q . Find the angle θ this vector makes with the positive x axis – the x component of the electric force will be $F \cos \theta$, the y component will be $F \sin \theta$.
6. **Sum the x components** from each charge Q to get the resultant x component of the electric force.
7. **Sum the y components** from each charge Q to get the resultant y component of the electric force.
8. **Find the total resultant force** from the total x and y components, using the Pythagorean theorem and trigonometry to find the magnitude and direction:

$$|F_{\text{tot}}| = \sqrt{|F_x|^2 + |F_y|^2} \text{ and } \tan \theta = \frac{F_y}{F_x}$$

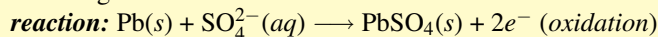
Charge transfer and chemical reactions: batteries



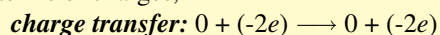
From one point of view, this reaction is nothing more than charges being transferred from one species to another. If we only write down the charges involved, we would have this:



This is consistent with only electrons being transferred from one object to another – four electrons are being transferred to the four H^+ , including two from the SO_4^{2-} and two ‘free’ electrons. Charge is conserved in this reaction as well. Another example:



In terms of charges,



The above reactions are essentially what take place in a normal lead-acid car battery. Plates of lead (Pb) and lead oxide (PbO_2) immersed in a sulfuric acid (H_2SO_4) electrolyte. The Pb plate is oxidized, releasing two electrons per Pb atom, while the PbO_2 plate is reduced, accepting two electrons per molecule. Connecting the two plates together through a circuit lets electrons released from the Pb plate travel to the PbO_2 plate, which makes an electric current.

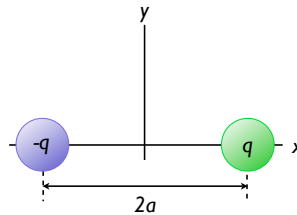
2.10 Problems

Solutions begin on page 246.

2.7. If two charges of $+1\ \mu\text{C}$ are separated by $1\ \text{cm}$ ($= 10^{-2}\ \text{m}$), what is the force between them?

2.8. A proton accelerates from rest in a uniform electric field of $800\ \text{N/C}$. At some time later, its speed is $1.2 \times 10^6\ \text{m/s}$. What is the magnitude of the acceleration on the proton?

2.9. Two point charges q and $-q$ are situated along the x axis a distance $2a$ apart as shown below. Show that the electric field at a distant point along $|x| > a$ along the x axis is $E_x = 4k_e qa/x^3$.



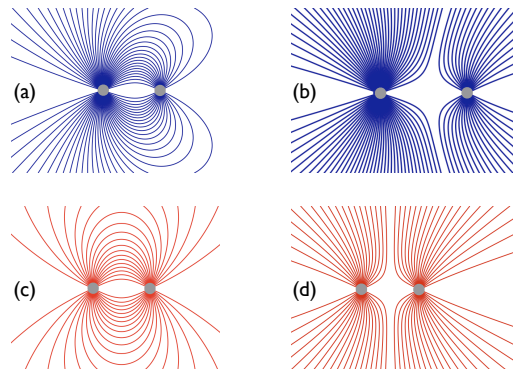
2.10. Two charges of $+1\ \mu\text{C}$ are separated by $1\ \text{cm}$. What is the magnitude of the electric field halfway between them?

2.11. A circular ring of charge of radius b has a total charge of q uniformly distributed around it. The magnitude of the electric field at the center of the ring is:

2.12. Two isolated identical conducting spheres have a charge of q and $-3q$, respectively. They are connected by a conducting wire, and after equilibrium is reached, the wire is removed (such that both spheres are again isolated). What is the charge on each sphere?

2.13. A single point charge $+q$ is placed exactly at the center of a hollow conducting sphere of radius R . Before placing the point charge, the conducting sphere had zero net charge. What is the magnitude of the electric field *outside* the conducting sphere at a distance r from the center of the conducting sphere? *I.e., the electric field for $r > R$.*

2.14. Which set of electric field lines below could represent the electric field near two charges of the *same sign*, but *different magnitudes*?



2.15. Referring again to the figure above, which set of electric field lines could represent the electric field near two charges of *opposite sign* and *different magnitudes*?

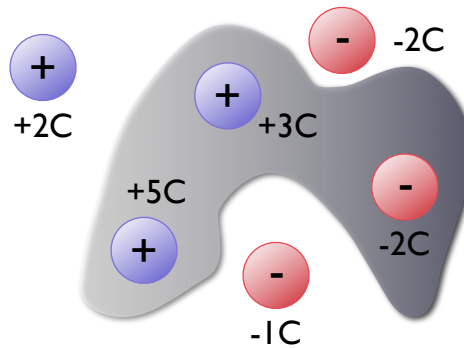
2.16. A “free” electron and a “free” proton are placed in an identical electric field. Which of the following statements are true? *Check all that apply.*

- Each particle is acted on by the same electric force and has the same acceleration.
- The electric force on the proton is greater in magnitude than the force on the electron, but in the opposite direction.
- The electric force on the proton is equal in magnitude to the force on the electron, but in the opposite direction.
- The magnitude of the acceleration of the electron is greater than that of the proton.
- Both particles have the same acceleration.

2.17. A point charge q is located at the center of a (non-conducting) spherical shell of radius a that has a charge $-q$ uniformly distributed on its surface. What is the electric field *for all points outside the spherical shell*?

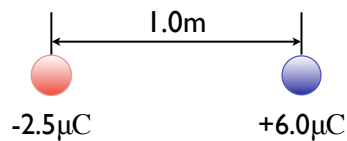
2.18. Referring to the previous problem, what is the electric field *inside* the same shell a distance $r < a$ from the center (*i.e.*, a point inside the spherical shell)?

2.19. What is the electric flux through the surface below, in terms of Coulombs per ϵ_0 ?



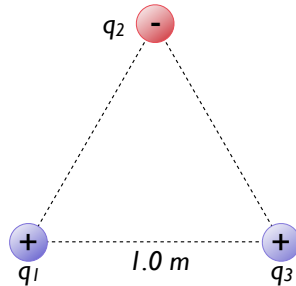
2.20. A spherical conducting object A with a charge of $+Q$ is lowered through a hole into a metal (conducting) container B that is initially uncharged (and is not grounded). When A is at the center of B , but not touching it, the charge on the inner surface of B is ...

2.21. Determine the point (other than infinity) at which the total electric field is zero and which is not between the two charges.



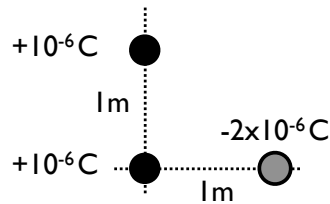
2.22. A flat surface having an area of 3.2 m^2 is rotated in a uniform electric field of magnitude $E = 5.7 \times 10^5\text{ N/C}$. What is the electric flux when the electric field is parallel to the surface?

2.23. Three charges are arranged in an equilateral triangle, as shown below. All three charges have the same *magnitude* of charge, $|q_1| = |q_2| = |q_3| = 10^{-9}\text{ C}$ (note that q_2 is negative though). What is the **force** on q_2 , magnitude and direction?

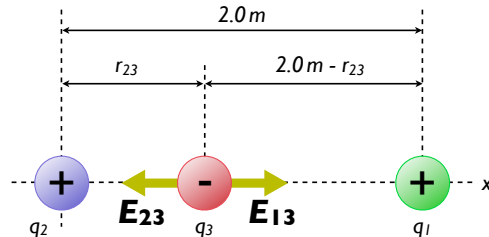


2.24. Suppose three positively charged particles are constrained to move on a fixed circular track. If all the charges were equal, an equilibrium arrangement would obviously be a symmetrical one with the particles spaced 120° apart around the circle. Suppose two of the charges have equal charge q , and the equilibrium arrangement is such that these two charges are 90° apart rather than 120° . What must be the relative magnitude of the third charge?

2.25. Two charges of $+10^{-6}\text{ C}$ are separated by 1 m along the vertical axis. What is the net **horizontal** force on a charge of $-2 \times 10^{-6}\text{ C}$ placed one meter to the right of the lower charge?



2.26. Three point charges lie along the x axis, as shown below. A positive charge $q_1 = 15\text{ }\mu\text{C}$ is at $x = 2\text{ m}$, and a positive charge of $q_2 = 6\text{ }\mu\text{C}$ is at the origin. Where must a *negative* charge q_3 be placed on the x -axis **between the two positive charges** such that the resulting electric force on it is zero?



2.27. Two solid spheres, both of radius R , carry identical total charges, Q . One sphere is a good conductor while the other is an insulator. If the charge on the insulating sphere is uniformly distributed throughout its interior volume, how do the electric fields outside these two spheres compare? Are the fields identical inside the two spheres?

Chapter 3

Electrical Energy and Capacitance

The principle of science, the definition, almost, is the following: The test of all knowledge is experiment. Experiment is the sole judge of scientific “truth.” But what is the source of knowledge? Where do the laws that are to be tested come from? Experiment, itself, helps to produce these laws, in the sense that it gives us hints. But also needed is imagination to create from these hints the great generalizations – to guess at the wonderful, simple, but very strange patterns beneath them all, and then to experiment to check again whether we have made the right guess. – Richard Feynman

Abstract Potential energy and the principle of conservation of energy often let us solve difficult problems without dealing with the forces involved directly. More to the point, using an energy-based approach to problem solving let us work with *scalars* instead of *vectors*. This way we get to deal with just plain numbers, which is nice.

In this chapter, we will learn that, as with the gravitational field, the electric field has an associated **potential** and **potential energy**. The electric potential will, in many cases, let us solve problems more easily than with the electric field and, as it turns out, electric potential is what we normally identify with ‘voltage’ in everyday life.

3.1 Electrical Potential Energy

The work done on an object by a conservative force, such as the electric force, depends only on the initial and final positions of the object, *not on the path taken between initial and final states*. For example, the work done by gravity depends only on the change in height. When a force is conservative, it means that there exists a **potential energy** function, PE , which gives the potential energy of an object subject to this conservative force which depends *only on the object’s position*. *Potential energy* is sometimes called the “energy of configuration” since it only depends on the position of objects in a system. Thus, for the conservative electric force, we can find a change in electrical potential energy just by knowing the starting and final configurations of the system we are studying – nothing in between matters.

As you know, potential energy is a scalar quantity, and the *change* in potential energy is equal to the work done by a conservative force.

Potential energy difference, ΔPE

$$\Delta PE = PE_f - PE_i = -W_F \quad (3.1)$$

where the subscripts $f(i)$ refer to the final (initial position), and W_F is the work done by the conservative force \vec{F} .

This is just how you dealt with gravity – moving an object of mass m through a vertical displacement h gives a changes in potential energy $\Delta PE = mgh$. Electrical forces and gravitational forces have a number of useful similarities, as you now know, and the same is true for their respective potential energies.

Consider a small positive test charge q in a uniform electric field \vec{E} , as shown in Figure 3.1. As the charge moves from point A to point B , covering a displacement $\Delta x = x_f - x_i$, the work done on the charge by the electric field is the component of the force $\vec{F}_e = q\vec{E}$ parallel to the displacement Δx :

¹Work done moving a charge q in a constant electric field \vec{E} :

$$\Delta W_{AB} = \vec{F} \cdot \Delta \vec{x} = |\vec{F}| |\Delta x| \cos \theta = qE_x (x_f - x_i) = qE_x \Delta x \quad (3.2)$$

where q is the charge, E_x is the component of the electric field \vec{E} along the direction of displacement, and θ is the angle between the force \vec{F} and the displacement $\Delta \vec{x}$ (of length Δx).

Note that q , E_x , and Δx can all be either positive or negative. Also recall that E_x is the x -component of the electric field \vec{E} , *not* the magnitude! Equation 3.2 is valid for the work done on a charge by *any constant electric field*, no matter what the direction of the field, or sign of the charge. Just remember that the angle between the field and displacement does matter!

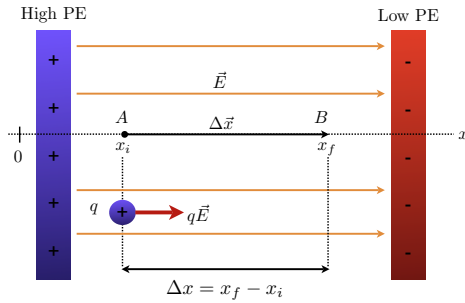


Fig. 3.1 When a charge q moves in a uniform electric field \vec{E} from point A to point B , covering a distance Δx , the work done on the charge by the electric force is $qE_x \Delta x$.

Now that we have found the work done by the electric field, the work-energy theorem gives us the potential energy change:

The change in electric potential energy ΔPE of an object with charge q moving through a displacement Δx in a constant electric field \vec{E} is:

$$\Delta PE = -W_{AB} = -q|\vec{E}| |\Delta \vec{x}| \cos \theta = -qE_x \Delta x \quad (3.3)$$

where the quantities are defined as in Eq. 3.2.

Remember, just like any other work, the work done involving the electric force only counts the displacement *parallel to the force*. You can find the component of the field parallel to the full displacement, or find the component of the displacement parallel to the field – it is the same thing. Figure 3.2 compares a charge moving in an electric field to a mass moving in a gravitational field. A positive charge moving in an electric field acts much like a mass moving in a gravitational field: the positive charge at point A falls in the direction of the field, just as the mass does. This lowers its potential energy, and increases its kinetic energy.

Assuming other forces are absent, we can also find the kinetic energy change through conservation of energy. Since both the electrical and gravitational forces are conservative, we can find the changes in kinetic and potential energy in both cases and compare them. In both situations, the change in potential energy must be equal and opposite the change in kinetic energy for energy to be conserved²:

¹ At this point you may want to remind yourself about the scalar or “dot” product, $\vec{A} \cdot \vec{B} = |\vec{A}| |\vec{B}| \cos \theta_{AB}$, where θ_{AB} is the angle between \vec{A} and \vec{B} .

² The subscripts i and f refer to initial and final, as usual.

$$KE_i + PE_i = KE_f + PE_f \quad (3.4)$$

$$(KE_f - KE_i) = -(PE_f - PE_i) \quad (3.5)$$

$$\Delta KE = -\Delta PE \quad (3.6)$$

$$\Delta KE + \Delta PE = 0 \quad (3.7)$$

For the gravitational case, we have done this a million times for an object of mass m starting at a height d and ending at a height defined as 0:

$$\Delta KE + \Delta PE_G = \Delta KE + (0 - mgd) = 0 \quad (3.8)$$

$$\implies \Delta KE = mgd \quad (3.9)$$

For the electrical case, it is not much more difficult. We will move a charge q through an electric field E :

$$\Delta KE + \Delta PE_E = \Delta KE + (0 - qE_d d) = 0 \quad (3.10)$$

$$\implies \Delta KE = qE_d d \quad (3.11)$$

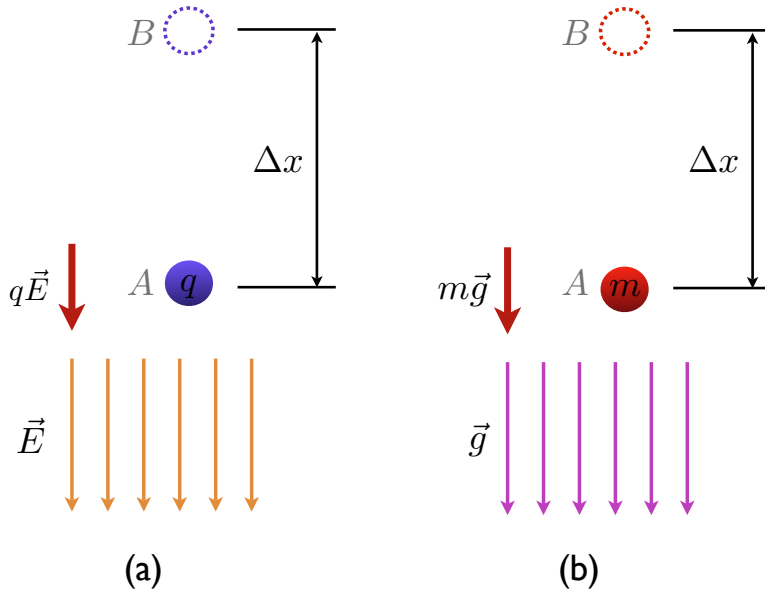


Fig. 3.2 (a) When an electric field \vec{E} is directed downward, point B has a lower electrical potential energy than point A . As a positive test charge moves from A to B , the electrical potential energy decreases. (b) An object of mass m moves in the direction of the gravitational field \vec{g} , the gravitational potential energy decreases.

Here d is the distance moved in the electric field \vec{E} , and E_d is the component of the electric field parallel to the direction of motion. For positive charges, electric potential energy works just like gravitational potential energy. Since mass comes only in one flavor, while charge comes in positive and negative varieties, this is not the whole story, however. For a negative charge, we have to substitute $-q$ for q in the equations above - rather than falling in the electric field like the positive charge, the negative charge wants to move *upward*. In other words, the negative charge “falls up” compared to a positive charge.

In order to make a negative charge move downward we would have to do work against the electric field. Remember that positive charges like to follow the direction of the electric field lines, while

negative charges like to go against them. For the positive charge in Figure 3.2, we are moving the charge in the direction it wants to go. For a negative charge in the same situation, we are moving the charge *against* the direction it wants to go. The negative charge has a *positive* change in electrical potential energy moving from **point A to point B**, meaning kinetic energy has to be lost to make this happen. The positive charge has a *negative* change in potential energy moving from **point A to point B**, meaning kinetic energy will be gained by doing this.

3.2 Electric Potential

In Chapter 2, it was convenient to define \vec{E} related to the electric force, viz., $\vec{F} = q\vec{E}$. This let us think about individual charges one at a time, even when our system was a collection of several charges, and discard the idea of “action at a distance.” For the same reasons, we would like to define a variation of the electrical potential energy *per unit charge*, so we may think about *how much potential energy would be gained or lost by a single charge present in an electric field*.³ This quantity is the **electric potential difference** ΔV , and it is related to potential energy by $\Delta PE = q\Delta V$.⁴

The electric potential difference ΔV between points A and B is the change in electric potential energy between those two points divided by the quantity of charge moving Q :

$$\Delta V = V_B - V_A = \frac{\Delta PE}{q} \quad \text{or} \quad q\Delta V = \Delta PE \quad (3.12)$$

where V_B is the potential at point B and V_A is the potential at point A.

Electric potential is measured in Joules per Coulomb, otherwise known as *Volts*. In fact, we will often refer to electric potential as “voltage,” the two are synonymous for our purposes. Just like gravitational potential, **electric potential is a scalar quantity**. It is essentially a measure of the change in electric potential energy per unit charge. By definition, it takes 1 J to move 1 C worth of charge between two points with a potential difference of 1 V. If a 1 C charge moves through a potential difference of 1 V, it gains 1 J of potential energy.

Units of V and ΔV : [J/C] (Joules per Coulomb) or [V] (Volts)

Consider the special case of a single charge q moving through a region of constant electric field, such as the area between two parallel charged plates (Fig. 2.8). If the displacement of the charge Δx is perfectly parallel to the electric field, we can divide Equation 3.3 by q to find the potential difference ΔV :

Single charge q in a constant electric field \vec{E}

$$\Delta V = \frac{\Delta PE}{q} = -|\vec{E}| |\Delta \vec{x}| \cos \theta = -E_x \Delta x \quad (3.13)$$

where the quantities are defined as in Eq. 3.2.

³ This is similar to the *chemical potential* in a way, if you are familiar with that.

⁴ The *gravitational potential* is the potential energy per unit mass, which is just gh for terrestrial cases, or $-\frac{Gm}{r}$ for the more general case. We would say that the *potential energy difference* between two points whose height differs by h is mgh , while the *potential difference* is just gh .

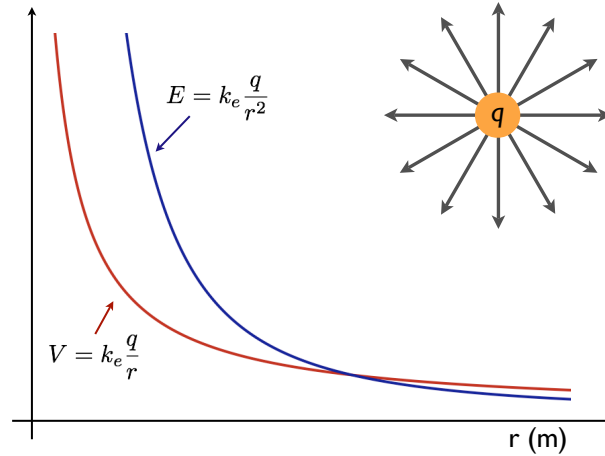
This lets us see that potential difference also has units of electric field times distance. This makes sense in a way, since for there to be an electrical potential difference we pretty much have to move through an electric field. Since electric field has the units of newtons per coulomb (N/C), we can make the following observation:

A newton (N) per coulomb (C) equals a volt (V) per meter (m): $1 \text{ N/C} = 1 \text{ V/m}$

If we release a positive charge, it spontaneously accelerates from regions of high potential to low potential - positive charges seek out minima in the electric potential. Conversely, negative charges seek out maxima in electric potential. Work must be done on positive charges to move them toward higher potential, work must be done on negative charges to move them to regions of lower potential.

3.2.1 Electric Potential and Potential Energy due to Point Charges

Fig. 3.3 The electric field \vec{E} and electric potential V versus the distance r from a point charge. Note V is proportional to $1/r$, while E is proportional to $1/r^2$.



As described briefly in Sect. 2.2.1.1, in electric circuits the zero point of electric potential ($V=0$) is defined by a “ground” wire connecting some point in the circuit to the earth. In a sense, defining a precise point at which $V=0$ through a ground wire is a bit like choosing an origin in a coordinate system. It can be anywhere you like, but you have to have one! For example, connecting the negative terminal of a 9 V battery to the ground would define the negative terminal as $V=0$, and the positive terminal would be at +9 V. If, on the other hand, we connected the positive terminal to ground, it would have $V=0$ and the negative terminal would have -9 V . In a way, the potential difference of the battery of 9 V well-defined, but the absolute potentials are *not* until a zero point is chosen.

For point charges, the electric field is defined throughout space, except *right at the charge*, and it works the same way for its electric potential. There is no obvious place to call “zero.” Further, we cannot connect a tiny ground wire to a single electron! (What could we make the wire out of ...) In the end, we nearly always, we define the potential for a point charge to be zero an infinite distance from the charge itself. This is actually convenient, believe it or not, and it makes clear the fact that the only way to get rid of the potential due to a point charge is to completely banish the charge itself. With this definition and some calculus, the electric potential of a point charge q at a distance r from the charge can be found as:

Electric potential created by a point charge:

$$V = k_e \frac{q}{r} \quad (3.14)$$

where r is the distance from the point charge q , and k_e is Coulomb's constant (Eq. 2.2).

This gives us the electric potential – work per unit charge – required to move the charge q from an infinite distance away to a point r . Figure 3.3 plots for comparison the electric field and electric potential for a point charge as a function of the distance from the charge. Keep in mind: you can only measure *differences* in electric potential. Some reference point must always be defined as $V=0$. For a point charge, this is $r=\infty$, for a circuit it is a specific point in the circuit.

One quick point, to clear up any later confusion: when dealing with point charges like electrons in electric fields, or atoms in a crystal (*e.g.*, in nuclear or atomic physics, and sometimes inorganic chemistry), we often use a more convenient unit of energy, the *electron volt*. We will encounter the electron volt more and more as time goes on, it turns out to be quite convenient when worrying about small numbers of charges.

An **Electron Volt [eV]** is the kinetic energy an electron gains when accelerated through a potential difference of 1 V.

$$1 \text{ eV} = 1.60 \times 10^{-19} \text{ C} \cdot \text{V} = 1.60 \times 10^{-19} \text{ J}$$

3.2.2 Energy of a System of Charges

Electric potential also obeys the superposition principle, just like the electric force. **The total electric potential at some point due to several point charges is just the sum of the electric potentials due to the individual point charges.** Since electric potential is a scalar, we do not need to worry about components, electric potentials are just numbers.

Figure 3.4 shows a “3-d” plot of the electric potential of an electric dipole (one positive charge and one negative charge close together, as in Fig. 2.5), where the color height scale represents the magnitude of the electric potential. As expected from the superposition principle, the potential is zero right between the two charges, and becomes very large near each charge, as does the electric field (Fig. 2.5).

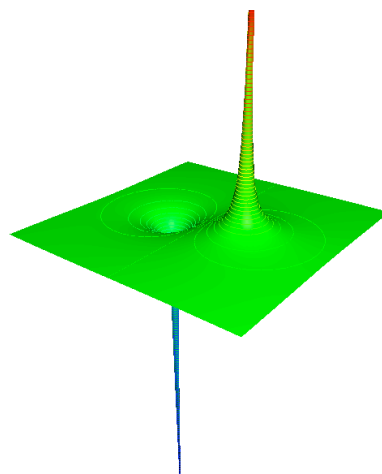


Fig. 3.4 The electric potential in a plane containing an electric dipole. The height (color) scale gives the electric potential. The lines represent equipotential contours.

From Eq. 3.12, we can see that it is easy to convert between electric potential and electric potential energy. What about the potential energy of two charges? If V_1 is the potential due to a charge q_1 at a point P , the work required to bring a charge q_2 from infinity to the point P is q_2V_1 , as shown in Fig. 3.5. That is, q_2V_1 is the energy it took to configure our system with charge q_2 at point P , and how much energy would be gained or lost by completely removing q_2 . Similarly, if q_2 is fixed in place, it takes q_1V_2 to bring q_1 in from an infinite distance to its final position.

This means that configuring two charges close to one another entails a gain or loss of energy – each charges feels the potential from the other. Bringing charges close together means energy is gained or lost to make that happen, and that energy is the potential energy of the pair of charges – how much energy is tied up in keeping those two charges where they are. For example, if two positive charges are to be kept close together against their natural repulsion, energy should be supplied to keep them together. If a positive and negative charge are to be kept together, energy should be supplied to keep them *apart*.

Now we see that potential energy really is the energy it takes to configure the system under study. Figure 3.5 also illustrates the difference between the *potential of a the separate point charges*, and the *potential energy of the pair of point charges*. If q_1 is already fixed its position, but q_2 is at infinity, the work that must be done to bring q_2 from infinity to its position near q_1 is $PE = q_2V_1 = k_e q_1 q_2 / r_{12}$. That is what the potential energy is, the energy of this configuration of charges *relative to just having q_1 all by itself*. If q_2 is fixed, it also takes $PE = k_e q_1 q_2 / r_{12}$ to bring in q_1 . Thus, it takes $PE = k_e q_1 q_2 / r_{12}$ to build our system of two charges, no matter how we do it:

$$PE_{\text{two charges}} = PE_{(1 \text{ due to } 2)} = PE_{(2 \text{ due to } 1)} = q_2V_1 = q_1V_2 = \frac{k_e q_2 q_1}{r_{12}} \quad (3.15)$$

As mentioned above, if the charges are of the same sign, PE is positive, and work must be done by an external force to bring the charges together. If they are of opposite charges, PE is negative, and negative work must be done to keep the charges from accelerating toward each other as they are brought together. In other words, work must be done to *keep* the charges apart. Another way to view the potential energy of the pair of charges is to think about how much kinetic energy would be gained if we let one of them loose again. If we have a pair of charges with an electrical potential energy of, say, 1 J with both charges fixed, the charges can gain between them 1 J of kinetic energy after being let loose. If one stays fixed, the other gets a full 1 J. If both charges are identical and both move, they each get 0.5 J.

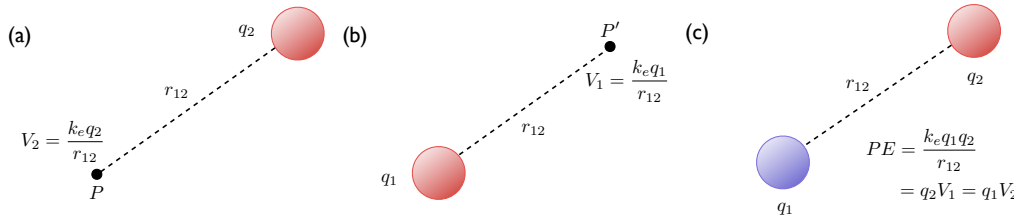


Fig. 3.5 (a) If the charge q_1 is removed, a *potential* $k_e q_2 / r_{12}$ exists at point P due to **charge** q_2 (b) Similarly, the charge q_1 gives a potential $k_e q_1 / r_{12}$ at point P' . (c) Either way we build our system of charges, the potential energy of the system of two charges is just $q_2V_1 = q_1V_2$, or $k_e q_1 q_2 / r_{12}$.

What if we have several charges? Just to be concrete, take the system of three point charges in Figure 3.6. We can obtain the total potential energy of this system by calculating the PE for every pair combination of charges and adding the results together. Since potential and potential energy are scalars, we don't need to worry about components – this is just an algebraic sum:

$$PE = PE_{1\&2} + PE_{2\&3} + PE_{1\&3} = PE_{2\&1} + PE_{3\&2} + PE_{3\&1} = k_e \left(\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \quad (3.16)$$

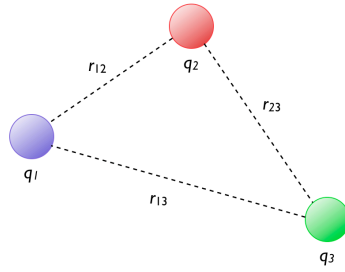


Fig. 3.6 A system of three point charges. Finding the total potential energy is just a matter of adding up the potential of pair combinations of charges.

Note that it doesn't matter what the order we sum them in, or if we transpose the labels – $PE_{1\&2}$ is the same thing as $PE_{2\&1}$, and r_{13} is the same as r_{31} , just like the example with two charges above.⁵

What does this really mean, physically? It is the same whether we have two charges or three or a million. What we are really summing up is the energy required to *build* this particular configuration of charges. Imagine that q_1 is fixed at the position shown in Figure 3.6, but that q_2 and q_3 are at infinity. The work that must be done to bring q_2 from infinity to its position near q_1 is $PE_{1\&2} = k_e q_1 q_2 / r_{12}$, which is the first term in Equation 3.16. The last two terms represent the work required to bring q_3 from infinity to its position near q_1 and q_2 , which involves the interaction with q_1 (the second term in Equation 3.16) and the interaction with q_2 (the third term in Equation 3.16). Compare this with Equation 3.15. Again, the result is independent of the order in which the charges are moved in from infinity.

We can write this more succinctly as a sum over all the charges:

$$PE = \frac{1}{2} \sum_{i=1}^3 \sum_{\substack{j=1 \\ j \neq i}}^3 \frac{k_e q_i q_j}{r_{ij}} \quad (3.17)$$

$$= \frac{1}{2} \left(\frac{k_e q_2 q_1}{r_{21}} + \frac{k_e q_3 q_1}{r_{31}} + \frac{k_e q_1 q_2}{r_{12}} + \frac{k_e q_3 q_2}{r_{32}} + \frac{k_e q_1 q_3}{r_{13}} + \frac{k_e q_2 q_3}{r_{23}} \right) \quad (3.18)$$

$$= k_e \left(\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \quad (3.19)$$

Here we color-coded the like terms for clarity. Basically, first we pick some charge j , and sum over all its pairings with the other charges i , making sure not to pair the charge with itself. Here we have the factor $\frac{1}{2}$ because the sum as written would count every pair of charges *twice* – since the pair 1&3 is the same as the pair 3&1. Think about that for a second, and reassure yourself that the factor $\frac{1}{2}$ is necessary. (If you are not familiar with summations, don't worry. We will only ever deal with a few charges at once.) For any arbitrary number of charges N , we can just change the limits on the sum:

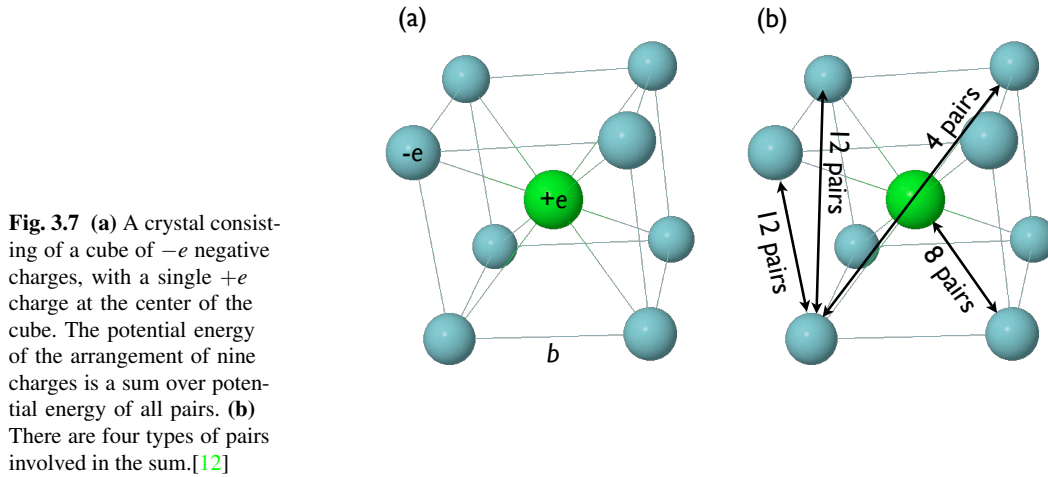
$$PE_{\text{total}} = \frac{1}{2} \sum_j^N \sum_{i \neq j}^N \frac{k_e q_i q_j}{r_{ij}} \quad (3.20)$$

The double-sum notation above means “take the charge $j=1$, and sum over all the other charges $i=2, 3, 4, \dots, N$, then take the charge $j=2$, and sum over the other charges $i=1, 3, 4, \dots, N$, and so on, until $j=N$.” Again, this counts every pair *twice*, hence the factor $\frac{1}{2}$.

3.2.2.1 Electrical Energy in a Crystal Lattice

What good is being able to find the energy of a large number of charges? Well, for one, this is one way to compute the stability of various crystal lattices. As an example, let us calculate the potential energy of eight negative charges on the corners of a cube of side b , with a single positive charge in the center. We will say each negative charge has $-e$, while the single positive charge is $+e$, Fig. 3.7

⁵ If you are into the math, that means we sum over all possible *combinations*, ${}^n C_k$, not *permutations*, ${}^n P_k$, so we do not count any pair more than once.



We can readily sum over all the possible pair interactions in the crystal, after a bit of geometry to figure out the distances between pairs.

For this crystal, we have 12 pairs of negative charges that are just one edge of the cube apart, twelve pairings between negative charges sideways across the cube faces, eight pairings between the negative corner charges and the central positive charge, and four corner to corner pairings of negative charges. This is illustrated in Fig. 3.7. Standard geometry tells us that the distance between edge charges is just b , the distance from corner to center is $\frac{\sqrt{3}}{2}b$, the corner-corner distance across a cube face is $\sqrt{2}b$, and finally the distance between opposite corner charges is $\sqrt{3}b$. The sum over all pairs is then:

$$PE_{\text{crystal}} = 8 \times \left[\frac{k_e(-e \cdot e)}{(\sqrt{3}/2)b} \right] + 12 \times \left[\frac{k_e e^2}{b} \right] + 12 \times \left[\frac{k_e e^2}{\sqrt{2}b} \right] + 4 \times \left[\frac{k_e e^2}{\sqrt{3}b} \right] \approx \frac{13.55 k_e e^2}{b} \quad (3.21)$$

Figure 3.7 shows where each term in the sum comes from. Though this seems a bit complicated, think about how hard it would be to compute the *forces* for every pair of charges and find the resultant vector force! We would have to do that for every stage of construction of the crystal, a tedious task at best. The relatively simple potential energy calculation above is a powerful way to address the amount of energy tied up in maintaining a particular charge distribution.

In this case, note that the total energy of this crystal lattice is *positive*, representing the fact that work had to be done on the crystal to assemble it in the first place. Left to its own devices, the charges in the crystal would want to disassemble. If we did let these charges move apart again, they would recover the potential energy as kinetic energy and speed away. This makes sense – it is silly to expect that real crystals are made of mostly negative charges, when we know that they are neutral overall. In reality, crystals are made of an equal number of positive and negative charges, which in many cases leads to a *negative* potential energy, indicating that the charges actually lower their energy by assembling into a crystal, and therefore favor doing so.

It is also curious that the potential energy sum for our cubic crystal ends up being a constant factor (about 13.55 times) what it would be for just a single pair of point charges separated by a distance b . In general, this is true for nearly any crystal lattice we can construct – the energy will always be some multiple of what for a single pair of charges. The multiple itself – in this case 13.55 – divided by the total number of charges is known as *Madelung's constant*, and every sort of crystal lattice has its own particular Madelung constant. The Madelung constant only depends on the geometric arrangement of the constituent ions in the crystal structure. Basically, the Madelung constant is something you look up in a table that takes care of all the nasty summing for you – someone has already done it! In general we can the potential energy of a crystal like this:

$$PE_{\text{crystal}} = \frac{1}{2}MN \frac{k_e z^2 e^2}{r} \quad (3.22)$$

here M is the Madelung constant, N is the number of charges we are considering, z is the charge of the ions in the lattice (± 1 in this case), and r is their separation. By inspection, you can see that for our cubic crystal, $13.55 = \frac{1}{2}MN$. Since there are $N=9$ charges in our example, our Madelung constant is $2(13.55)/10=2.71$.

If we take the structure of NaCl (common salt or rocksalt), the so-called face-centered cubic structure shown in Fig. 3.8a, the Madelung number ends up being about -1.75 if you carefully take the limit of the sum for very large N . The rocksalt structure has alternating positive Na^+ and Cl^- ions, arranged in a face-centered cubic structure. Overall, it is electrically neutral, and the negative potential energy reflects the stability of the structure. **The negative sign shows that work would have to be done to take the NaCl crystal apart – it is intrinsically stable.** This is in contrast to our fictitious body-centered cubic case above. Since our cubic crystal is mostly made of negative charges, it is not stable, and work has to be done to assemble it. The NaCl structure, however, has an equal number of positive and negative charges, and the negative potential energy sum explains the cohesion of the crystal and the fact that NaCl spontaneously assembles when Na and Cl are mixed. The Na and Cl constituents can lower their overall energy by assembling into a crystal, and that is what they do when given half a chance. The more negative the Madelung constant, the more stable the crystal is, if everything else is the same.

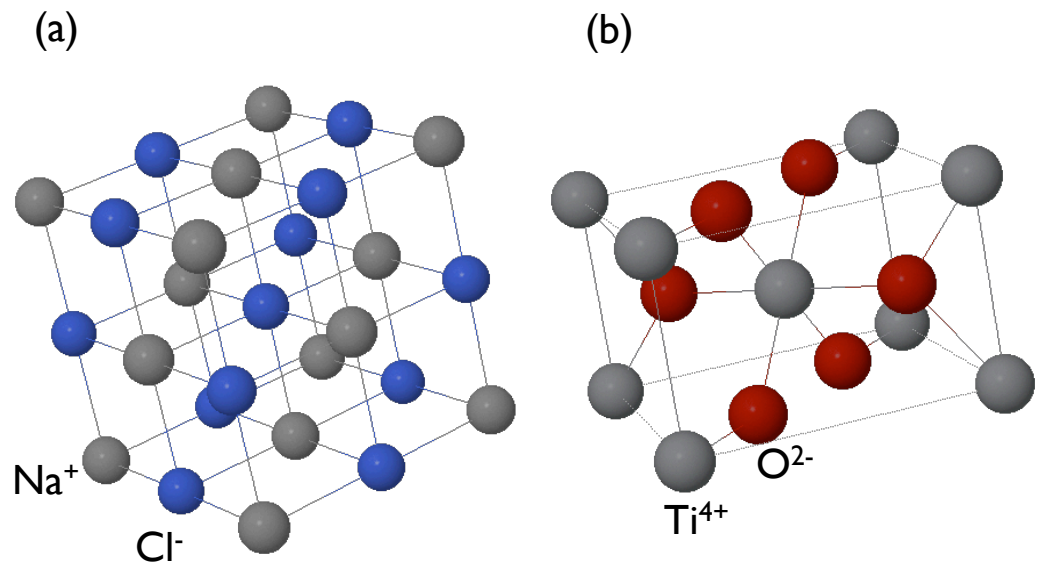


Fig. 3.8 (a) The NaCl or rocksalt structure. There are an equal number of Na^+ and Cl^- ions, the crystal is neutral overall. (b) The Rutile (TiO_2) structure. There are twice as many O^{2-} ions as Ti^{4+} to maintain neutrality.[12]

As another example, consider the Rutile (TiO_2) structure in Fig. 3.8b. In this case, the Madelung number is -4.82 , suggesting that rutile structure materials should be quite stable, and they generally are. There is one problem with all of this, however. Based on the analysis above, shrinking the distance b between charges in the crystal should make the potential energy even *more* negative. In other words, the smaller the spacing, the more stable the crystal would be. If that were true, why would the crystal not just keep shrinking until it collapsed? In fact, it can be shown that no system of stationary charges can be in a stable equilibrium according to classical physics. We need quantum physics to explain why, *e.g.*, salt crystals do not spontaneously shrink, and how crystals are stable in the first place.

3.3 Potentials and charged conductors

So the work done on a charge by an electric force is related to the change in electric potential energy of the charge. We also know that the change in electric potential energy between points A and B must be related to the potential difference between those two points. Putting these two facts together, we can easily relate work and potential difference:

Work and electrical potential for a charge moving from point A to B :

$$-W = \Delta PE = q(V_B - V_A) \quad (3.23)$$

where V_B is the electrical potential B , and V_A is the electrical potential A .

In Chapter 2, we said that for a conductor in electrostatic equilibrium, net charge resides only on the conductor's surface. Moreover, we said that the electric field just outside the surface of the conductor is perpendicular to the surface, and that the field inside the conductor is zero. This also means that **all points on the surface of a charged conductor in electrostatic equilibrium are at the same potential**.

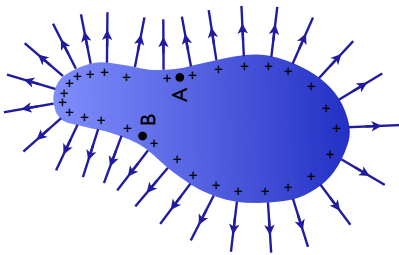


Fig. 3.9 An arbitrarily shaped conductor carrying a positive charge. When the conductor is in electrostatic equilibrium, all of the charge resides at the surface, $\vec{E} = 0$ inside the conductor, and the direction of \vec{E} just outside the conductor is perpendicular to the surface. The electric potential is constant inside the conductor and is equal to the potential at the surface. Note from the spacing of the positive signs that the surface charge density is nonuniform.

Equation 3.23 gives us a very general result: **no net work is required to move a charge between two points which are at the same electric potential**. Mathematically, $W = 0$ whenever $V_B = V_A$.

Consider the path connecting points A and B along the surface of the conductor in Figure 3.9. If we move only along the conductor's surface, the electric field \vec{E} is always perpendicular to our path. Since the electric field and displacement are always perpendicular, no work is done when moving along the surface of a conductor. Equation 3.23 then tells us that if the work is zero, points A and B must be at the same potential, $V_B - V_A = 0$. Since the path we have chosen is completely arbitrary, this means it is true for any two points on the surface.

Potentials and charged conductors

1. electric potential is a constant on the surface
2. electric potential is constant inside, and has the same as the value at the surface
3. no work is required to move a charge from the interior to the surface, or between two points on the surface

Of course, this only holds for perfect conductors. If other dissipative (or non-conservative) forces are present, this is not true, and work *is* required to move the charge in the presence of a dissipative

force. The electrical analog of friction or viscosity is *resistance*, which will be treated in the next chapter.

3.4 Equipotential Surfaces

A surface on which all points are at the same electric potential is called an *equipotential surface*. The potential difference between any two points on the surface is zero, hence, **no work is required to move a charge at constant speed on an equipotential surface**. The surface of a conductor is therefore an equipotential surface. Equipotential surfaces have a simple relationship to the \vec{E} field: the field is perpendicular to the equipotential surface at every point. Figure 3.10 shows equipotential surfaces and electric field lines for a single point charge, a dipole, and two like charges. Notice that once you have drawn electric field lines, drawing equipotential surfaces is trivial, and *vice versa*.

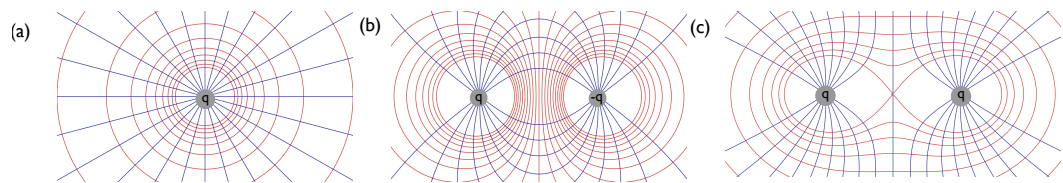


Fig. 3.10 The blue lines electric field lines, and the red lines are equipotential surfaces for (a) a single point charge, (b) an electric dipole, and (c) two like charges. In each case, the equipotential surfaces are perpendicular to the electric field lines at every point. (Again, arrows are left off of the field lines for simplicity. Equipotential lines do not need arrows, since potential is a scalar.)

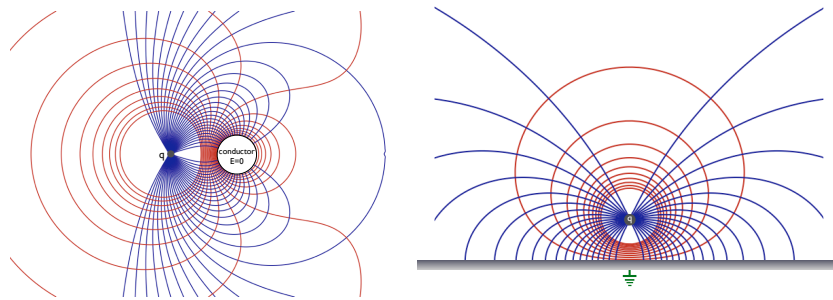


Fig. 3.11 The blue lines electric field lines, and the red lines are equipotential surfaces for **left** a conducting sphere near a point charge q , and **right** a point charge suspended above a long grounded conducting plate.

More examples are given in Fig. 3.11, which include conductors. For a conductor, we know the electric field inside is zero, and the electric potential is constant. Add to this the fact that electric field lines and equipotential lines are *always* perpendicular where they meet, and you should be able to explain all of the examples shown here. This why in the right-hand example, a single charge above a ground plane, the electric field lines all intersect the ground plane at perfect right angles, and in the left-hand example, there are no lines inside the conducting sphere. Compare these figures with Fig. 2.8 – the relationship between electric field lines and equipotential lines should be clear. Appendix ?? might give you a bit more insight as to why the electric field lines and equipotential lines behave the way they do. Recall from Sect. 2.5 that *a conductors are mirrors for electric field lines, the same is true for the equipotential lines*.

3.5 Potential Difference Sources as Circuit Elements

How do we actually change the electric potential – which we will usually just call voltage henceforth – of one object relative to another? Charging by induction or conduction are two ways, but somewhat cumbersome. A device known as a *voltage source* is a circuit element with two terminals, where a constant potential difference is supplied between these two terminals. Whatever you connect to the “negative” terminal of the source will have a voltage ΔV lower than the “positive” terminal. Using a “ground” point (recall Sect. 2.2.1.1), one can also experimentally define one of the terminals as $V = 0$. If we “ground” the negative terminal, then the negative terminal is $V_{\text{neg}} = 0$, and the positive terminal has $V_{\text{pos}} = \Delta V$. We will see much more of this in the coming chapters, and it will begin to make more sense!

Batteries are one example of a constant voltage source, which we will cover in more detail in Chapter 5, and the wall outlets in your house are another example of a voltage source (though this voltage is not strictly constant, see Chapter 8). Ideal textbook voltage sources *always* supply a constant potential difference, ΔV . Real voltage sources always have restrictions, a primary one being the amount of power that can be sourced. Below are circuit diagram symbols for constant voltage sources: the first two represents batteries, the last is a generic symbol for any more complicated sort of voltage source:

Circuit diagram symbol for voltage sources:

Batteries:

General constant voltage source:

Now that we know a bit about voltage and conductors, we are moving closer to being able to describe simple electric circuits. Presently, we will introduce our first real circuit element, the capacitor.

3.6 Capacitance

A **capacitor** is an electronic component used to store electric charge, it is used in essentially any electric circuit you can name. Capacitors are at the heart of both Random Access Memory (RAM) and flash memory, besides being crucial for nearly any sort of power supply. It is one of the fundamental building blocks for electronics, and the first we will meet. Figure 3.12 shows a typical design for a capacitor – two metal plates with some special stuff in between. It is hard to believe complicated devices like computers rely on such a simple construction, but it is true!

A typical capacitor consist of two parallel metal plates, separated by a distance d . When used in a circuit, the plates are connected to the positive and negative terminals of a voltage source such as a battery. An ideal voltage source insists that the two plates have a voltage difference of ΔV , and this has the effect of pulling electrons off of one plate, leaving it with a net positive charge $+Q$, and transferring these electrons to the second plate, leaving it with a net negative charge $-Q$. The charge on both plates is equal, but opposite in sign. Essentially, putting the two plates at different potentials means electrons want to migrate to the plate with higher potential, and leave the plate with lower potential deficient.

The transfer of charge between the plates stops when the potential difference across the plates is the same as the potential difference of the voltage source. *The capacitor stores this potential difference, and hence stores electrical energy, until some later time when it can be reclaimed for a specific application.* You can think of this as energy storage from one point of view, or a time-delayed response from another.

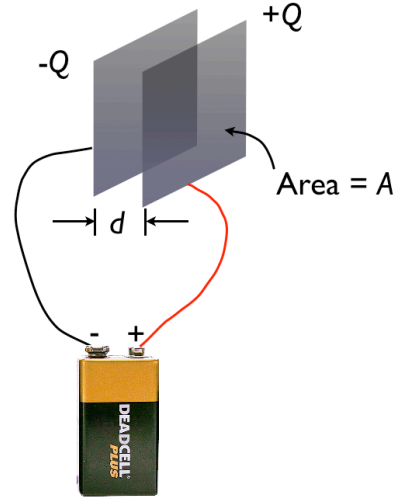


Fig. 3.12 A parallel-plate capacitor consists of two conducting plates of area A , separated by a distance d . The capacitance of this structure is $C = \epsilon_0 A / d$.

Keep in mind (again): you can only *measure* differences in electric potential. Some reference point must always be defined as $V = 0$. In the case of the capacitor connected only to a battery (without any ground points), the potential is zero half way between the two plates.⁶

Definition of Capacitance:

The capacitance C is the ratio of the charge stored on one conductor (or the other) to the potential difference between the conductors:

$$C \equiv \left| \frac{Q}{\Delta V} \right| \quad (3.24)$$

C is always positive, and has units of farads [F], or coulombs per volt [C/V].

3.6.1 Parallel-Plate Capacitors

The capacitance of a particular arrangement of two conductors depends on their geometry and relative arrangement. One common (and simple) structure is the **parallel plate** capacitor, as shown in Figure 3.12. In Chapter 2, we stated without proof (but not without good reason) that the electric field between two parallel plates is constant. But what is the field in between the plates?

First, we assume that the two plates are identical, such that they have the same charge on them – one has $+Q$ and one has $-Q$. Second, we assume the plates area A is large compared to their spacing d , such that we can ignore the edge regions where the field “fringes” (see Fig. 2.8 and 3.13). Finally, we will connect the plates to a battery with total voltage V .

In Sect. 2.8.4, we found that the electric field above a flat conducting plate is given by $E = \sigma_E / \epsilon_0$, where σ_E is the charge per unit area on the plate. Since the total charge on each plate is just Q , the charge per unit area is $\sigma_E = Q/A$, and $Q = \sigma_E A$. This leads us to a more useful expression for the field: $E = \frac{Q}{A\epsilon_0}$. Again, this is not valid near the edges of the plates where the field is not really constant.

Now where the field *is* constant, we know that the potential difference between the two plates is $\Delta V = Ed$, where d is the distance between the two plates. Combining this with the facts above, we can find the capacitance of the parallel plate capacitor from Equation 3.24:

⁶ The potential is also zero infinitely far away of course, but this is hardly useful or reassuring when wiring a circuit.

$$C = \frac{Q}{\Delta V} = \frac{\sigma_E A}{Ed} = \frac{\sigma_E A}{(\sigma_E/\epsilon_0)d} = \frac{\cancel{\sigma_E} A}{(\cancel{\sigma_E}/\epsilon_0)d} = \epsilon_0 \frac{A}{d} \quad (3.25)$$

Capacitance of a parallel plate capacitor:

$$C = \epsilon_0 \frac{A}{d} \quad (3.26)$$

where d is the spacing between the plates, and A is the area of the plates.

We can see from Equation 3.26 that capacitors can store more charge when the plates become larger. The same is true when the plates get close together. When the plates are closer together, the opposing charges exert a stronger force on each other, allowing more charge to be stored on the plates. From Equation 3.24, **a capacitor of value C at a potential difference of ΔV stores a charge $Q = C\Delta V$.**

Figure 3.13 shows more realistic field lines for a parallel plate capacitor. In between the two plates, the field is very nearly constant, but much less so near the edges of the plates. So long as the plates are relatively large compared to their separation, we can for practical purposes ignore this complication, and our capacitance calculated from Eq. 3.26 will be very accurate.

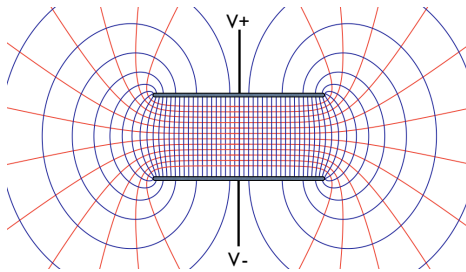


Fig. 3.13 (a) The electric field (blue) and equipotential (red) lines near and between the plates of a parallel-plate capacitor. The potential and field are both uniform near the center, but nonuniform near the edges.

Capacitors form the basis for several types of Random Access Memory (RAM) in modern computers. Dynamic random access memory (DRAM) is one type of random access memory that stores each bit of data in a separate capacitor. One capacitor in a DRAM structure holds one *bit* of information (a “1” or a “0”). When the capacitor has charge stored in it, the bit is a “1,” and when there is no charge stored the bit is a “0.” Flash memory works in a roughly similar manner.

3.6.2 Energy stored in capacitors

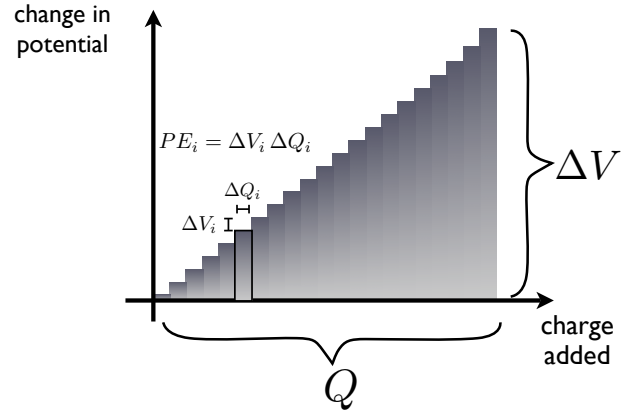
Capacitors store electrical energy. Anyone who has worked with electronic equipment long enough has verified this one painful way or another.⁷ If the plates of a charged capacitor are connected to a conducting object, the capacitor will transfer charge from one plate to another until it is discharged. This is often seen as a “spark” if the capacitor was charged to a high enough voltage. Given that humans are reasonably good conductors at high voltages, this can be a problem.

Charged capacitors store energy, and that energy is the work required to move the charge onto the plates. If a capacitor is initially uncharged (both plates neutral), very little work is required to move a charge ΔQ from one plate to another across the separation d . As soon as this charge is moved, however, a potential difference $\Delta V = \Delta Q/C$ appears between the plates. This potential difference

⁷ I once burned a small hole in my thumb by accidentally discharging a high-voltage capacitor across it while repairing a TV, for example. Capacitors can store dangerous amounts of energy if released at the wrong time!

3.1. An ideal parallel plate capacitor is completely charged up, and then disconnected from a battery. The plates are then pulled a small distance apart. What happens to the capacitance, C , and charge stored, Q , respectively?

Fig. 3.14 Each bit of charge ΔQ_i transferred through a voltage ΔV_i contributes a bit of potential energy $PE_i = \Delta V_i \Delta Q_i$. Summing all those contributions to get the total energy stored is the same as finding the total area of the shaded region. If we make ΔV_i and ΔQ_i tiny enough, the area is basically a triangle, and in total $PE = \frac{1}{2} Q \Delta V$.



$$\Delta PE = -W = \sum_i \Delta Q_i \Delta V_i = \text{area under curve} = \frac{1}{2} Q \Delta V$$

means that work must be done to move additional charges onto the plates. Combining what we know so far, and assuming a constant electric field between the plates, the work that needs to be done to move the first bit of charge ΔQ has to be:

$$\Delta PE = -\Delta W \quad (3.27)$$

$$= \Delta Q \cdot E \Delta x \quad (3.28)$$

$$= \Delta Q \cdot E d \quad (3.29)$$

$$= \frac{1}{\epsilon_0} \Delta Q \sigma_E d \quad (3.30)$$

But we know that $\sigma_E = \frac{\Delta Q}{A}$, and thus $\Delta Q = \sigma_E A$, which simplifies things:

$$\Delta PE = \Delta Q \Delta Q \frac{d}{A \epsilon_0} \quad (3.31)$$

Since $C = \frac{\epsilon_0 A}{d}$ for our parallel plate capacitor,

$$\Delta PE = \frac{(\Delta Q)(\Delta Q)}{C} \quad (3.32)$$

If we keep doing this with more and more ΔQ s, until we build up the total charge Q , we can find the total work. As illustrated in Fig. 3.14, each little bit of charge ΔQ_i adds a bit of potential energy $\Delta V_i \Delta Q_i$. If we sum up all those contributions, we are really just finding the shaded area of the triangle on the graph. The area of a triangle is just $\frac{1}{2}(\text{base})(\text{height})$, so the total change in potential energy is just:⁸

$$|W| = |\Delta PE| = \frac{1}{2} Q \Delta V \quad (3.33)$$

Remember that $Q = C \Delta V$ must still be true, so we can write the energy stored in the capacitor in three different ways, as shown below (noting that energy stored = work done). For example, you can verify that a 5 μF capacitor charged with a 120 V source stores 3.6 mJ ($3.6 \times 10^{-3} \text{ J}$).

Energy stored in a capacitor:

⁸ This is a bit of a hand-waving derivation, but it doesn't require any calculus like the more rigorous version does.

$$\text{Energy stored} = \frac{1}{2} Q \Delta V = \frac{1}{2} C (\Delta V)^2 = \frac{Q^2}{2C} \quad (3.34)$$

Remember that the units of energy are **Joules**.

Is there an analogy for electrical energy storage? One way to store *gravitational energy* is simply to pump a large mass m of water up to a height Δy , see Figure 3.15. Releasing the water at a later time releases the stored potential energy $mg\Delta y$, which could be used to, *e.g.*, rotate a turbine. In fact, this is one way to store excess energy generated at off-peak times in power plants for later reclamation.

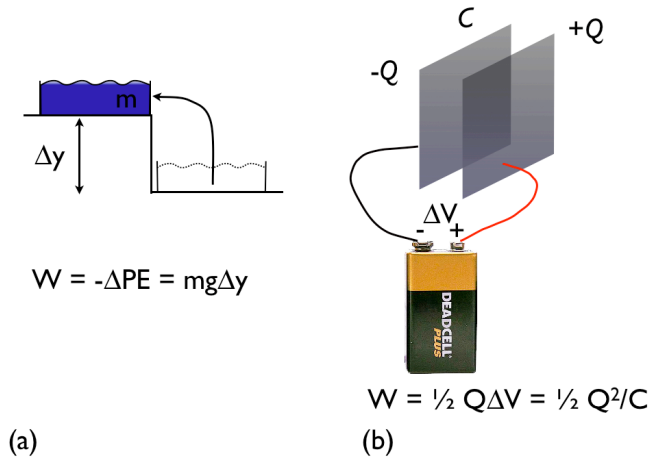


Fig. 3.15 (a) Raising a mass m of water to a height Δy above the ground stores an energy $mg\Delta y$. (b) Charging a capacitor C with a potential difference ΔV stores an energy $\frac{1}{2} Q \Delta V = \frac{Q^2}{2C}$.

3.6.3 Capacitors as Circuit Elements

Now that we know about a second circuit element, we can begin to make some simple circuits. As you might have gathered above, capacitors are often used in electrical circuits as energy-storage devices. As we will find out later, they can also be used to filter out high- and low-frequency signal selectively. The circuit diagram symbol for a capacitor is a reminder of the parallel plate geometry:

Circuit diagram symbol for a capacitor:

What can we do only knowing about two circuit components, capacitors and batteries? Well, we can hook up a capacitor to a battery, as shown in Fig. 3.16

What does this circuit do? The moment we connect the battery to the capacitor, charges will start to flow from one plate to another for time, until both plates are fully charged. Fully charged means that the potential difference between the two plates is the same as that at the battery terminals, ΔV . After that ... nothing. The capacitor will just happily store these charges. If the capacitor is disconnected from the battery, the charges will remain on the two plates since they have no path to escape. The capacitor stays charged, thereby storing energy, so long as it is truly isolated. If one of the plates had a path to ground, for instance, the charges would leak away *via* this ground connection, and the energy would dissipate. In a rough sense, FLASH memory works by storing charges on very tiny, isolated conducting plates.

We cannot do very much with only capacitors and batteries, but we will remedy this in subsequent chapters. For now, there are a few more things we can figure out about capacitors.

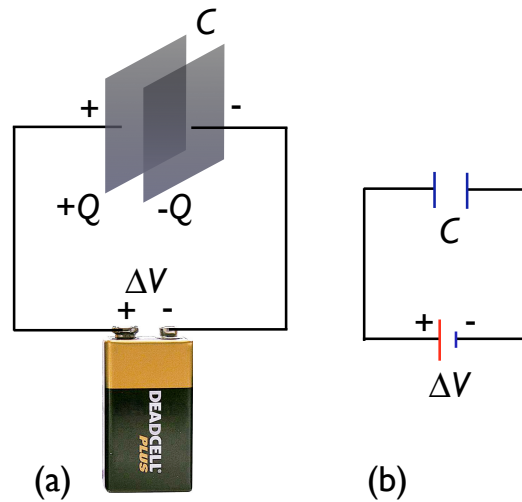


Fig. 3.16 (a) A parallel plate capacitor of value C connected to a battery supplying a voltage difference ΔV (b) Circuit diagram for this configuration.

3.6.4 Combinations of Capacitors

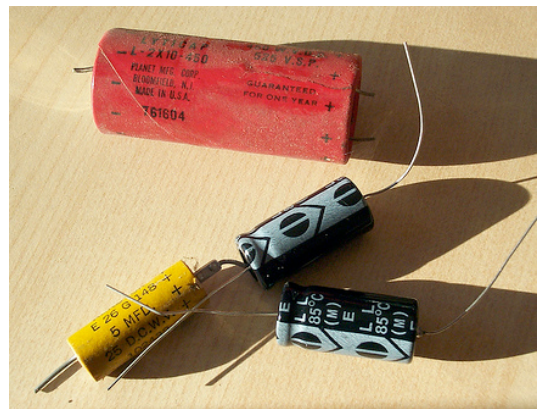


Fig. 3.17 A picture of several common types of capacitors.[13]

Two or more capacitors can be combined in circuits in many possible ways, but most reduce to two simple configurations: **parallel** and **series**. Two capacitors in series or in parallel can be reduced to a single equivalent capacitance, and more complicated arrangements can be viewed as combinations of series and parallel capacitors.

3.6.4.1 Parallel Capacitors

Capacitors are manufactured with standard values, and by combining them in different ways, any non-standard value of capacitance can be realized. Figure 3.18 shows a parallel arrangement of capacitors. The left plate of each capacitor is connected by a wire (black lines) to the positive terminal of a battery, while the right plate of each capacitor is connected to the negative terminal of the battery.⁹ This means that **the capacitors in parallel both have the same potential difference ΔV across them**, the voltage supplied by the battery.

When the capacitors are first connected, electrons leave the positive plates and go to the negative plates until equilibrium is reached - when the voltage on the capacitors is equal to the voltage of the battery. The internal (chemical) energy of the battery is the source of energy for this transfer. In

⁹ In circuit diagrams like these, the wires are assumed to be perfect.

this configuration, both capacitors charge independently, and the total charge stored is the sum of the charge stored in C_1 and the charge stored in C_2 . We can write the charge on the capacitors using Equation 3.24:

$$\begin{aligned} Q_1 &= C_1 \Delta V \\ Q_2 &= C_2 \Delta V \\ Q_{\text{total}} &= Q_1 + Q_2 = C_1 \Delta V + C_2 \Delta V = (C_1 + C_2) \Delta V \end{aligned}$$

What this equation shows is that **two capacitors in parallel behave as one single capacitor** with a value of $C_1 + C_2$. In other words, “capacitors add to each other in parallel.” We call $C_1 + C_2$ the “*equivalent capacitance*”, $C_{\text{eq}} = C_1 + C_2$

Two Capacitors in Parallel:

$$C_{\text{eq}} = C_1 + C_2 \quad (3.35)$$

Three or More Capacitors in Parallel:

$$C_{\text{eq}} = C_1 + C_2 + C_3 + \dots \quad (3.36)$$

The key point for capacitors in parallel is that **the voltage on each capacitor is the same**. One way to see this is that they are both connected to the battery by the same perfect wires, so they pretty much *have* to have the same voltage. This is true in general, as we will find out, so long as we have perfect textbook wires. It follows readily that **the equivalent capacitance of a parallel combination is always more than either of the individual capacitors**.

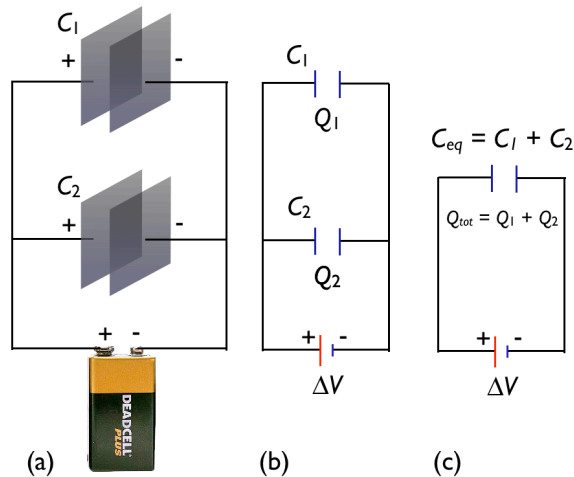


Fig. 3.18 (a) A parallel connection of two capacitors to a battery (b) The circuit diagram for the parallel combination. (c) The potential differences across the capacitors are the same, and the equivalent capacitance is $C_{\text{eq}} = C_1 + C_2$

3.6.4.2 Series Capacitors

Figure 3.19a shows the second simple combination, two capacitors connected in *series*. **For series capacitors, the magnitude of charge is the same on all plates**. Consider the left-most plate of C_1 and right-most plate of C_2 in Figure 3.19. Since they are connected directly to the battery, they must have the same magnitude of charge, $+Q$ and $-Q$ respectively.

Since the middle two plates (the right plate of C_1 and the left plate of C_2) are not connected to the battery at all, *together they must have no net charge*. On the other hand, the left and right plates

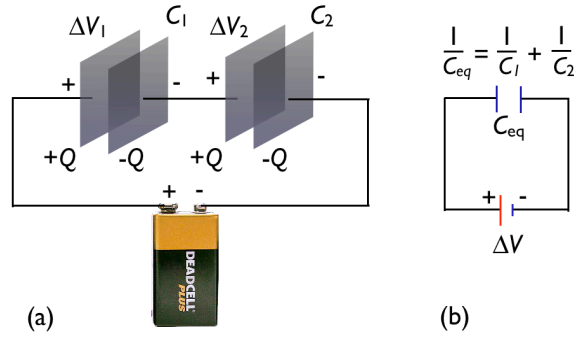


Fig. 3.19 (a) A series combination of two capacitors. The charges on the capacitors are the same. (b) Circuit diagram corresponding to (a). The equivalent capacitance be calculated from the reciprocal relationship, $\frac{1}{C_{eq}} = \frac{1}{C_1} + \frac{1}{C_2}$

of the *same capacitor* have to have the same magnitude of charge, so this means *all plates have a charge of either $+Q$ or $-Q$ stored on them*. **All of the right plates have charge $-Q$, and all the left plates have a charge $+Q$**

Can we reduce this series combination to a single equivalent capacitor, like we did for the parallel case? Sure, with a little math. A single capacitor equivalent to the series capacitors, Figure 3.19b, must have a charge of $+Q$ on its right plate, and $-Q$ on its left plate, so the total charge stored is still $\pm Q$ on each plate.. Further, it must have a potential difference equal to that of the battery, ΔV . Using Equation 3.24:

$$\Delta V = \frac{Q}{C_{eq}} \quad (3.37)$$

We can also apply Equation 3.24 to each of the individual capacitors:

$$\Delta V_1 = \frac{Q}{C_1} \quad \Delta V_2 = \frac{Q}{C_2} \quad (3.38)$$

Conservation of energy requires that all of the potential difference of the battery ΔV be “used up” somewhere. Since our wires are assumed to be perfect, the only place the potential can go is onto the capacitors. Therefore, for the series case the voltage on C_1 and C_2 must together total that of the battery:

$$\Delta V = \Delta V_1 + \Delta V_2 \quad (3.39)$$

This, combined with Equations 3.37 and 3.38, gives us:

$$\Delta V = \frac{Q}{C_{eq}} = \frac{Q}{C_1} + \frac{Q}{C_2} \quad (3.40)$$

Canceling the Q ’s, we can come up with the equivalent capacitance for series capacitors:

Two Capacitors in Series:

$$\frac{1}{C_{eq}} = \frac{1}{C_1} + \frac{1}{C_2} \quad (3.41)$$

Three or More Capacitors in Series:

$$\frac{1}{C_{eq}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} \dots \quad (3.42)$$

It follows that **the equivalent capacitance of a series combination is always less than either of the individual capacitors**. The key point for capacitors in series is that the *charge* on each capacitor is the same, and the same as the charge on the equivalent capacitor.

What to do for more complex combinations of capacitors?

1. **Combine** capacitors that are in parallel or series in to single equivalent capacitors, using (3.35) and (3.41).
2. **Parallel** capacitors all have the same potential difference ΔV across them.
3. **Series** capacitors all have the same charge Q , which is the same as the charge on their equivalent capacitor.
4. **Redraw** the circuit after every combination.
5. **Repeat** the first two steps until there is only equivalent one capacitor left.
6. **Find the charge** on this equivalent capacitor using (3.24).
7. **Reverse** your steps one by one to find the charge and voltage drop on each equivalent capacitor along the way, until you recreate the original diagram.

3.6.4.3 Example of a complex capacitor combination

The easiest way to see how one can use the rules for series and parallel capacitors to reduce *any* complex combination of capacitors to a single equivalent capacitor is by example. For example, consider the combination of capacitors in Figure 3.20 below.

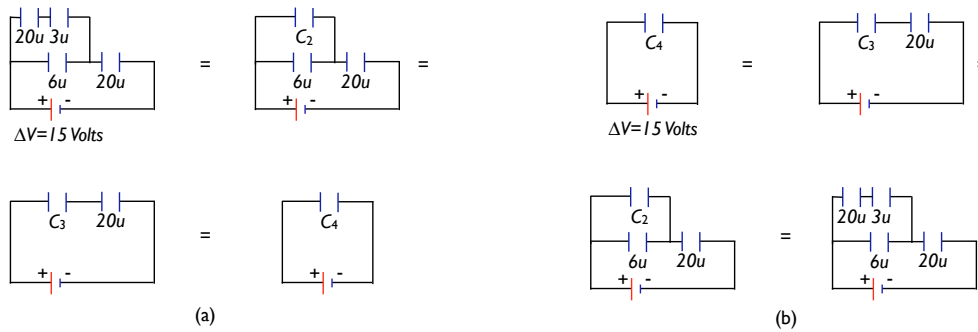


Fig. 3.20 (a) Reducing the complex combination to a single equivalent capacitor. (b) Working backwards to find the charge on each capacitor.

Finding the equivalent capacitor

First, we notice from Figure 3.20a that the only purely series or parallel combination to start with is the $20\mu\text{F}$ and $3\mu\text{F}$ capacitors in series. We can combine those into an equivalent capacitance, C_2 , using Equation 3.41:

$$\frac{1}{C_2} = \frac{1}{20\mu\text{F}} + \frac{1}{3\mu\text{F}} \quad (3.43)$$

$$C_2 = \frac{1}{\frac{1}{20\mu\text{F}} + \frac{1}{3\mu\text{F}}} = \frac{3 \cdot 20}{3 + 20} \quad (3.44)$$

$$C_2 = 2.6\mu\text{F} \quad (3.45)$$

Redraw the circuit to reflect this change, and we arrive at the second diagram in Figure 3.20a. Now we have the equivalent capacitor C_2 purely in parallel with the $6\mu\text{F}$ capacitor. Using Equation 3.35, we can combine those two into another equivalent capacitance C_3 :

$$C_3 = C_2 + 6\mu\text{F} = 8.6\mu\text{F} \quad (3.46)$$

Redraw the circuit, and we arrive at the third diagram in Figure 3.20a. Now we only have C_3 in parallel with $20\mu\text{F}$ left, which we can now combine into a final overall equivalent capacitance C_4 . Again using Equation 3.41, we have

$$\frac{1}{C_4} = \frac{1}{C_3} + \frac{1}{20\mu\text{F}} \quad \text{or} \quad C_4 = 6.02\mu\text{F} \quad (3.47)$$

So the equivalent capacitance of the four capacitors we started with is about $6\mu\text{F}$.

Finding the charge on each capacitor

Now we have to work *backwards* from our single equivalent capacitor and deduce the charge and voltage on each individual capacitor, following Figure 3.20b. First, we know the charge on C_4 , the equivalent capacitor, once we know the value of C_4 (above) and ΔV (given):

$$Q_4 = C_4 \Delta V = (6.02\mu\text{F})(15\text{V}) = 90.3\mu\text{C}$$

Now C_3 and the $20\mu\text{F}$ are in series. Two series capacitors must both have the *same charge but different voltages*. Further, the charge on series capacitors is *the same as the charge on the equivalent capacitor*. Therefore, both the $20\mu\text{F}$ and C_3 have to have the same charge that C_4 has. So

$$Q_3 = Q_{20\mu} = Q_4 = 90.3\mu\text{C}$$

Now we get to the third diagram. We know that the $6\mu\text{F}$ and C_2 *together* have Q_4 worth of charge. *Parallel capacitors both have the same voltage, but different charges*. If we call the voltage on these two capacitors V , the charge on the $6\mu\text{F}$ is $6\mu\text{F} \cdot V$, and the charge on C_2 is $C_2 \cdot V$, which gives us Q_4 :

$$Q_4 = 90.3\mu\text{C} = (C_2)V + (6\mu\text{F})V$$

Since $C_2 = 2.6\mu\text{F}$, this gives $V = 10.47$ Volts, so

$$Q_{6\mu} = (6\mu\text{F})V = 62.9\mu\text{C} \quad \text{and} \quad Q_2 = (C_2)V = 27.4\mu\text{C}$$

Note that the voltage V and the voltage on the lower $20\mu\text{F}$ capacitor must together equal the battery voltage, so the voltage on the lower $20\mu\text{F}$ capacitor must be $15.00 - 10.47 = 4.53\text{V}$. Now for the last step. You now know the charge on C_2 , which is the same as the total charge on the $20\mu\text{F}$ and $3\mu\text{F}$ capacitors. Since they are in series, they both have the same charge, and the both have to have Q_2 . Thus $Q_{3\mu} = Q_{20\mu} = 27.4\mu\text{C}$. We can find the voltage on each by noting that

$$V_{3\mu} = \frac{Q_{3\mu}}{C_{3\mu}} = 9.13\text{V} \quad \text{and} \quad V_{20\mu} = \frac{Q_{20\mu}}{C_{20\mu}} = 1.37\text{V}$$

Further, we know that that $V_{3\mu} + V_{20\mu}$ has to equal the voltage on the equivalent capacitor C_2 , *viz.* 10.47V . So, in the end, the charge on the $20\mu\text{F}$ is the same as that on the effective capacitance, the charge on $20\mu\text{F}$ and the $3\mu\text{F}$ are the same, and the charge on the $6\mu\text{F}$ is about halfway in between either of those. The charge, capacitance, and voltages are summarized in Table 3.1.

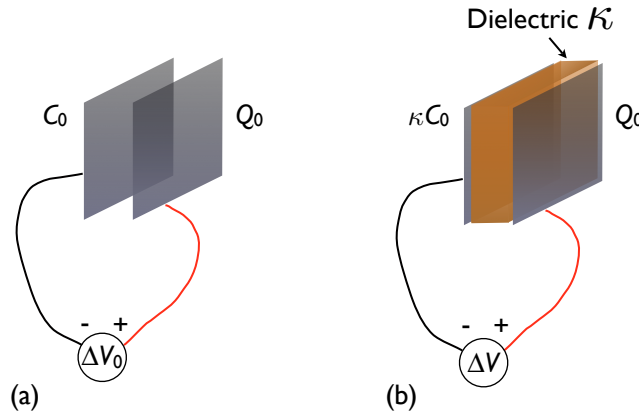
3.6.5 Capacitors with (non-conducting) stuff inside

What if we separate the plates of our parallel plate capacitor with something other than air? As you might expect, this changes the capacitance. A **dielectric** is another name for an insulating material

Table 3.1 Equivalent capacitances, charges, and voltages for Figure 3.20.

Capacitor [μF]	Charge [μC]	Voltage [V]
top 20 μ	27.4	1.37
$C_2 = 2.6$	27.4	10.47
$C_3 = 8.6$	90.3	10.47
$C_4 = 6.02$	90.3	15
6 μ	63	10.47
3 μ	27.4	9.13
lower 20	90	4.53

Fig. 3.21 (a) With air between the plates, the voltage across the capacitor is ΔV_0 , the capacitance is C_0 , and the charge is Q_0 . (b) With a dielectric inside, the charge is still Q_0 , but the voltage and capacitance change.



(like rubber, or most ceramics and plastics). When we put a dielectric between the plates of our capacitor, the capacitance increases. If the dielectric totally fills the region between the plates, the increase is proportional to a constant κ , the dielectric constant. We note that sometimes you will see the dielectric constant is written as ϵ_r rather than κ , but it is the same thing.

Figure 3.21 shows the effect of a dielectric inserted in a parallel plate capacitor. Without the dielectric, we know that $\Delta V_0 = Q_0/C_0$. If we now insert the dielectric, the voltage is *reduced* to:

$$\Delta V = \frac{\Delta V_0}{\kappa} = \frac{\Delta V_0}{\epsilon_r} \quad (3.48)$$

What happens is that part of the potential difference originally across the plates of the capacitor is now spent on the dielectric itself. Being an insulator, the dielectric can support regions of charge, unlike a conductor. When it is inserted into the capacitor, the part of the dielectric near the $+Q_0$ plate builds up a partial *negative* charge in response, and the part near the $-Q_0$ plate builds up a partial *positive* charge. This has the effect of “canceling” part of the $+Q$ and $-Q$ charges on the plates, so the battery supplies more charges to compensate! This goes on until an equilibrium is reached, and the dielectric can steal no more charge.

In the end, since the dielectric “steals” a bit of extra charge, the capacitor with a dielectric inside stores more charge than the capacitor without the dielectric. The total amount of charge present, including the “extra” bit “stolen” by the dielectric, is proportional to κ , so the capacitance of the new structure is *increased* by a factor of κ :

$$C = \frac{Q_0}{\Delta V} = \frac{\kappa Q_0}{\Delta V_0} = \frac{\epsilon_r Q_0}{\Delta V_0} \quad (3.49)$$

For a parallel plate capacitor, this means:

Parallel plate capacitor with a dielectric between the plates:

$$C = \kappa \epsilon_0 \frac{A}{d} = \epsilon_r \epsilon_0 \frac{A}{d} \quad (3.50)$$

the dielectric *increases* the capacitance by a factor κ , the dielectric constant. The dielectric constant is also sometimes called ϵ_r .

This is not an insignificant effect - the value of κ can range from ~ 1 for air to a few thousands – adding a good dielectric layer can increase the amount of charge stored by hundreds or thousands! For vacuum, the value is exactly 1, so Equation 3.50 just reduces to Equation 3.26. The value of κ is **always** greater than 1 ($\kappa > 1$), so the capacitance always increases when a dielectric is included. Why this is true microscopically is treated in the next section. Table 3.2 lists the dielectric constants for a few common materials.

Table 3.2 Dielectric constants of materials at $T_0 = 20^\circ\text{C}$ [14]

Material	κ	Material	κ
Vacuum	1		
Air	1.00054	Teflon [®]	2.1
Polyethylene	2.25	Paper	3.5
Silicon dioxide	3.7	Pyrex	4.7
Rubber	7	Methanol	30
Silicon	11.68	Water (distilled)	80.1
SrTiO ₃	310	BaTiO ₃	~ 1000

This trick for making larger capacitors does not work indefinitely. Every dielectric has a “dielectric strength,” the maximum tolerated value of the electric field inside that particular material. If the electric field inside the dielectric exceeds this value, the dielectric breaks down, which usually means a spark jumps across (or through) it. Exceeding the dielectric strength is a catastrophic failure, and usually results in “magic smoke” being released from the device in question.

3.7 Dielectrics in Electric Fields

Somehow or another, dielectrics inside a capacitor are able to dramatically increase the amount of charge that can be stored and *decrease* the voltage across the capacitor. Our explanation so far is that the dielectric itself partly charges, which both increases the amount of charge stored and decreases the net voltage. How does this work? In order to understand what is really going on, we have to think a bit about the microscopic nature of the dielectric.

The dielectric itself contains a large number of atomic nuclei and electrons, but overall there are equal numbers of positive nuclei and electrons to make the dielectric overall neutral. We have said that charges in insulators are not mobile, so electrons and nuclei remain bound. What, then, are the induced charges in the dielectric? Despite being bound, both electrons and nuclei in a dielectric can move very slightly without breaking their bonds. Electrons will attempt to move in the direction opposite the electric field between the plates, and nuclei will attempt to move in the opposite direction. As a result, tiny dipoles are formed inside the dielectric, which will be aligned along the direction of the electric field (see Figure 3.22). Random thermal motion of the atoms or molecules will limit the degree of alignment to an extent. In most materials the degree of alignment and the induced dipole strength are directly proportional to the external electric field. Essentially, an electric field induces a charge separation within the atom or molecule.

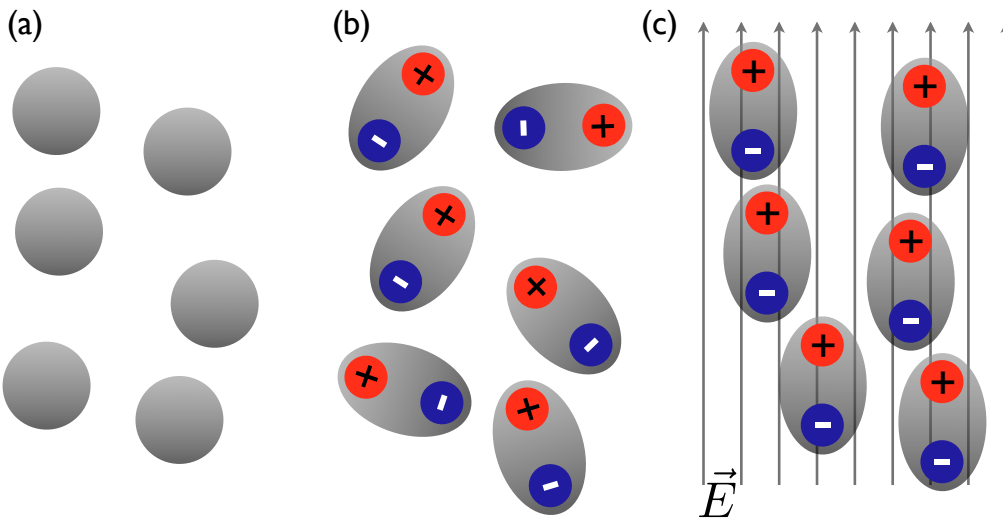


Fig. 3.22 (a) Atoms and many nuclei have no net charge separation without an electric field present. (b) Some “polar” molecules have a permanent electric dipole moment. Usually, these moments are oriented randomly from molecule to molecule, and the net moment is zero. (c) In an electric field, non-polar molecules can have an induced dipole moment, due to electrons and nuclei wanting to move in opposite directions in response to the field. Permanent dipoles remain bound, but can move or rotate slightly to align with the electric field. Either way, an overall dipole moment results.

Some molecules have a natural charge separation or dipole moment already built in, so-called *polar* molecules such as water or CO_2 . In these kinds of dielectrics, the built-in dipole moments are usually randomly aligned, and cancel each other out overall. An electric field exerts a torque on the dipoles, which tries to orient them along the electric field. Once again, random thermal motion works against this alignment, but the overall effect of the electric field is a net alignment, the degree of which is proportional to the applied electric field. Thus, in both polar and non-polar dielectrics, there is a net orientation of dipoles when an electric field is applied. The net dipole strength is far stronger in polar materials, and in the rest of the discussion below we will assume that our dielectric is made of polar molecules.

Now, what happens when we place our dielectric between two conducting plates? With no voltage applied between the plates, there is no electric field, and the tiny dipoles are randomly oriented, Fig. 3.23b. Once a voltage is applied to the plates, a constant electric field is created between them, which serves to align the dipoles, Fig. 3.23c. The net alignment of dipoles within a dielectric leads to the surfaces of the dielectric being slightly charged, Fig. 3.23. Within the bulk of the dielectric, dipoles will be aligned head-to-tail, and their electric fields will mostly cancel (Fig. 3.23a). At the surfaces of the dielectric, however, there will be an excess of positive charge on one side, and an excess of negative charge on the other. In this situation, the dielectric is said to be *polarized*. The dielectric is still electrically neutral on the whole, an equal number of positive and negative charges still exist, they have only separated due to the applied electric field.

These surface charges from the aligned dipoles look just like sheets of charge, in fact. This is the origin of our earlier statement that the dielectric picks up an induced charge on its surface – the part of the dielectric near the positive plate *does* build up a partial *negative* charge, and the part near the negative plate *does* build up a partial *positive* charge. What we missed in our initial analysis was the fact that in reality we are *aligning charges throughout the dielectric, even though only the surfaces have a net charge*. Not only are we storing energy in the surface charges, we are also storing energy by creating the aligned configuration of the dipoles! It took energy to orient them, so keeping them aligned is in a sense storing energy for later release. In a sense, we actually store energy in the whole volume of the dielectric, not just at the surfaces.

The electric field due to these effective sheets of charge is *opposite* that of the applied electric field, and thus the total electric field – the sum of the applied and induced field – is smaller than if there were no dielectric. Thus, the dielectric reduces both the applied voltage and the electric field.

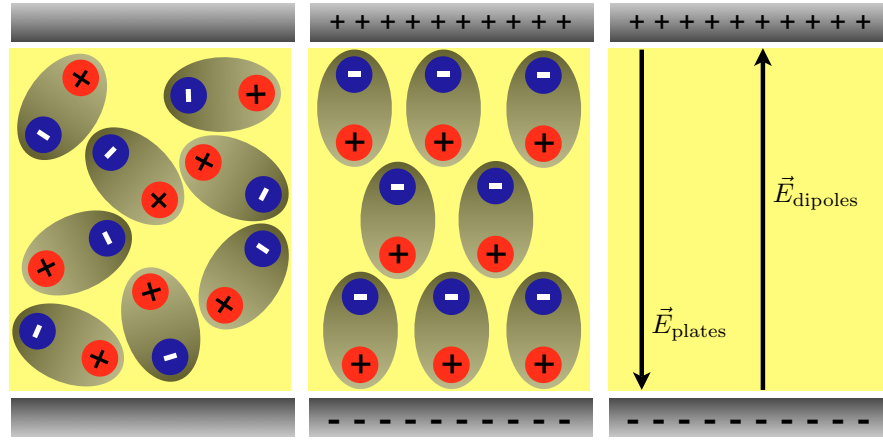


Fig. 3.23 (a) When no voltage is applied between the plates, the polar molecules align randomly, and there is no net dipole moment. (b) A voltage applied across the plates creates an electric field, which aligns the molecules. (c) The electric field from the voltage applied across the plates is partially canceled by the field due to the aligned dipoles.

The electric field due to the oriented dipoles inside the dielectric is usually proportional to the *total* electric field they experience:

$$E_{\text{dipoles}} = \chi_E E_{\text{total}} \quad (3.51)$$

where the constant of proportionality χ_E is called the *electric susceptibility*. It represents the relative strength of the dipoles within the material, or more accurately, how easily a material polarizes in response to an electric field. The total electric field the dipoles experience is not just the field due to voltage applied across the plates, but must also include *the field of all the other dipoles* as well:

$$E_{\text{total}} = E_{\text{plates}} - E_{\text{dipoles}} \quad (3.52)$$

$$E_{\text{total}} = E_{\text{plates}} - \chi_E E_{\text{total}} \quad (3.53)$$

$$(1 + \chi_E) E_{\text{total}} = E_{\text{plates}} \quad (3.54)$$

$$\Rightarrow E_{\text{total}} = \frac{1}{1 + \chi_E} E_{\text{plates}} \quad (3.55)$$

Thus, the field inside the plates is *reduced* by a factor $\frac{1}{1 + \chi_E}$ by the presence of the dielectric (χ_E is always positive). We already know that for a parallel plate capacitor, $\Delta V = Ed$, where d is the spacing between the plates, so we can also readily find the effect of the dielectric on the voltage between the plates:

$$\Delta V_{\text{total}} = \frac{1}{1 + \chi_E} E_{\text{plates}} d = \frac{1}{1 + \chi_E} \Delta V_0 = \frac{\Delta V_0}{\kappa} \quad (3.56)$$

here we again use ΔV_0 for the voltage on the plates without the dielectric. This result agrees precisely with Eq. 3.48, if we make the substitution $\kappa = 1 + \chi_E$, as we have in the last term in the equation above. We can go further and calculate the capacitance, just as we did for Eq. 3.50:

$$C = (1 + \chi_E) \epsilon_0 \frac{A}{d} = \kappa \epsilon_0 \frac{A}{d} = \kappa C_0 \quad (3.57)$$

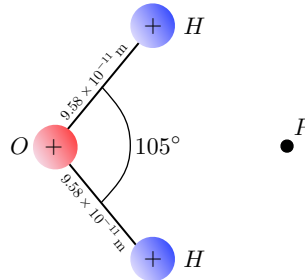
where C_0 is the capacitance without the dielectric present. Thus, our “dielectric constant” is simply related to the dielectric susceptibility, the ability of the dielectric to polarize in response to an electric field. This makes sense in a way – the more easily polarized the dielectric, the more easily it affects the capacitance. Also, since $\kappa = 1$ for vacuum, $\chi_E = 0$, which also makes sense as the vacuum is not polarizable (so far as we know). The result we obtain using this more sophisticated model is exactly the same as earlier, but now we have a plausible *microscopic* origin for the effect of dielectrics

in capacitors, and we know *why* the electric field and voltage are reduced, and the capacitance increased.

3.8 Problems

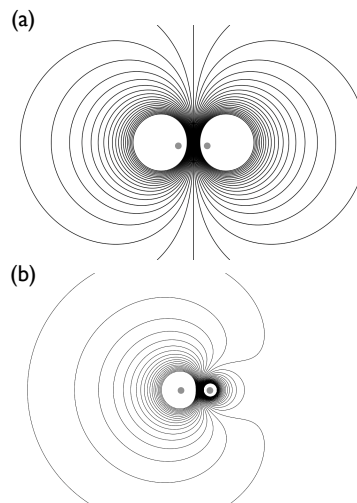
Solutions begin on page 258.

3.2. The distance between the oxygen nucleus and each of the hydrogen nuclei in an H_2O molecule is $9.58 \times 10^{-11} \text{ m}$, and the bond angle between hydrogen atoms is 105° . **(a)** Find the electric field produced by the nuclear charges (positive charges) at the point P a distance $1.2 \times 10^{-10} \text{ m}$ to the right of the oxygen nucleus. **(b)** Find the electric potential at P .



3.3. The figure below shows the **equipotential** lines for two different configurations of two charges (the charges are the solid grey circles). Which of the following are true?

- The charges in (a) are of the same sign and magnitude, the charges in (b) are of the same sign and different magnitude.
- The charges in (a) are of opposite sign and of the same magnitude, the charges in (b) are of the opposite sign and different magnitude.
- The charges in (a) are of the same sign and magnitude, the charges in (b) are of the opposite sign and the same magnitude.
- The charges in (a) are of the opposite sign and different magnitude, the charges in (b) are of the same sign and different magnitude.



3.4. An isolated conductor has a surface electric potential of 10 Volts. An electron on the surface is moved by 0.1 m. How much work must be done to move the charge? (e is the electron charge.)

3.5. An electron initially at rest is accelerated through a potential difference of 1 V, and gains kinetic energy KE_e . A proton, also initially at rest, is accelerated through a potential difference of -1 V, and gains kinetic energy KE_p . Is the electron's kinetic energy larger, smaller, or the same compared to the protons?

3.6. A parallel plate capacitor is shrunk by a factor of two in every dimension – the separation between the plates, as well as the plates' length and width are all two times smaller. If the original capacitance is C_0 , what is the capacitance after all dimensions are shrunk?

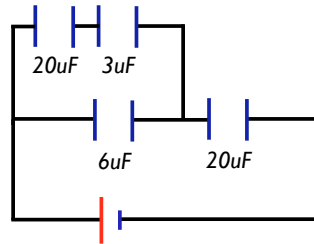
3.7. A capacitor with air between its plates is charged to 120 V and then disconnected from the battery. When a piece of glass is placed between the plates, the voltage across the capacitor drops to 30 V. What is the dielectric constant of the glass? (Assume the glass completely fills the space between the plates.)

3.8. Electrons in a TV tube are accelerated from rest through a potential difference of 2.00×10^4 V from an electrode towards the screen 25.0 cm away. What is the magnitude of the electric field, if it assumed to be constant over the whole distance? You may assume that the electron moves parallel to the electric field at all times.

3.9. A proton moves 1.5 cm parallel to a uniform electric field of $E = 240$ N/C. How much work is done by the field on the proton?

3.10. It takes 3×10^6 J of energy to fully recharge a 9 V battery. How many electrons must be moved across the 9 V potential difference to fully recharge the battery?

3.11. What is the effective capacitance of the four capacitors shown below?



3.12. Calculate the speed of a proton that is accelerated from rest through a potential difference of 104 V.

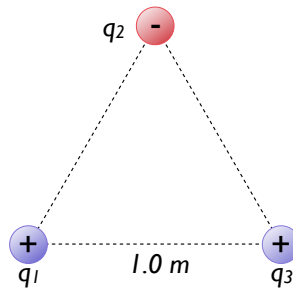
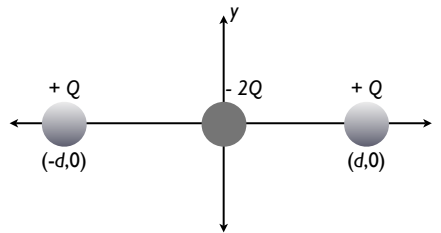
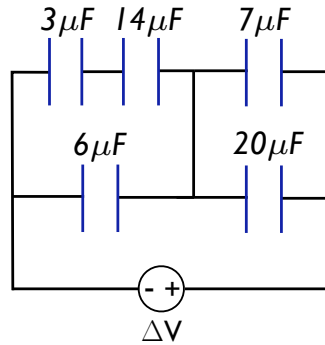
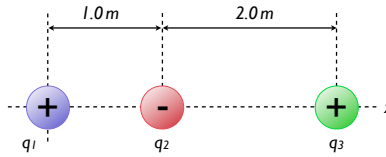
3.13. A proton at rest is accelerated parallel to a uniform electric field of magnitude 8.36 V/m over a distance of 1.10 m. If the electric force is the only one acting on the proton, what is its velocity in km/s after it has been accelerated over 1.10 m?

3.14. Three charges are positioned along the x axis, as shown below. All three charges have the same *magnitude* of charge, $|q_1| = |q_2| = |q_3| = 10^{-9}$ C (note that q_2 is negative though). What is the total **potential energy** of this system of charges? We define potential energy zero to be all charges infinitely far apart.

3.15. Two identical point charges $+q$ are located on the y axis at $y = +a$ and $y = -a$. What is the electric potential for an arbitrary point (x, y) ?

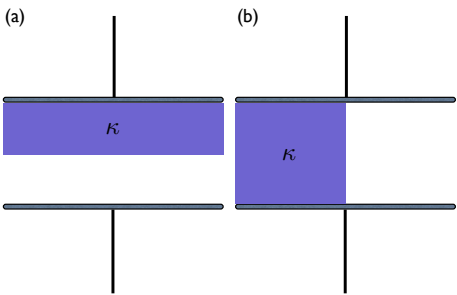
3.16. What is the equivalent capacitance for the five capacitors below?

3.17. The charge distribution shown is referred to as a linear quadrupole. What is the electric potential at an arbitrary point on the y axis?



3.18. Three charges are arranged in an equilateral triangle, as shown below. All three charges have the same *magnitude* of charge, $|q_1| = |q_2| = |q_3| = 10^{-9} \text{ C}$ (note that q_2 is negative though). What is the total **potential energy** of this system of charges? Take the zero of potential energy to be when all charges are infinitely far apart.

3.19. A parallel plate capacitor has a capacitance C when there is vacuum between the plates. The gap between the plates is half filled with a dielectric with dielectric constant κ in two different ways, as shown below. Calculate the effective capacitance, in terms of C and κ , for both situations. *Hint: try breaking each situation up into two equivalent capacitors.*



Chapter 4

Current and Resistance

*A little learning is a dangerous thing; Drink deep, or taste not
the Pierian spring – Alexander Pope (Essay on Criticism)*

Abstract Current is something that we use and hear about every day, but few of us stop to think about what it really is. What is an electric current? An electric current is nothing more than the net flow of charges through some region in a conductor.

4.1 Electric Current

If we take a cross section of a conductor, such as a circular wire, *an electric current is said to exist if there is a net flow of charge through this surface*. The *amount* of current is simply the *rate* at which charge is flowing, the number of charges per unit time that traverse the cross-section. Strictly speaking, we try to choose the cross-sections for defining charge flow such that the *charges flow perpendicular to that surface*, somewhat like we did for Gauss's law. Figure 4.1 shows a cartoon depiction of how we define current.

Current is a flux of charge through a wire in the same way that water flow is a flux of water through a pipe. As we shall see, this is a reasonable way to think about electric circuits as well – current always has to flow somewhere, and you don't want an open connection any more than you would want an open-ended water pipe. Voltage is more like a pressure gauge – you can have a voltage even when nothing is flowing, it just means there is the *potential* for flow (nerdy pun intended).

If a net amount of charge ΔQ flows perpendicularly through a particular surface of area A within a time interval Δt , we define the electric current to be simply the amount of charge divided by the time interval:

Electric Current: if a net amount of charge ΔQ flows perpendicularly through a surface of area A in a time interval Δt , the electric current I is:

$$I \equiv \frac{\Delta Q}{\Delta t} \quad (4.1)$$

In other words, current is charge flow per unit time.

This represents a conservation law as well. Charge can neither be created or destroyed. If we have some steady stream of charge pouring into of a region of fixed volume, then the charge density inside would continually grow (tending toward infinity!) if there were not also some compensating flow of charges out of the volume. Putting it the other way around, if a steady stream of charges were *leaving* the fixed volume, the charge density would also become infinitely large if there were not some other source of charges to replace those lost. But creating charges out of thin air is the one thing that definitely will not happen! Therefore, the *change* in the total number of charges in a

volume at any time has to equal the net flow of current through that volume, otherwise we would require spontaneous generation of charge.¹

Units of electric current I : Coulombs per second [C/s] or Amperes [A].

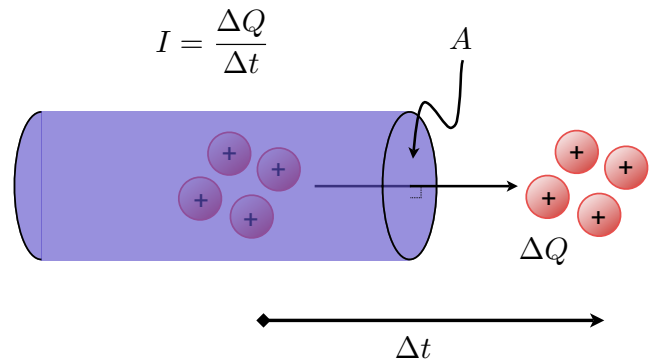


Fig. 4.1 A net amount of charge ΔQ goes through a surface (perpendicularly) of area A in a time Δt . The electric current through the surface A is $I = \Delta Q / \Delta t$.

We should get one thing out of the way right off the bat: the definition for the current *direction* is somewhat confusing. The historical definition is that **current flow is defined as the direction that positive charges would be moving**. Of course, at this point we know that usually it is really electrons doing all the work,² but the definition of electric current came before we knew about electrons. Figure 4.2 is a small exercise to help you understand the calculation of current.

In our previous investigations of electrostatics, we showed that the electric potential (“voltage”) must be the same everywhere in a conductor. **This is not true when currents are flowing**. When currents flow, we no longer have electrostatic equilibrium – its very definition was that no charges were moving! When currents flow, the electric potential continuously decreases from the point of the current’s source to its sink.³

Direction of Current Flow: the direction of current flow is defined as the direction of net *positive* charge flow. The flow of *electrons*, which are usually responsible for the current, is opposite the direction of current flow due to their negative charge. It is a source of much confusion. See <http://xkcd.com/567/>.

4.2 Getting Current to Flow

Current in real conductors is due to the (net) motion of microscopic charge carriers. How much current flows depends on the average speed of these charge carriers, the number of charge carriers

¹ We have waved our hands a bit here, since we should talk about current *density* and charge *density*, but the essential points are the same.

² In certain ionic conductors, of the sorts important for batteries for instance, the flow of positive ions does play a key role.

³ Except in superconductors, where the voltage is *always* zero. That is another story entirely, though.

per unit volume (the density of charge carriers), and how much charge is carried by each. But how do we get charges to flow in the first place?

Recall what happened when we *charged* conductors (Sec. 2.2.1 and 2.2.2). If we take a piece of conductor, and deposit some charge on it, those charges will flow to distribute themselves evenly on the conductor's surface. If we then connect the conductor to ground, the excess charge will flow down through the wire – this is a current! So this is one way to make electric currents – put some excess charge on an isolated conductor, and take it away by connecting it to ground. Of course, this is a somewhat cumbersome method . . .

More generally, what are we doing when we put excess charge on the conductor? We are *changing its electric potential* relative to the ground. Let's say we take electrons away from our conductor, which makes its potential positive relative to ground. Once we connect the conductor to a ground wire, electrons in the ground wire are attracted to our conductor and its relatively positive potential, and they flow up from the ground into the conductor until the conductor is at the same electric potential as the ground. Any time we can make one conductor, or part of a conductor, at a different potential than another, current will try to flow between them. Try being the operative word.

Whether any current *does* is another story. It depends on how we connect conductors which are at different potential. If we make the potential difference big enough, though, the electrons will always find a way to flow and make a current. Think about how this explains “static” shocks, or the sparks from a Van de Graaff generator, for example.

So this is our answer – in order to get a net flow of charges, we need to provide a potential difference (voltage!).⁴ The presence of a voltage gives rise to an electric field across the conductor, which in turn causes an electric force, which accelerates the charges. The effectiveness of a potential difference to cause a current depends on the density of charge carriers, their average speed, and microscopic properties of the conductor itself.

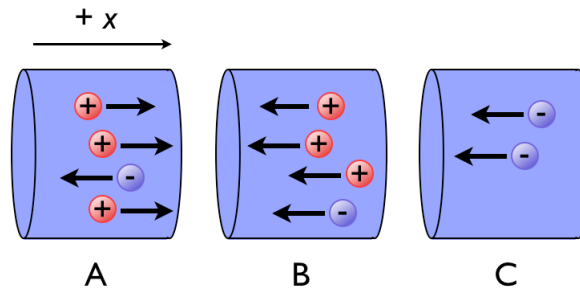


Fig. 4.2 (a) There are three positive charges moving to the right(+x direction), which count as +3, and one negative charge moving to the left (−x direction). A negative charge moving backwards is the same as a positive charge moving forwards, which counts as $-(-1) = +1$, so the total relative current is $3 + 1 = 4$. (b) Three positive charges moving to the right count as +3, and one negative charge moving left counts as +1, for a total relative current of $-3 + 1 = -2$. (c) Two negative charges moving to the left count as $-(-2) = 2$, so the total relative current is 2.

The free charges in conductors are extremely numerous and fairly mobile, as we already know. Inside a normal conductor, like copper, there is a fantastic density of charge carriers, $\sim 10^{22}$ electrons per cm^3 ! [15] So many, in fact, that they continuously scatter off of each other and the fixed atoms in the conductor (about once every 10^{-14} sec or so, even in a good conductor!). Typical drift speeds in copper are $\sim 10^{-3} - 10^{-4}$ m/s for moderate electric fields, compared to the speed of random thermal electron motion of $\sim 10^5$ m/s. [16] Any particular charge carrier has a hard time getting anywhere. Even though the charges are mobile, and able to move at fantastic speeds, the time it takes to actually get anywhere is quite a bit longer than expected. A bit like pachinko.

One result of all these collisions is that the carriers in, *e.g.*, copper, cover huge *distances* but have a very small *displacement* – most of their movement is wasted, and they end up close to where they started out, so their *net* velocity is very small. Even when we apply a potential difference, the net

⁴ From now on, we will interchangeably use the phrases “potential difference” and “voltage.” From our point of view, they are the same thing.

flow of charges is more sluggish than we might expect, due to all these collisions. It is a bit like trying to get 90,000 people out of Bryant-Denny stadium – even though there is a lot of commotion inside, the *net* flow of people out the exits is disappointingly small.

The *net* velocity of charge flow⁵ we call the *drift velocity*, v_d . In normal conductors, like copper, this drift velocity is more or less proportional to the voltage applied, a point which we will explore in depth presently.

4.3 Drift Velocity and Current

Our conceptual physical picture of current in conductors is basically complete. A voltage induces an electric field, which gives the carriers a net velocity in one direction, which is an electric current. This drift motion along the electric field is superimposed upon the random thermal motion of the charge carriers (just like the random thermal motion in an ideal gas). From here, all we need to do is apply our knowledge of electric forces and fields and kinematics to come up with a relationship between current, field, and voltage. There are many steps yet, but none too difficult.

So first: given a drift velocity v_d , through a conductor of cross section A , what is the current? The number of charges that flow through our cross section A in the time Δt is just the free charge which is physically close enough to *reach* the surface A within that time. Those charges close enough must cover the distance Δx in the time Δt . Since the average speed of the carriers is v_d , then we must have $\Delta x = v_d \Delta t$. In other words, in a certain amount of time Δt , the charge carriers will have, on average, covered a distance Δx , such that $v_d = \Delta x / \Delta t$. This is illustrated schematically in Figure 4.3.

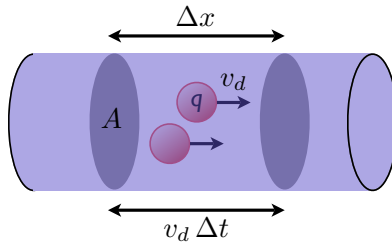


Fig. 4.3 A small piece of a conductor of cross-sectional area A . The charge carriers move with a speed v_d , and are displaced by $\Delta x = v_d \Delta t$ in a time interval Δt . The number of carriers in a section of length Δx is, on average, $nAv_d \Delta t$, where n is the density of the charge carriers.

The number of charges which cross the surface A , those close enough to reach it in a time Δt , is just the number contained within the volume $A \cdot \Delta x$, or $Av_d \Delta t$. A bit more mathematically, we can write this:

$$\text{number of charge carriers} \equiv N = \text{charge density} \times \text{volume} \quad (4.2)$$

$$= \text{charge density} \times \text{area} \times \text{distance covered in time } \Delta t \quad (4.3)$$

$$= nA\Delta x = nAv_d \Delta t \quad (4.4)$$

Here we have used n to represent the number of charges per unit volume, the carrier density. The total amount of charge is the number of charge carriers times how much charge each one carries, which we'll call q . The current then is just the total amount of charge, Nq divided by the total amount of time, Δt :

Current flow related to drift velocity:

⁵ Distinct from and not to be confused with the random thermal motion, see below.

$$I = \frac{\Delta Q}{\Delta t} = \frac{Nq}{\Delta t} = \frac{nqAv_d\Delta t}{\Delta t} = nqAv_d \quad (4.5)$$

here A is the cross-sectional area of the conductor, n the number of charge carriers *per unit volume* (their density), q is the charge on each, and v_d is their drift velocity.

We can see that the drift velocity and resulting current are larger when the carriers carry more charge q , or when their mass is small. However, it would be nice to have expressions that didn't directly involve the cross-sectional area of the conductor, so we can calculate general properties independent of any particular conductor shape or size. For this reason, it is common to introduce *current density*, J , which is just the current per unit area. Rewriting Eq. 4.5 in terms of current density, we come up with a simpler and more general expression:

Current density related to drift velocity:

$$J \equiv \frac{I}{A} = nqv_d \quad (4.6)$$

Now we can calculate the current density for any given material of *arbitrary geometry*, and later specify a cross-sectional area to determine absolute currents.

The units of current density: J is current per unit area, and has units of amperes per square meter [A/m²].

4.4 Resistance and Ohm's Law

From Equation 4.5, we saw that the current through a conductor can be expected to scale with the drift velocity. You might expect that the effect of increasing the applied voltage across a conductor ΔV is to increase the drift velocity. This is basically true, but justifying that statement will require a few more steps.

More accurately, the presence of a potential difference between two points on the conductor means that those two points are at different potential energies. Recall that negative charges want to move from regions of lower potential to regions of higher potential. In a conductor, even when a current flows, the charges like to spread out as evenly as possible. This even and moving distribution of charge gives rise to a uniform *electric field*. If the potential difference ΔV is applied over some distance l , and the electric field is uniform, we know from Equations 3.12 and 3.13 that the electric field along the length of the conductor must be given by:

$$E = \frac{\Delta V}{l} \quad (4.7)$$

The presence of the electric field causes an acceleration of the charge carriers:

$$a = \frac{F_e}{m} = \frac{q}{m}E \quad (4.8)$$

Thus the acceleration of the charge carriers depends only on the electric field and their charge-mass ratio, q/m , about 1.76×10^{11} C/kg for electrons. In order to figure out how much current will flow for a given potential difference, we need to find a way to take into account the dissipative effect of all the collisions the carriers are constantly undergoing. In a sense, the collection of charge carriers is a bit like an ideal gas, and our treatment here is reminiscent of an ideal gas law derivation. The

analogy is a close one (and useful if you are a chemist) – the innumerable electrons in a conductor are often called an *electron gas*.

4.4.1 Drift Velocity and Collisions

If we assume the charge carriers are electrons, of mass m_e (and charge $-e$), then each has an *average* momentum $p = m_e v_d$.⁶ We expect on average that each collision an electron experiences will completely destroy all forward momentum – they are stopped cold by every single collision. This makes some sense, since most of the collisions will be with the atoms making up the conductor, which are very heavy compared to electrons, rather than with other electrons. If all forward momentum is destroyed, then the electron is left with only its random thermal motion. If there were no electric force present to accelerate the electrons, the random thermal motion of all the electrons will cancel out, and there is no net flow or current.

We can easily find the thermal velocity of the carriers just like we do for an ideal gas – the thermal energy of the electrons is $\frac{3}{2}k_B T$, where k_B is Boltzmann's constant, and we equate this to the carriers' kinetic energy:

$$\frac{3}{2}k_B T = \frac{1}{2}m v_{th}^2 \quad (4.9)$$

$$\Rightarrow v_{th} = \sqrt{\frac{3k_B T}{m}} \sim 10^5 \text{ m/s (at 295 K)} \quad (4.10)$$

Here we use v_{th} to specify the thermal velocity distinctly from the electric-field-induced *drift* velocity. As it turns out, the thermal velocity typically greatly exceeds the drift velocity (by ten million times or so!) – the acceleration of the carriers by the electric field induces only a tiny velocity compared to that given by the random thermal motion of the carriers. Again, this is what leads to carriers covering huge *distances* but having very small *displacements*. The overall motion is terribly chaotic, and even fairly large electric fields only alter the carrier velocity in conductors by parts per million at best. Still, the random thermal velocities do not contribute to the electric current,⁷ it is only the *tiny* field-induced drift velocity that gives rise to electric current.

Boltzmann's Constant:

$$k_B = 1.38 \times 10^{-23} \text{ J/K} = 8.62 \times 10^{-5} \text{ eV/s} \quad (4.11)$$

We should also keep in mind that the collisions the carriers undergo are not continuous, but happen one after another with some average time between them τ .⁸ In that time interval, the electron loses its momentum $m_e v_d$ due to a collision, and thereafter regains it due to the action of electric field present, only to lose it again about τ seconds later. As stated above, the presence of the electric force F_e gives the electron an acceleration $a = F_e/m_e$, which allows it to regain its former drift velocity. From kinematics, we would expect a *mean* displacement $v_d \approx a\tau$.⁹

The starting and stopping motion of the carriers gives us an average rate at which the electrons are losing momentum due to the collisions and associated impulse forces. We can straightforwardly find this momentum change as:

⁶ Since we are talking about zillions of collisions in every possible random direction, there is no need to carry around the vector baggage here. We will just deal with scalar magnitudes.

⁷ They do give rise to electrical *noise*, however.

⁸ For Cu, we can estimate[15] $\tau \sim 2 \times 10^{-14} \text{ s}$.

⁹ Depending on the method of derivation, there may be a factor of 2 in this expression, but the physics is the same.

$$\left(\frac{\Delta p}{\Delta t}\right)\bigg|_{\text{loss}} = \frac{m_e v_d}{\tau} \quad (4.12)$$

Remember, $\Delta p/\Delta t$ is also a *force* – we are still dealing with kinematics, even though we have involved electricity. Once the scattering event is over, the electron regains momentum through the action of the electric *force* caused by the electric field. We can easily write down the momentum gained up until the next collision:

$$\left(\frac{\Delta p}{\Delta t}\right)\bigg|_{\text{gain}} = F_e = qE = -eE \quad (4.13)$$

Now, the total momentum loss has to equal the total momentum gain for there to be a steady state. If this were not true, the momentum would quickly build up, and the whole wire would start to move! So we must impose conservation of momentum:

$$\left(\frac{\Delta p}{\Delta t}\right)\bigg|_{\text{loss}} = \left(\frac{\Delta p}{\Delta t}\right)\bigg|_{\text{gain}} \quad (4.14)$$

$$\frac{m_e v_d}{\tau} = -eE_x \quad (4.15)$$

$$v_d = \frac{-e\tau}{m_e} E \quad (4.16)$$

Now we have an expression for the average drift velocity of electrons flowing along the wire, in terms of the average time between carrier collisions:

Drift velocity and electric field:

$$v_d = \frac{-e\tau}{m_e} E \quad (4.17)$$

here τ is the mean time between collisions, and E the electric field.

The minus sign makes sense here, by the way. Since electrons are negatively charged, they move in the opposite direction that the electric field lines point. It is also reassuring that the drift velocity *increases* as τ increases, since more time between collisions means more time spent accelerating, and that in principle lighter carriers would have a higher velocity since they are more easily accelerated. Finally, the proportionality with the electric field is what we expect.

For typical metals, we can estimate^[16] drift velocities of about 5×10^{-3} m/s for a moderate electric field of 1 V/m, about *eight orders of magnitude below the thermal velocity!* Really, the effect of the electric field is quite negligible in one sense, but it has profound consequences.

Another way of seeing this is as an application of Newton's laws – $\Delta p/\Delta t$ is nothing more than force, and the equations above are also in some sense a force balance between the electric force, and the impulse force due to the collision.

4.4.1.1 Mean Free Path and Mobility

Instead of dealing with the mean time between collisions, we could just as easily have started with the mean *distance* that electrons travel before undergoing a collision.¹⁰ This quantity is known as the *mean free path*, λ_{mfp} , and it has essentially the same meaning as it does in the kinetic theory of gasses. The shorter the time between collisions, the smaller the mean free path, and vice versa. The mean time and mean free path are easily related through kinematics:

$$\lambda_{\text{mfp}} = \tau(v_d + v_{th}) \approx \tau v_{th} \quad (4.18)$$

¹⁰ Here we do mean the *distance* covered between collisions, not the *displacement*

Here we are considering the total distance covered not just the net displacement, so we need to use the *total* velocity, $v_d + v_{th}$. For the last relationship, we have made use of the fact that $v_{th} \gg v_d$. What this means is that the mean distance (and mean time) between collisions does *not* really depend on the applied electric field, but really *only* comes from the random thermal motion of the carriers.

As another aside, the proportionality constant between drift velocity and electric field in Eq. 4.17 is often called the carrier *mobility*, which is just what it sounds like. In this case, we write $v_d = \mu E$, where μ is the mobility:

Carrier mobility:

$$v_d = \mu E \quad \text{with} \quad \mu = \frac{q\tau}{m} \quad (4.19)$$

where q is the charge of the carrier, m its mass, and v_d its drift velocity. Mobility relates the drift velocity of carriers to the applied electric field. The units for mobility are $\text{m}^2/\text{V}\cdot\text{s}$.

From the units of μ ($\text{m}^2/\text{V}\cdot\text{s}$) and E (N/C or V/m), we can see that mobility is a quantity that tells us how far a charge is able to move per second per unit of electric field (V/m). Now we have a nice expression for *exactly* what we mean by mobility, rather than just a vague notion.

4.4.2 Current, Electric Field, and Voltage

From here, the rest is easy, we already derived Eq. 4.6 above, relating drift velocity and current density! Plugging Eq. 4.17 into Eq. 4.6:

Relation between current density and electric field:

$$J = \frac{I}{A} = nqv_d = -ne \frac{-eE\tau}{m_e} = \frac{ne^2\tau}{m_e} E \equiv \frac{1}{\rho} E \quad (4.20)$$

where A is the cross-sectional area of the conductor, E is the electric field, q is the charge per electron of $-e$, n is the density of electrons in the material, τ is the average time between electron collisions, and ρ is a constant of proportionality known as the **resistivity** of the conductor.

In the end, it turns out that current density (or current) and electric field are simply proportional. We could almost have guessed this in the first place, but now we have a formal relationship between the two, and we even know the constant of proportionality. In this regard, we have sneakily defined a new quantity ρ , the *electrical resistivity*, which is the constant of proportionality between current density and electric field:¹¹

Resistivity ρ for “free electrons”

$$\rho = \frac{m_e}{ne^2\tau} \quad (4.21)$$

where n is the density of electrons in the material, and τ is the average time between electron collisions.

¹¹ We will use a slightly different rho character for resistivity, ρ , to distinguish it from the one we use for *mass density*, ρ .

Units of resistivity: ρ has units of Volt-meters per amp [V·m/A], or Ohm-meters, [$\Omega \cdot m$].

Resistivity represents the effectiveness with which a given electric field or potential difference causes a current to flow, and is a (strongly) material-dependent property – it is a measure of the resistance of a material to current flow. We see that the resistivity gets larger when the time between electron collisions gets smaller, just as we would expect, and it gets larger when we increase the density of free carriers. We will return to the resistivity of various materials shortly. We can go further in our analysis by noting that the potential difference and electric field are simply related by Eq. 4.7, $E = \Delta V/l$, which leads us to:

$$J = \frac{I}{A} = \frac{1}{\rho} \frac{\Delta V}{l} \quad \text{or} \quad \Delta V = \frac{\rho l}{A} I = \rho l J \quad (4.22)$$

In other words, we find $J \propto I \propto \Delta V$ – the current flow in a conductor is proportional to the magnitude of the applied voltage, and the amount of current one gets for a particular applied voltage depends on the conductor's resistivity and geometry. We can make this simpler by introducing a new constant of proportionality $R = \frac{\rho l}{A}$. This, along with the definition of current density ($J = I/A$), will allow us to relate I and ΔV directly. **This new constant of proportionality R between I and ΔV is known as the resistance of the conductor**, and it allows us to connect ΔV and I in the traditional form known as Ohm's¹² law:

Ohm's Law:

Current through and voltage across a conductor are proportional, the constant of proportionality is the **resistance** of the conductor.

$$\Delta V = IR \quad \text{or} \quad I = \frac{\Delta V}{R} \quad \text{or} \quad R = \frac{\Delta V}{I} \quad (4.23)$$

R is different for different conductors and geometries.

4.4.3 Resistance

The presence of resistance does not mean that conductors “lose” current, greater resistance just lessens the ability of a given ΔV to create a current. **Resistance is somewhat analogous to viscosity for a liquid or kinetic friction – it just makes it harder for charge to flow, a larger resistance requires a larger potential difference for the same current.**

The units for resistance R are volts per ampere [V/A], or Ohms [Ω]

Not all conductors follow Ohm's law, it is only valid for certain materials (it is valid for most metals). Those conductors that *do* follow Ohm's law give a specific resistance R for a given ΔV and I , and are called “ohmic.” Those that do not follow Ohm's law are simply “non-ohmic.” A **Resistor** is a circuit element made out of such an ohmic conductor, which provides a specific value of R for use in a circuit. The circuit diagram symbol for a resistor is shown below, and Fig. 4.4 shows a picture of a common type of resistor.

¹² After Georg Simon Ohm (1789–1854) a German physicist who first found the relationship between current, voltage, and resistance.



Circuit diagram symbols for resistors:  or 

Fig. 4.4 A picture of a common type of resistor.



A non-ohmic device is characterized by a current-voltage plot that is *not* a straight line, current and voltage are not proportional. What this means is that our simple model of relatively free electrons drifting along in a conductor does not apply in those cases. Often this means that the electrons are interacting in a more complicated way with other electrons or atomic nuclei. One common example non-ohmic behavior is a device called a *diode*, which we will encounter in our laboratory experiments shortly. A diode is a semiconductor device that only lets charge pass through in one direction – in essence it is a “check valve” for electrons. In the “reverse” direction, ideal diodes do not allow current to flow at all. In the “forward” direction, diodes allow current to flow as soon as a certain threshold voltage has been reached. Perhaps without knowing it, you encounter diodes every day, in the form of Light Emitting Diodes (LEDs) commonly found on electronic panels (and these days in newer traffic lights). In LEDs, the onset of current flow beyond the threshold voltage is commensurate with the onset of light output.

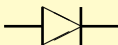
Circuit diagram symbol for a diode: 

Figure 4.5 shows current (I) as a function of voltage (V) for a $200\,\Omega$ resistor (Ohmic), and a red light-emitting diode (LED; non-Ohmic). For an Ohmic device, the slope of an I vs. V curve is $\Delta I/\Delta V = 1/R$. The higher the slope on the plot, the lower the resistance. The resistor shows a constant slope, as expected, while the diode shows a slope which dramatically decreases at higher applied voltages – the resistance decreases dramatically as V increases. Note that this measurement is for a “forward biased” LED, the threshold voltage of $\sim 1.5\text{ V}$ is clearly visible. For negative voltages, essentially zero current flows through the LED (and there is no light output).

4.4.4 Resistors as Circuit Elements

What good is a device like a resistor that apparently restricts current flow? In Sect. 3.5, we realized that in order to place different objects at different potentials – thus creating current flow – we need a voltage source. What happens if we want to put several objects at several different voltages? This is one thing resistors can do, they can be used to control voltage levels. In this capacity, you have probably used dozens of resistors already today – dimmer switches, volume controls, and a million other things. Another useful function of resistors is to divide a single current up in to two or more, which we will learn about in the next chapter.

In order to start to see the utility in resistors, we should think a bit about what happens when we source a current through a resistor. In Fig. 4.6a, there is a current I in the resistor of value R . Ohm’s law (Eq. 4.23) tells us that the potential difference due to the current I in the resistor R has to be $\Delta V = IR$.

When thinking of a resistor as an actual, physical component, what this means is that *the voltage difference between the two ends of the resistor is $\Delta V = V_b - V_a$* . When charges flow from point a to point b in the resistor, they *lower* their potential energy by ΔV . This might be more clear if we

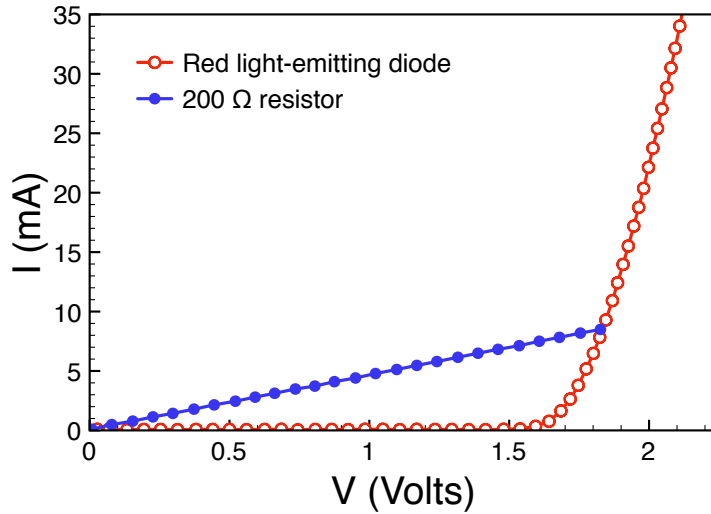


Fig. 4.5 I vs. V for a resistor and LED. The resistor clearly displays ohmic behavior, while the diode is highly non-ohmic, as evidenced by its nonlinear $I(V)$ characteristic. The slope of an $I(V)$ plot is $1/R$, from Ohm's law, so the higher the slope, the lower the resistance. For the diode, which is "forward biased" (current in the direction that gives light), the resistance decreases substantially at higher applied voltages. (These curves were actually measured with the new laboratory hardware you have been using!)

purposely ground point a , Fig. 4.6b, which defines $V_a = 0$. Now charges start out at zero potential, and lose ΔV after traversing the resistor, and end up with potential $V_b = -\Delta V$. Similarly, if we ground point b , Fig. 4.6c, then $V_b = 0$, and $V_a = \Delta V$. What we have just figured out is one of the basic functionalities of a resistor – controlling the relative voltages between points a and b when a current is flowing.

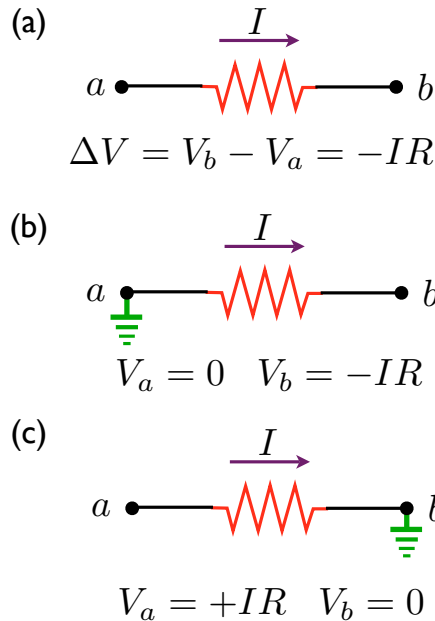


Fig. 4.6 (a) When a current I is sourced through a resistor R , the voltage difference between the ends of the resistor is $\Delta V = IR$. Alternatively, when a voltage ΔV is applied across the ends of the resistor, a current $I = \frac{\Delta V}{R}$ flows. (b) Grounding point a on the resistor means $V_a = 0$, and thus $V_b = -IR$. (c) Similarly, grounding point b means that $V_a = IR$. Resistors can thus be used to control voltage at specific points in a circuit, or the amount of current flow.

Incidentally, this all still works if we are using a voltage source instead of a current source. If instead we applied a voltage difference of ΔV between points a and b of the resistor using a voltage source, the potential difference would create a current $I = \frac{\Delta V}{R}$, in accordance with Ohm's law. Whether we source constant voltage, constant current, or some combination of the two, Ohm's law is still valid for resistors.

4.4.5 Resistivity of Materials

So far, we have mathematically derived the expected relationship between current, voltage, and electric field. We have even found a way to relate the proportionality constants, resistance and resistivity, to materials properties like the mean time between carrier collisions and mean free path. Do the dependencies we found make any sense, though, and how do we relate this to what we will actually measure in the lab? What resistance should we find for a particular copper wire, for example?

We can figure out if what we have makes sense qualitatively. From the discussion above, we can see that the drift velocity, and resulting current, get larger if we apply our potential difference between points very close together (make l small). We might also expect that the current depends on how big the cross sectional area A of the conductor is – the larger A is, the more charge carriers per unit time will be able to flow through comfortably.

We would expect that for a given ΔV , applied over a conductor of length l , that $I \propto A$ and $I \propto l^{-1}$. This mostly makes sense (and it is what we have derived above) – we need thick wires to carry large currents, and for long wires we need larger ΔV to make the same current flow. So how do we find out the “intrinsic” resistance of a material, independent of size? In fact, this is exactly what the resistivity ρ is. If we know the resistivity of a conductor, and its dimensions, we can calculate the expected resistance. And vice versa – if we know the resistance and dimensions of a conductor, we can find its resistivity if we recall the definition of resistance based on Eq. 4.22.

Resistivity, resistance, and geometry:

$$\rho = \frac{RA}{l} \quad \text{or} \quad R = \frac{\rho l}{A} \quad (4.24)$$

where ρ is the material’s resistivity, R is the resistance of the conductor, A is the cross-sectional area of the conductor, and l is its length.

Resistivity is material dependent. Copper, for example, is a better conductor than steel, which is one reason why we use it for the vast majority of wiring (also, it is reasonably cheap!). Resistivity does depend on extrinsic parameters, however. For given samples of, *e.g.*, copper, the resistivity can vary wildly depending how pure the copper is, its microstructure, and many other factors. Comparing resistivities between different materials absolutely is only truly valid for extremely pure, perfect crystals at low temperatures. One can make a very dirty sample of copper that is a worse conductor than steel at room temperature, for example. The resistivity for many common conducting materials is listed in Table 4.1.

Table 4.1 Resistivity of common materials at $T_0 = 20^\circ\text{C}$ [17]

Material	ρ_0 [$\Omega \cdot \text{m}$]	α_ρ [$^\circ\text{C}^{-1}$]
Silver	1.59×10^{-8}	0.0038
Copper	1.72×10^{-8}	0.0039
Gold	2.44×10^{-8}	0.0034
Aluminium	2.82×10^{-8}	0.0039
Iron	1.0×10^{-7}	0.005
Platinum	1.1×10^{-7}	0.00392
Lead	2.2×10^{-7}	0.0039
Carbon	3.5×10^{-5}	-0.0005
Silicon	6.4×10^2	-0.075
Glass	$10^{10} - 10^{14}$	nil
Hard rubber	$\sim 1 \times 10^{13}$	nil
Quartz (fused)	8×10^{17}	nil
Teflon	$\gtrsim 1 \times 10^{22}$	nil

4.4.6 Variation of Resistance with Temperature

The resistivity ρ , and therefore the resistance R , of a conductor depends on many factors. One primary consideration is the purity and morphology of the conductor, as discussed briefly above. Another primary factor is temperature. For most metals, resistivity *decreases* as temperature *decreases*. In fact, many use this as a (loose) definition of “metallic behavior” when discussing resistivity.

In conductors, this can be qualitatively understood in simple terms. Much of the resistance of a conductor comes from the myriad collisions between the electrons and the atoms of the conductor. As the temperature of the conductor increases, its constituent atoms vibrate with greater amplitudes. As a result, the electrons find it increasingly difficult to navigate through the conductor at higher temperatures. One way to envision this is to view a plucked guitar string. As the string vibrates, it moves faster than your eye can keep up with, and it appears to be larger and fuzzier. As the temperature increases, the atoms appear slightly “larger and fuzzier” from the electron’s point of view. This increases the scattering of the electrons, the result of which is increased resistance.

Over a limited temperature range, the resistivity of many conductors increases linearly with temperature, according to:

Temperature variation of resistivity in a conductor:

$$\rho = \rho_0 [1 + \alpha_\rho (T - T_0)] \quad (4.25)$$

where α_ρ is called the **temperature coefficient of resistivity**. The values of ρ_0 and α_ρ are listed for many materials in Table 4.1.

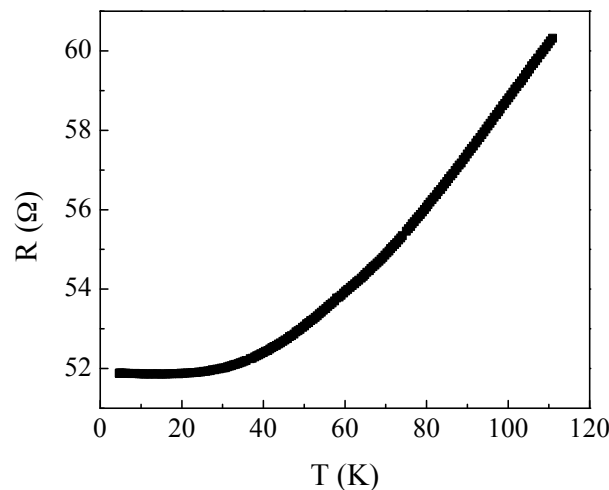


Fig. 4.7 Resistance vs. temperature for a thin film of Cobalt between 4.2 and ~120 K. Above ~80 K, resistance is roughly linear with temperature, as implied by Eq. 4.25.

A positive value for α_ρ is expected for metals, based on increased vibration of atoms at higher temperatures. Certain materials obey this empirical relationship far better than others. For platinum, it is an exceptionally good approximation over a wide temperature range. As a result, the resistance variation of carefully fabricated platinum pieces is often used to actually *measure* temperature. Figure 4.7 shows the roughly linear behavior of resistance with temperature for a thin film of Cobalt. At the lowest temperatures, the resistance varies very little – as thermal fluctuations become tiny, the dominant contribution to resistance is actually imperfections and impurities. At higher temperatures, above ~80 K in this case, the observed resistance is roughly linear with temperature.

A negative value of α_p is observed for semiconductors. This is because conduction in semiconductors is fundamentally different from that in metals. At lower temperatures, charges in semiconductors are weakly bound to host atoms and not very mobile, leading to a high resistivity. At higher temperatures, random thermal motion of the charges overcomes this weak bonding, and the charges actually become more mobile at higher temperatures. Based on Eq. 4.19, we would expect a larger mobility to lead to a larger drift velocity for the same applied electric field (or voltage), and hence a larger current. A larger current for the same voltage or electric field means a lower resistance, and the resistance of semiconductors *decreases* as temperature increases.¹³

4.5 Electrical Energy and Power

Now we know that when a potential difference is applied between the ends of a conductor, an electric field is set up within the conductor, creating an electric force on the electrons which drives a current. This decreases the potential energy of the carriers, and increases their kinetic energy. The repeated collisions, which result in the relatively small drift velocity, transfer (kinetic) energy from the carriers to the conductor's atoms. This carrier's kinetic energy is converted into primarily *vibrational* energy of the atoms, which corresponds to a temperature increase in the conductor. In short: all of the collisions in a conductor dissipate energy as heat, and this is a power loss.

Where does the energy come to cause this heat? From the power source driving the current, which could be the chemical energy stored in a battery. How much power is lost? Let us consider a resistor of value R in a circuit with a constant current through it, resulting in a potential difference $\Delta V = IR$. When an amount of charge ΔQ , corresponding to some number of electrons, passes through a resistor, it passes through a potential difference of ΔV , so *lowers* its potential energy by $\Delta PE = \Delta Q \Delta V$. This lost potential energy is what can be converted into heat inside the resistor.

If it takes an amount of time Δt for the charge packet ΔQ to go through the resistor, the *rate* of potential energy loss is:

$$\frac{\Delta PE}{\Delta t} = \frac{\Delta Q \Delta V}{\Delta t} = I \Delta V \quad (4.26)$$

Here we used Eq. 4.1. The rate at which the charges lose potential energy is equal to the rate at which the internal energy of the resistor rises. Energy change per unit time is nothing more than *power* (which we will denote by a fancy scripted \mathcal{P} to avoid confusion with pressure).

Power delivered in a circuit

$$\mathcal{P} = I \Delta V \quad (4.27)$$

In fact, *Equation 4.27 is valid for any type of device, Ohmic or not.* We didn't make any special assumptions, only that the packet of charge ΔQ passes through a net potential difference of ΔV , so this result works for any sort of electronic device, not just resistors. If we *do* have an Ohmic device, say just a plain resistor, we know the relationship between I and ΔV from Equation 4.23. Substituting that into the expression above:

Power delivered to a resistor

$$\mathcal{P} = I^2 R = \frac{\Delta V^2}{R} \quad (4.28)$$

¹³ We are ignoring the fact that the number of carriers also increases as temperature increases in a semiconductor, which is significant and also causes resistance to decrease as temperature increases.

4.6 Problems

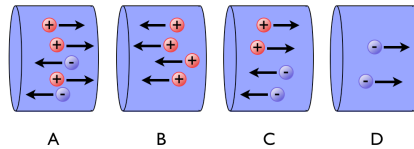
Solutions begin on page 267.

4.1. An electric current of $1\ \mu\text{A}$ flows through a conductor, which results in a 1.5 mV potential difference. What is the resistance of the conductor?

4.2. Which of the following do **not** obey Ohm's law?

- A resistor
- A slab of Copper
- A diode
- An insulator
- A capacitor

4.3. Consider the positive and negative charges moving horizontally through the four regions below. Which one has the highest current? The lowest? Consider the direction of positive current flow to be to the right.

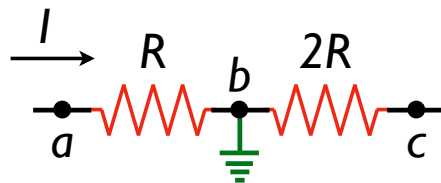


4.4. The drift velocity of charges in a typical copper wire is very small, $\sim 10^{-3}\text{ m/s}$. At this rate, it would take about 15 minutes after flipping the switch for your lights to come on. Why do your lights actually come on almost instantaneously?

4.5. Suppose a current-carrying wire has a cross-sectional area that gradually becomes smaller along the wire, so that the wire has the shape of a very long cone. How does the drift speed vary along the wire?

4.6. If the number of carriers in a conductor n decreases by 100 times, but the carriers' drift velocity v_d increases by 5 times, by how much does its **resistance** change?

4.7. A current I flows through two resistors in series of values R and $2R$. The wire connecting the two resistors is connected to ground at point b . Assume that these resistors are part of a larger complete circuit, such that the current I is constant in magnitude and direction. What is the electric potential relative to ground at points a and c , or V_a and V_c , respectively? *Hint: what is the potential of a ground point?*



4.8. Suppose a (cylindrical) electrical wire is replaced with one of the same material, but having every linear dimension doubled – the length and radius are twice their original values. What is the new value of resistance compared to the original?

- 4.9.** A potential difference of 11 V is found to produce a current of 0.45 A in a 3.8 m length of wire with a uniform radius of 3.8 mm. What is the resistivity of the wire?
- 4.10.** In a time interval of 1.37 sec, the amount of charge that passes through a light bulb is 1.73 C. How many electrons pass through the bulb in **5.00 sec**?
- 4.11.** A toaster is rated at 550 W when connected to a 130 V source. What current does the toaster carry?
- 4.12. (a)** A high-voltage transmission line with a diameter of 1.60 cm and a length 200 km carries a steady current of 1000 A. If the conductor is copper wire with a free charge density of $n = 8.20 \times 10^{28}$ electrons/m³, how long does it take one electron to travel the full length of the line?
- (b)** A high-voltage transmission line carries 1000 A starting at 600 kV for a distance of 150 mi. If the resistance in the wire is 0.5 Ω /mi, what is the power dissipated due to resistive losses?

Chapter 5

Direct-Current Circuits

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong. – Richard Feynman

Abstract What do we mean by *direct current*? “Direct current” is a constant flow of charges in the single direction, for example electrons from a low to high potential.¹ Charge flow whose magnitude varies in time will be considered at the end of this chapter (Section 5.6), and time variation of both *magnitude* and *direction* of charge flow will be considered in later chapters.

The circuit components we know about so far – batteries, resistors, capacitors, and diodes – can be used in various combinations to construct circuits. These circuits direct and control the flow of electrical energy. In this chapter, we will learn some rules that allow the analysis of a variety of simple direct current circuits to deduce their behavior. These principles follow from conservation of energy and conservation of charge. Most of the circuits we consider in this chapter are presumed to be “steady-state,” that is, the currents are constant in magnitude and direction. Toward the end of the chapter, we will consider some circuits where the *magnitude* of the current varies in time, but its direction is constant.

5.1 Sourcing Voltage

A current can only be maintained in a closed circuit by a source of electrical energy. The simplest way to generate a current in a circuit is to use a *voltage source*, such as a battery. As discussed in Sect. 3.5, a voltage source essentially raises or lowers the potential energy of charges that pass through it. The amount of energy gained *per charge* that passes through a device is the potential difference that the voltage supplies, ΔV , measured in Joules per Coulomb (J/C), *i.e.*, Volts (V). Though voltage is strictly an energy per unit charge, it is often useful to think of a voltage as a “pressure” of sorts, which tries to force charges through an electric circuit. Just like hydrostatic pressure, the presence of a voltage does not necessarily lead to a current, this only occurs when a completed circuit is present.

In this way of thinking, a voltage source is a sort of generalized power supply which can be thought of as a “charge pump” that tries to force charges to move within an electric field inside the source. Many batteries, for instance, are “electron pumps” in which negatively charged electrons move opposite to the direction of the electric field. In an idealized voltage source, the output terminals provide a constant potential difference ΔV , and can pump any amount of charge through any closed circuit connected to the output terminals.

Direct current: a constant flow of charges in the single direction, voltages and currents do not change in time. It is often abbreviated **dc** or **DC**. *dc* is preferred.

¹ Remember, as we mentioned last chapter, that long after this term was invented, scientists realized that electrons, negative charges, are what actually “flow” in a current, and the direction of current is *opposite* the direction of electron flow. It is confusing and annoying. So it goes.

In reality, pure voltage sources do not exist. Real voltage sources always have internal resistances, which “use up” some of voltage, and they have power limits which restrict the amount of current that can be sourced. A real voltage source is one which can supply, at best, a specified voltage, but the actual output may be less. Let us make this clearer by example.

Real batteries always have some internal resistance r , as illustrated in Figure 5.1. Real batteries therefore behave as a voltage source ΔV in *series* with an internal resistance r . This has the effect that the voltage at the battery terminals is always less than the rated value. Consider the simple circuit in Figure 5.2a, a battery specified to provide ΔV Volts connected to a resistor of value R .² If we neglect the internal resistance of the battery, the potential difference across the battery terminals is ΔV as rated. The rated voltage of a battery is the *idealized* terminal voltage of the battery in the limit that the battery itself has no internal resistance.

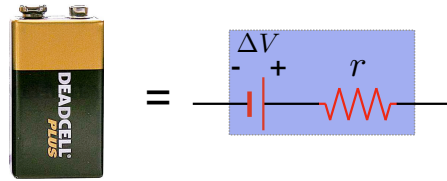


Fig. 5.1 A real battery provides a voltage ΔV , but has an internal resistance r . The actual output voltage developed at its terminals depends on r and the resistance of the circuit hooked up to the battery.

Now let us analyze the circuit of Figure 5.2b, the circuit diagram representation of the pictorial version in Figure 5.2a. The battery itself is everything inside the blue rectangle, and is modeled as a source of voltage ΔV in series with an internal resistance r .

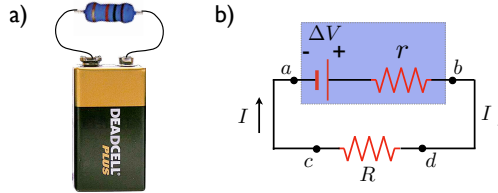


Fig. 5.2 (a) A circuit consisting of a resistor connected to the terminals of a battery. (b) A circuit diagram of a source of voltage ΔV having an internal resistance r , connected to an external resistor (load) R .

First, consider what happens to a positive charge moving *through the battery* from point a to b . As the charge goes from the negative to positive terminal of the battery, its potential increases by ΔV . Once it goes through the internal resistor r , however, its potential *decreases* by an amount Ir , where I is the current in the circuit. Thus, the voltage at the battery output terminals, points a and b , is the raw voltage ΔV minus the loss of due to the internal resistance:

$$\Delta V_{\text{terminals}} = V_b - V_a = \Delta V - Ir \quad (5.1)$$

This makes it clear that the voltage across the battery terminals $\Delta V_{\text{terminals}}$ is the same as the rated voltage ΔV when the current is zero. This is why another name for the rated voltage is the **open-circuit voltage** – rated and actual voltages are only the same for a real battery when nothing is connected and no current flows.

Now consider the effect of connecting an external resistor in Figure 5.2b. The external resistor (or resistive device, such as a light bulb) you are trying to power is often called the **load resistance**. Since it is directly connected to the battery terminals, and the wires are assumed to be perfect, it must have a potential difference across it of $\Delta V_{\text{terminals}}$. The potential difference across the load resistor and the current through it must also follow Ohm’s law, hence $\Delta V_{\text{terminals}} = IR$ (using Eq. 4.23).

² We are assuming, as we almost always will, that wires connecting to the battery have no resistance.

Combining this with Equation 5.1, we can relate the rated battery voltage to the internal resistance and the load resistance:

$$\Delta V_{\text{terminals}} = IR = \Delta V - Ir \quad (5.2)$$

$$\implies \Delta V = IR + Ir \quad (5.3)$$

The total rated voltage of the battery is partly spent on the load resistor and partly spent on the internal resistance. This is just conservation of energy – charges must go all the way around a closed loop and come back with the same energy. If not, we would have a perpetual motion device, gaining energy out of thin air! Every bit of potential gained by charges from a voltage source must be lost somewhere else in the circuit loop – in a resistor for example. We will see later that this is part of a more general rule – the sum of voltage sources and voltage losses in a closed loop must be zero.

Now that we have related the battery's rated voltage ΔV to the internal and load resistances, we can solve Eq. 5.3 for the current I through the battery and resistor:

Current supplied by a voltage source:

$$I = \frac{\Delta V}{R + r} \quad (5.4)$$

where R is the resistance of the load connected to the battery, r is the internal resistance of the battery, and ΔV is the rated open-circuit battery voltage.

Now it is clear that the current delivered by the battery through the resistor actually depends on both the resistor's value *and* the internal resistance of the battery. If $R \gg r$, of course we do not need to worry about the internal resistance of the battery. When the load resistance is high enough that we can neglect the internal resistance, nearly all of the rated voltage is developed across the load resistor. We can explicitly write down the actual voltage developed across the load resistor R to make this more clear:

Voltage delivered to a load by a voltage source:

$$\Delta V_{\text{load}} = IR = \Delta V \frac{R}{r + R} \quad (5.5)$$

where the quantities are the same as in Eq. 5.4. *The voltage delivered to the load depends on the value of the load and internal resistances.*

Now it is even easier to see that when r is small enough to be neglected, the battery operates as nearly an ideal voltage source, supplying almost the whole ΔV to the load itself. This is usually how things work out.³ In a nutshell, to operate properly voltages sources like high load resistances compared to their internal resistance.

Batteries: neither constant I nor V

- Equations 5.1 and 5.5 indicate that the terminal voltage of a battery depends on its own internal resistance *and* the load resistance, so a battery is not a constant voltage source.
- Equation 5.4 indicates that the current supplied depends on the load resistance, so a battery is not a constant current source.

³ If not ... you probably have a badly designed circuit, and very quickly, a dead battery!

We can also find the power output of our battery by multiplying Equation 5.3 by I . Keep in mind that the *total* power output is the total voltage ΔV times the total current I .

$$\mathcal{P} = I\Delta V = I \cdot I(r + R) = I^2(R + r) = I^2R + I^2r \quad (5.6)$$

The total power output $I\Delta V$ of the battery is delivered both to the resistor and the battery's internal resistance, at the rate I^2R to the resistor *and* I^2r within the battery itself. Again, if $R \gg r$ we do not need to worry about the power lost in the battery itself, and this is usually the case. One thing to keep in mind: should you connect too small a load to the battery (for example by short-circuiting it), such that $r \sim R$, you will immediately notice the I^2r power dissipated within the battery itself – in the form of heat.

Just to be complete, we can also write the power output in another way, using Eq. 5.4:

$$\mathcal{P} = \Delta V \frac{R}{r + R} \cdot \Delta V = \Delta V^2 \frac{R}{r + R} \quad (5.7)$$


The expressions above tell us what power is delivered by the battery or voltage source. Batteries and other voltage sources typically have power ratings (in Watts) which tell you the maximum \mathcal{P} that can be delivered. From the equations above, it is straightforward to calculate the proper resistances, voltages, and currents within a given power rating.

Everything above applies not just to batteries, but to any sort of voltage source. All real voltage sources have an internal resistance, and are subject to the same considerations above. Batteries, however, have an additional constraint that they have a limited capacity. The available capacity of a battery depends upon the rate at which it is discharged – if a battery is discharged at a high rate, the available capacity will be lower. Conversely, discharging a battery at a low rate prolongs its life. Batteries are usually given a capacity rating of A·h or mA·h along with their rated voltage.⁴ From the rated voltage and capacity, we can calculate a product of power and hours which tells us how long a battery can deliver a certain power:

$$\mathcal{P} \cdot \text{hours} = \text{capacity} \cdot \Delta V \quad \text{or} \quad \text{hours} = \frac{\text{capacity} \cdot \Delta V}{\mathcal{P}} \quad (5.8)$$

5.2 Sourcing Current

A current source is nothing more than a device that delivers and absorbs a constant current, sourcing and sinking a constant number of charges per unit time. An ideal current source (which exists only on paper) delivers a constant current to any closed circuit connected to its output terminals, no matter what the voltage or load resistance.⁵ Though a battery provides a simple example of a voltage source, there is no correspondingly simple realization of a current source.

Circuit diagram symbol for a current source: 

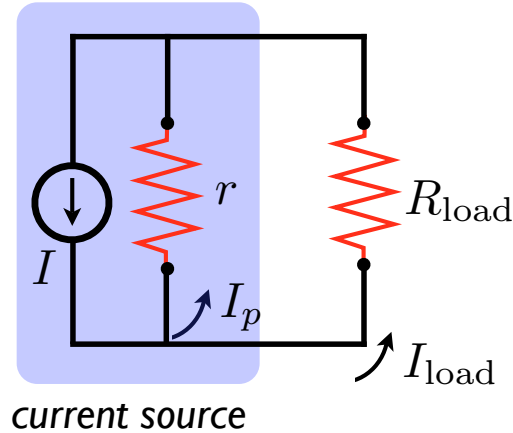
We can *approximate* a current source, however, with a single battery and resistor. In the circuit of Fig. 5.2, a battery with internal resistance connected to a load resistor, the current through the load is given by Eq. 5.4. If we make the load resistor very small (or equivalently, make the internal resistance of the battery very large), $r \gg R_{\text{load}}$, then the current through the load resistor is $I \approx \Delta V / r$. This does provide a roughly constant current, but the power loss in the internal resistor will be severe, and it is generally impractical to construct a current source in this way.

⁴ For example, a typical alkaline AA battery has a capacity of ~ 2.85 A·h at its rated voltage of 1.5 V.

⁵ You may have already guessed that this is a serious oversimplification. You would be right.

How more realistic constant current sources work internally is a bit beyond the scope of our discussion. However, that does not prevent us from seeing how they behave when connected to a circuit. In the same way that a real voltage source can be considered an ideal voltage source in series with a resistor, a real current source can be considered an ideal current source in parallel with a resistor, as shown in Fig. 5.3.

Fig. 5.3 A real current source can be considered as an ideal current source in parallel with an internal resistance r . The internal resistance “steals” some of the current, depending on the value of the load resistor.



The current source attempts to supply a current I to its own internal resistance and the load resistor. When a stream of charges tries to leave the current source, it quickly encounters a junction between the internal resistor and the load resistor. As we will find out in more detail in Sect. 5.3.2, when a current encounters a junction it splits up and takes *both* paths, inversely proportional to their resistance. That is, more current goes through the smaller of the two resistors, but some current goes through the larger resistor too. Since charge must be conserved, the sum of the currents in the two resistors must equal the total current before the junction.

If the internal resistance is very large, almost all of the current goes through the load, and the current source is nearly ideal. If the load resistance becomes comparable to the internal resistance, however, a significant portion of the current takes the “parasitic” path (I_p in the figure) through the internal resistance, and the source is no longer close to ideal. You will figure out how to calculate the actual current through the load in Sect. 5.3.2, but for now we will quote you the result:

$$I_{\text{load}} = I \frac{r}{r + R} \quad (5.9)$$

As you can see, the current through the load is independent of the load resistance R and nearly equal to the source current I when $r \gg R_{\text{load}}$. In other words, current sources want low load resistances, in contrast to voltage sources. This brings up one answer to a common question: is it better to source current or voltage? If the load you are trying to source has a large resistance, sourcing voltage is generally better. If the load is small, sourcing current is generally better.⁶

Source voltage or current?

- Current sources have high internal resistances and like low load resistances.
- Voltage sources have low internal resistances and like high load resistances.
- If the load has a large resistance, sourcing voltage is generally better.

⁶ For sources, internal resistance is often called “output resistance.” Good laboratory current sources can have internal resistances above $10^{14} \Omega$, while good laboratory voltage sources can have internal resistances below 1Ω , so with good equipment either I or ΔV can usually be sourced without issues. Noise is what usually determines which is actually used, but even so, the rule of thumb stated is still valid.

- If the load has a small resistance, sourcing current is generally better.

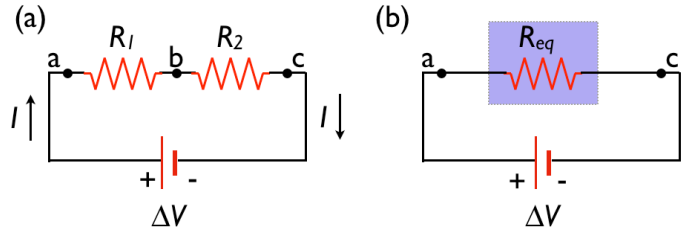
5.3 Combinations of Resistors

Based on section 5.1, we can now understand a bit more clearly what happens when we connect a battery to a single resistor. What about combinations of resistors? In section 3.6.4, we learned that complicated combinations of capacitors could usually be reduced to a single *effective capacitance*, based on the rules for two distinct pairings, series and parallel. The same is true for resistors. Moreover, these two combinations of resistors end up being useful circuits in their own right.

5.3.1 Resistors in Series

Figure 5.4 shows two resistors connected in series with a battery. The resistors could be, *e.g.*, light bulbs or heaters, or just plain resistors. When the resistors R_1 and R_2 are connected to the battery, **the current through each resistor is the same**. This makes sense – there is only one single path in the circuit, so there can only be one current. This is because every charge that flows through R_1 must also flow through R_2 and back to the battery. This is just **conservation of charge**, in the same way we say that any water flowing into a pipe has to come out again.

Fig. 5.4 (a) Two resistors R_1 and R_2 connected in series with a battery. (b) The currents in the resistors are the same, and the equivalent resistance of the combination is $R_{\text{eq}} = R_1 + R_2$.



We know the current is the same through both resistors, and conservation of energy tells us that the potential difference between points a and c must equal the battery voltage ΔV , Equation 5.1. The potential difference between a and c we can break up into the sum of the potential difference between a and b and the potential difference between b and c : $\Delta V_{ac} = \Delta V_{ab} + \Delta V_{bc}$. What is the potential difference between points a and b ? This is just the potential drop across the resistor R_1 , IR_1 . Similarly, the potential difference between points b and c is IR_2 . Conservation of energy tells us that the potential drop across both resistors together must equal the battery voltage:

$$\Delta V = IR_1 + IR_2 = I(R_1 + R_2) = IR_{\text{eq}} \quad \text{with} \quad R_{\text{eq}} = R_1 + R_2 \quad (5.10)$$

The right hand side of this equation shows us that the potential drop across both resistors is the same as it would be for a single resistor of $R_{\text{eq}} = R_1 + R_2$. In other words, in series, resistors just add together. No matter how many we have, the equivalent resistor of a series combination is just the sum of the individual resistances. Notice, however, that the current through resistors in series is the same. Further, since series resistors must have the same current, they all have the same current as their equivalent resistance as well.

Two Resistors in Series:

$$R_{eq} = R_1 + R_2 \quad (5.11)$$

Three or More Resistors in Series:

$$R_{eq} = R_1 + R_2 + R_3 + \dots \quad (5.12)$$

The current through resistors in series is the same.

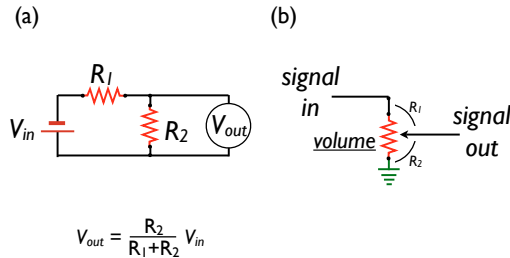


Fig. 5.5 (a) Two resistors in series can act as a ‘voltage divider,’ providing two different voltages from a single supply. (b) Using a variable resistor, this is one way to control audio volume. A sliding bar on the variable resistor changes the values of R_1 and R_2 while keeping their sum constant. When R_1 is small and R_2 large, most of the signal reaches the output (sliding bar at the top). When R_1 is large and R_2 small, most of the signal is sent to the ground, and the output is small.

The potential difference (voltage) across each resistor is unless the resistors are identical, hence **series resistors are often called “voltage dividers.”** Figure 5.5 shows how a voltage divider can be used as an audio volume control, using a single variable resistor. This can be seen in Figure 5.4 – from the battery voltage ΔV we have generated two different (lower) voltages, $\Delta V_1 = IR_1$ between points a and b , and $\Delta V_2 = IR_2$ between points b and c .

Perhaps this will help?

Resistors in series add like capacitors in parallel, and *vice versa*.

5.3.2 Resistors in Parallel

Resistors in series, we found, added like capacitors in parallel. As you might expect, *resistors in parallel add like capacitors in series*. Consider the parallel combination of resistors in Figure 5.6. Both resistors are connected directly to the battery terminals, so **the potential difference across both resistors is the same**. The currents are *not* the same, unless the resistors are identical.

Current in parallel circuits behaves just like a “T” in a system of water pipes. Current coming out of the positive pole of the battery flows to point a , and splits into two parts, I_1 flowing through R_1 and I_2 flowing through R_2 . The larger portion of the current goes through the the smaller resistor.

Junctions in circuits: current splits up to take all possible paths at once, divided up inversely proportional to the resistance of the path.

Because charge has to be conserved, just like water flowing into a network of pipes eventually has to come out, **the current I that enters point a must equal the total current leaving that point:** $I = I_1 + I_2$. Since the potential drop must be the same across both resistors, we can easily find I_1 and I_2 :

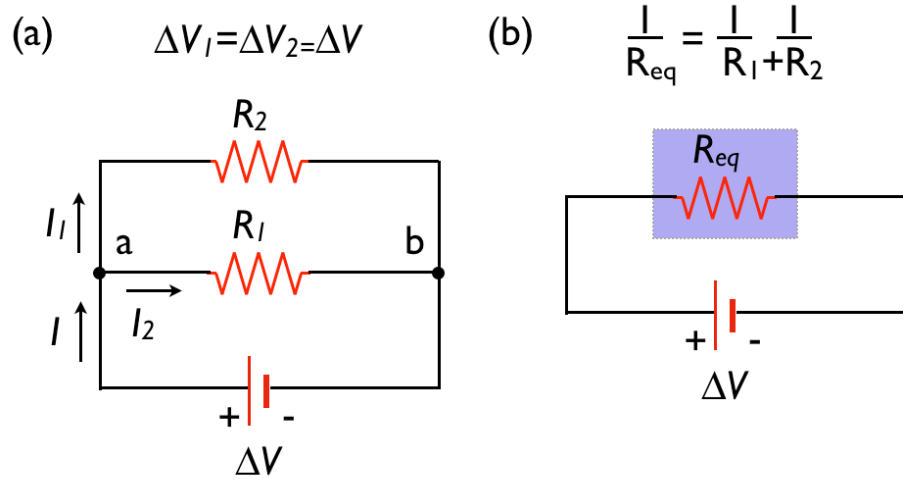


Fig. 5.6 (a) Two resistors R_1 and R_2 connected in parallel with a battery. The potential differences across R_1 and R_2 are the same. (b) The equivalent resistance of the combination is given by $1/R_{eq} = 1/R_1 + 1/R_2$.

$$I_1 = \frac{\Delta V}{R_1} \quad \text{and} \quad I_2 = \frac{\Delta V}{R_2} \quad (5.13)$$

We want to find a single equivalent resistor R_{eq} , such that $I = \Delta V/R_{eq}$. First, we just write down the expression for the total current, and rearrange it a bit:

$$I = I_1 + I_2 \quad (5.14)$$

$$= \frac{\Delta V}{R_1} + \frac{\Delta V}{R_2} = \Delta V \left[\frac{1}{R_1} + \frac{1}{R_2} \right] \quad (5.15)$$

$$= \Delta V \left[\frac{R_2}{R_1 R_2} + \frac{R_1}{R_1 R_2} \right] = \Delta V \left[\frac{R_2 + R_1}{R_1 R_2} \right] \quad (5.16)$$

$$= \frac{\Delta V}{R_{eq}} \quad (5.17)$$

Now we can equate the right-hand sides of Equations 5.16 and 5.17 to find out what R_{eq} is:

$$\frac{\Delta V}{R_{eq}} = \Delta V \left[\frac{R_2 + R_1}{R_1 R_2} \right] \quad (5.18)$$

$$\frac{1}{R_{eq}} = \left[\frac{R_2 + R_1}{R_1 R_2} \right] \quad (5.19)$$

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} \quad \text{or} \quad R_{eq} = \frac{R_1 R_2}{R_1 + R_2} \quad (5.20)$$

So now we have derived that resistors in parallel add inversely, just like capacitors in series. *The potential difference (voltage) across resistors in parallel is the same, and the equivalent resistance is always less than the smallest resistance in the group.*

Two Resistors in Parallel:

$$\frac{1}{R_{eff}} = \frac{1}{R_1} + \frac{1}{R_2} \quad \text{or} \quad R_{eq} = \frac{R_1 R_2}{R_1 + R_2} \quad (5.21)$$

Three or More Resistors in Parallel:

$$\frac{1}{R_{\text{eq}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \dots \quad (5.22)$$

Resistors in parallel add like capacitors in series and *vice versa*.

The current through each is different, hence **parallel resistors are often called “current dividers.”** This can be seen from Figure 5.6 – from a single current I , we have generated two different and smaller currents I_1 and I_2 .

What happens this time if one of the resistors fails? The other continues to be powered this time. Household circuits are wired in parallel, so that each device operates independently of the others. Further, all devices operate at the same voltage when wired in parallel. If they were connected in series, the voltage seen by each device would depend on how many devices were connected and their individual resistances. Parallel wiring is why the lights do not dim when you turn on the TV!

The disadvantage to parallel wiring is that when one device fails, the others would suddenly see a larger current – if R_1 failed in Figure 5.6, R_2 would suddenly see the full current I , not just I_2 , which could cause serious problems. In reality, circuit breakers are inserted in series with each device, which limit the current to some maximum value (typically 15 or 20 A).

What to do for more complex combinations of resistors?

1. **Combine** resistors that are in parallel or series in to single equivalent resistors, using (5.21) and (5.11).
2. **Series** resistors all have the same current, and $R_{\text{eq}} = R_1 + R_2 + R_3 + \dots$
3. **Parallel** resistors all have the same voltage, and $\frac{1}{R_{\text{eq}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \dots$
4. **Redraw** the circuit after every combination.
5. **Repeat** the first two steps until there is only equivalent one resistors R_{eq} left.
6. **Use Ohm’s law**, $\Delta V = IR_{\text{eq}}$ (Equation 4.23), to find the current in R_{eq} .
7. **Reverse** your steps one by one to find the current and voltage for each equivalent resistors along the way, until you recreate the original diagram.

5.3.3 Example: a Complex Resistor Combination

Just as we saw with complex capacitor combinations (Sec. 3.6.4.3), once you know the rules for series and parallel resistors, you can simplify most complicated resistor combinations.

Consider the example in Figure 5.7, where we have resistors R_1 through R_4 connected to a battery⁷ supplying a voltage ΔV . Now trace the wires from the negative pole of the battery. A current I will be present in the single wire leaving the battery, and it will split up into I_1 and I_2 when it encounters the first junction. The currents I_1 and I_2 will recombine at the junction just before R_4 , and the current I goes back to the battery. Conservation of charge tells us $I = I_1 + I_2$.

We start to simplify by combining the simple series pair R_2 and R_3 into R_{2-3} , as shown in Figure 5.7a-b. Equation 5.11 tells us:

$$R_{2-3} = R_2 + R_3 \quad (5.23)$$

Once we have done that, Figure 5.7b, we have a simple parallel combination of R_{2-3} and R_1 . Equation 5.21 tells us that the equivalent resistance of these two, R_{1-2-3} (Figure 5.7c) is:

⁷ We will assume that the battery’s internal resistance is negligible compared to any of the resistors R_1 - R_4 so we may neglect it. If we want to include it, we can always do that by including a resistor r in series with the voltage source, like in Figure 5.1.

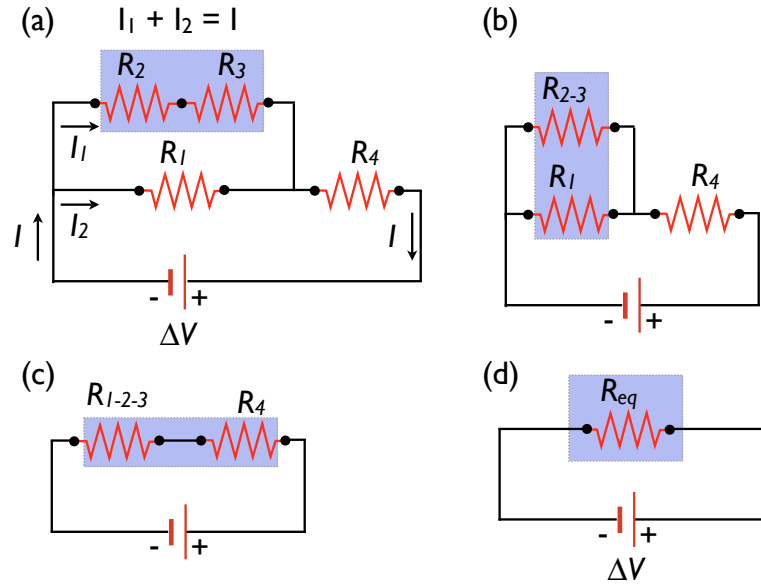


Fig. 5.7 (a) The current I leaving the battery splits up into I_1 and I_2 when it encounters the junction, and recombines at R_4 . Conservation of charge tells us $I = I_1 + I_2$. We start to simplify by combining the simple series pair R_2 and R_3 into R_{2-3} . (b) Now we can combine the simple parallel pair R_{2-3} and R_1 into R_{1-2-3} . The current through R_{2-3} is still I_1 , and the current through R_1 is still I_2 . (c) Now we are left with only a simple series pair, R_{1-2-3} and R_4 , which we combine into R_{eq} . The current through R_{eq} is just I . Now we can work backwards and find I_1 , I_2 , and the voltage drop across each resistor.

$$\frac{1}{R_{1-2-3}} = \frac{1}{R_1} + \frac{1}{R_{2-3}} \quad \text{or} \quad R_{1-2-3} = \frac{R_1 R_{2-3}}{R_1 + R_{2-3}} \quad (5.24)$$

Now we are left with only a simple series pair (Figure 5.7c), R_{1-2-3} and R_4 , which we combine into R_{eq} (Figure 5.7d). Note that the current through R_{eq} is just I .

$$R_{eq} = R_{1-2-3} + R_4 \quad (5.25)$$

Once we have a single resistor, Figure 5.7d, we know that $\Delta V = IR_{eq} = I(R_{1-2-3} + R_4)$. If we are given the values of the resistors and ΔV , we can calculate the current I , and the voltage drop across R_4 :

$$I = \frac{\Delta V}{R_4} \quad \text{and} \quad \Delta V_4 = IR_4 \quad (5.26)$$

Working backwards to Figure 5.7c, we know that the total voltage drop across R_{1-2-3} and R_4 together is ΔV . Since the voltage drop across R_4 alone is $\Delta V_4 = IR_4$, and the total voltage in the whole circuit has to be ΔV , the voltage across R_{1-2-3} has to be $\Delta V - \Delta V_4$. This is just conservation of energy again.

Now going back to Figure 5.7b, we know that since R_1 and R_{2-3} are in parallel, they have the same voltage drop, which has to be $\Delta V - \Delta V_4$. This gives us immediately the current I_2 in R_1 :

$$I_2 = \frac{\Delta V - \Delta V_4}{R_1} = \frac{\Delta V - IR_4}{R_1} \quad (5.27)$$

This then gives us I_1 by conservation of charge:

$$I_1 = I - I_2 = \left[\frac{\Delta V}{R_4} \right] - \left[\frac{\Delta V - \Delta V_4}{R_1} \right] = \left[\frac{\Delta V}{R_4} \right] - \left[\frac{\Delta V - IR_4}{R_1} \right] \quad (5.28)$$


Finally, back to Figure 5.7a, since R_2 and R_3 are in series, they have the *same current* I_1 given above. That gives us the voltage drops across R_2 and R_3 as $I_1 R_2$ and $I_1 R_3$, respectively. And now we know everything about this circuit! Well, except what possible use it might have ... but that is another topic entirely.

5.4 Current and Voltage Measurements in Circuits

At this point, we know how to make some simple (but useful) circuits, and properly source either current or voltage. What we have not really touched on so far is how to *measure* currents and voltages properly in circuits. As with sourcing, each type of measurement has its own non-idealities.

5.4.1 Measuring Voltage

A voltmeter is just what it sounds like – a device that measures voltage, or potential difference, between two points. A typical voltmeter has two input terminals, and one simply connects wires from these input terminals to the points within a circuit between which one wants to know the potential difference. If we wish to measure the potential difference across a particular component in a circuit, we connect the voltmeter in *parallel* with that component.

Circuit diagram symbol for a voltmeter: 

Of course, the idea is to measure the potential difference while disturbing the circuit as little as possible. For this reason, voltmeters have very high internal resistances (see Fig. 5.9a),⁸ and no current flows through an ideal voltmeter. As an example, Fig 5.8a shows an incorrect use of a voltmeter – connecting the voltmeter in *series* with the resistor and battery. No current flows through an ideal voltmeter, so connecting the voltmeter in this way essentially opens the circuit and nothing is measured. Figure 5.8b shows the proper use of a voltmeter – in *parallel* with the component to be measured, a resistor in this case. The voltmeter probes the potential on both sides of the resistor, but since no current flows through it, it does not affect the circuit.

Real voltmeters are not ideal, you might have guessed. A real voltmeter has a finite input resistance, and, even when connected properly, draw a small amount of current. As shown in Fig. 5.9, a voltmeter connected properly forms a parallel resistor network with the load resistor R_{load} . What the voltmeter really measures then is not just the load, but the equivalent resistance of the load in parallel with its own internal resistance r .

Put another way, the voltmeter forms a *current divider* with the load, and “steals” part of the current through the load. The voltmeter “stealing” part of the current obviously leads to inaccurate results, and the measured voltage drop across the resistor is no longer IR_{load} like we expect. We should try to figure out how bad this problem is! If we assume there is a current I in the wire leading to the resistor, we can readily calculate the voltage measured by the voltmeter:

$$\Delta V_{\text{measured}} = IR_{\text{eq}} = \frac{rR_{\text{load}}}{r + R_{\text{load}}} I = \frac{IR_{\text{load}}}{1 + \frac{R_{\text{load}}}{r}} = \frac{\Delta V_{\text{expected}}}{1 + \frac{R_{\text{load}}}{r}} \quad (5.29)$$

The ratio between the measured voltage and the expected value is $1/(1 + \frac{R_{\text{load}}}{r})$, which tells us two things. First, the measured value is always *smaller* than the true value, since $1/(1 + \frac{R_{\text{load}}}{r}) \leq 1$.

⁸ Good laboratory voltmeters can have internal resistances on the order $10^{10} \Omega$ or more. For voltmeters, internal resistance is often called “input resistance.”

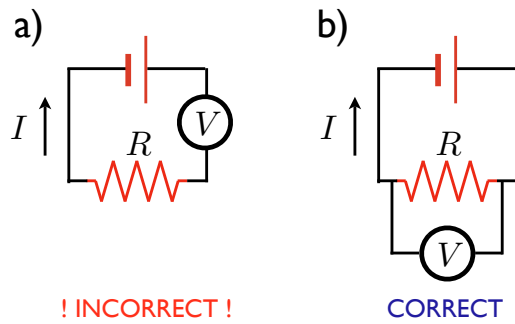


Fig. 5.8 (a) Incorrect connection of a voltmeter. Voltmeters have enormous internal resistances, current will not flow through them. (b) Correct connection of a voltmeter. The two input terminals of the voltmeter connect across the resistor, and draw no current.

Second, so long as the load resistor is small compared to the internal resistance of the meter, $R_{\text{load}} \ll r$, the measured and expected values will be very close. Given the enormous internal resistance of most modern voltmeters, this is usually the case, but one must still exercise caution. Using a meter with insufficient internal resistance is known as “measuring the meter,” and is something you will encounter in your laboratory experiments.

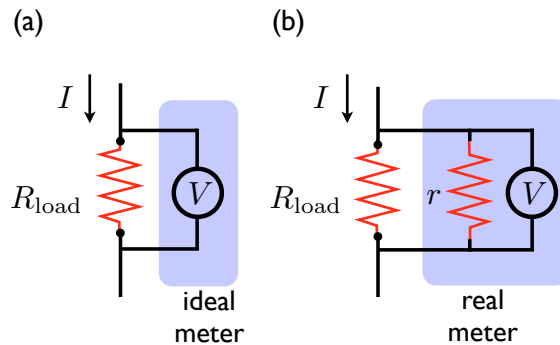



Fig. 5.9 (a) An ideal voltmeter has an infinite internal resistance, and no current flows through it. Hence, it measures the true voltage drop across the resistor, $\Delta V = IR$. (b) A real voltmeter has a finite internal resistance r , and forms a voltage divider with the load resistor. Some current flows through the voltmeter itself if R_{load} is comparable to r , and the measured voltage is *less* than the true voltage on the resistor.

5.4.2 Measuring Current

An ammeter is the device that measures current, and it behaves rather differently than a voltmeter. Measuring the flow of charge has similarities with measuring the flow of fluids. A flow meter measures fluid flow by allowing the fluid of interest to pass through it. Similarly, an ammeter measures charge flow by allowing current to pass through it. Ammeters therefore connect in *series* with the device of interest.

Circuit diagram symbol for an ammeter: —  —

Ammeters have tiny internal resistances, and current flows readily through them. If an ammeter is connected incorrectly in *parallel* with the load, it will create current divider (parallel resistor network) with the load resistor. The small internal resistance of the ammeter will “steal” all of the current from the load resistor, and an improper measurement results. Since the ammeter resistance is

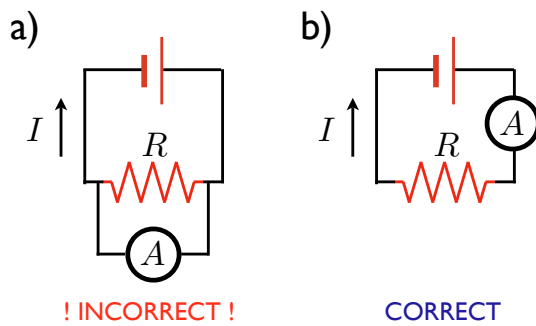


Fig. 5.10 (a) Incorrect connection of an ammeter. Ammeters have tiny internal resistances, current flows readily through them. In this case, the ammeter would “steal” all of the current from the resistor. (b) Correct connection of an ammeter. The ammeter connects in series with the resistor to measure the current, and creates no additional voltage drop.

so small, it can be connected in *series* with the load, and the voltage drop across the ammeter itself is negligible – it measures the current without disturbing the circuit as shown in Fig. 5.10

A simple ammeter can be constructed using a precise resistor and a good voltmeter, as shown in Fig. 5.11.⁹ A precise resistor placed in series with the device to be measured (in place of the ammeter in Fig. 5.10, for instance), and a voltmeter measures the voltage drop across this precise resistor.

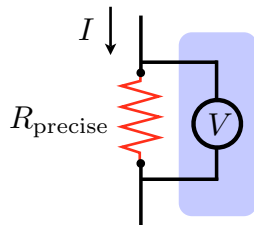


Fig. 5.11 A simple ammeter can be constructed from a precise resistor and a good voltmeter. Since the value of the resistance is known, the measured voltage drop across it yields the current.

Since the value of the resistor is known precisely, the measured voltage drop across it yields the current *via* Ohm’s law:

$$I = \frac{\Delta V_{\text{measured}}}{R_{\text{precise}}} \quad (5.30)$$

In this way currents can be measured reasonably accurately, but this is far from an ideal ammeter. First, this technique of current measurement brings in all the non-idealities associated with real voltmeters as discussed above. Second, placing a resistor within the circuit of interest introduces an additional voltage drop, which can affect other components. Care must be exercised when using this technique. The precise resistor can be chosen carefully as not to introduce a sufficiently large voltage drop to alter the circuit too much, the voltages on other components in the circuit must be independently measured to take this effect into account, or the circuit must be designed from scratch to account for this additional voltage drop.

Voltmeters and Ammeters

- Voltmeters have a high internal resistance and draw little current.
- Voltmeters connect in *parallel* with the device to be measured.
- Ammeters have a low internal resistance, and current passes through readily.
- Ammeters connect in *series* with the device to be measured.

⁹ This is how we will measure currents in our laboratory sessions, see Appendix ??

5.5 Kirchhoff's Rules and Complex dc Circuits

We have just seen that simple circuits can be analyzed using Ohm's law and the rules for series and parallel resistors. However, there are ways in which resistors can be connected such that there is not a single equivalent resistance. For these more complicated cases, we use two simple rules, known as **Kirchhoff's rules**:

Kirchhoff's Rules:

1. The sum of currents entering any junction must equal the sum of the currents leaving that junction. *a.k.a.* the “junction rule.”
2. The sum of the potential differences across all the elements around any *closed* circuit loop must be zero. *a.k.a.* the “loop rule.”

The junction rule is nothing more than *conservation of charge*. Whatever charge flows into a given point in a circuit has to flow out again, charges are neither created nor destroyed. Figure 5.12 illustrates this rule, that the current entering the junction (I_1) has to be the same as the sum of the currents leaving the junction ($I_2 + I_3$), that is, $I_1 = I_2 + I_3$. Again using our fluid analogy, this would be the same as a “tee” in a water pipe. The flow rate into the pipe equals the total flow rate out of the two branches.

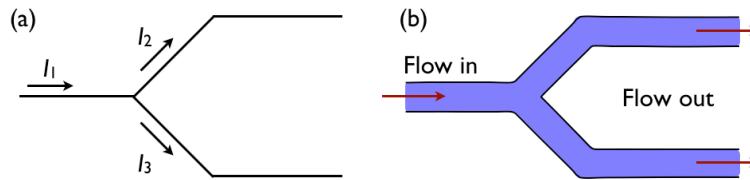


Fig. 5.12 A schematic illustrating Kirchhoff's “junction rule.” (a) Conservation of charge requires that the current entering the junction equal the sum of the currents leaving the junction, $I_1 = I_2 + I_3$. (b) An everyday analogy is a “tee” in a water pipe – the net flow in must equal the net flow out.

The loop rule is nothing more than *conservation of energy*. Like we saw in Section 5.3.3, any charge that moves around a closed loop in a circuit must gain as much energy as it loses. Charges gain energy when going through a source of voltage, and lose energy by way of a potential drop across a resistor. Charges also lose energy by going *backwards* into a source of voltage. As one example, potential energy of the charge is converted into chemical energy when charging a battery.

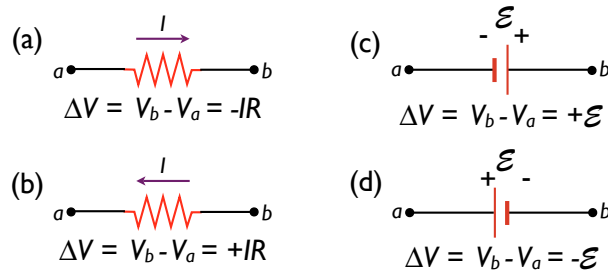


Fig. 5.13 Rules for determining potential differences across resistors and batteries. Here we assume the battery has no internal resistance.

In general you can use the junction rule one time fewer than the total number of junction points in the circuit. All this means is that one point has to be left such that your resulting circuit is still

a closed loop. The loop rule can be used as often as needed, until you have only one loop left. **Solving any particular problem requires as many unique independent equations as you have unknowns.** Figure 5.13 illustrates the rules for determining whether the voltage difference across a resistor or battery is positive or negative when applying Kirchhoff's rules.

5.5.1 Example: analyzing a simple parallel or series circuit

In order to understand the power of Kirchhoff's laws in analyzing complicated circuits (ones for which the simple series and parallel rules will not work, for example), we should first analyze a circuit we already understand. In that light, we will re-examine our parallel resistor circuit in Fig. 5.6 according to our 'Using Kirchhoff's rules' box below.

Using Kirchhoff's Rules:

1. Assign symbols and directions to the currents in all branches of the circuit.
2. If you guess the direction of the current wrong at first, don't worry. The magnitude will be correct, but the *sign* will come out negative, indicating that the direction is opposite what you expected.
3. Use the loop rule on as many loops as you can to come up with more equations to solve the problem.
4. When you use the loop rule, chose a direction for going around the loop and stick with it. Pick, *e.g.*, clockwise or counterclockwise, and always do that.
5. As you go around a loop, add up voltage drops and rises according to these rules (see Figure 5.13)
 - a. Going across a resistor with the current *lowers* the potential by $-IR$
 - b. Going across a resistor *against* the current *raises* the potential by $+IR$
 - c. If you cross a voltage source from $-$ to $+$, the change in potential is $+\Delta V$
 - d. If you cross a voltage source from $+$ to $-$, the change in potential is $-\Delta V$
6. Use the junction rule as often as you can to generate new equations to solve the problem
7. Solve the equations you generate for the unknown quantities
8. Plug your answers back into the original equations to check for consistency.

In this case, we have already assigned the proper symbols and labeled the currents in each branch. The next step is to use the loop rule as many times as possible. For this circuit, there are only two loops present - an upper one, containing R_2 and I_1 , and a lower one containing R_1 and I_2 . For the upper loop, the loop rule says that the sum of all voltage sources and drops around the loop must be zero. Traversing the loop clockwise from point **a** (an arbitrary choice), we first go through R_2 in the direction of the current, giving a voltage *drop*, and then through R_1 against the current, giving a voltage *increase*. The wires themselves are still assumed to be perfect, and give no voltage changes. Accordingly, using the labeled current in each resistor:

$$0 = -I_1 R_2 + I_2 R_1 \quad (5.31)$$

$$\implies I_1 = \left(\frac{R_1}{R_2} \right) I_2 \quad (5.32)$$

Now consider the bottom loop, again traversing clockwise from point **a**. We first go through R_1 in the direction of the current, giving a voltage drop, and then go through the battery from $-$ to $+$, giving a voltage *increase*. Thus:

$$0 = -I_2 R_1 + \Delta V \quad (5.33)$$

$$\implies \Delta V = I_2 R_1 \quad (5.34)$$

$$\text{and } I_2 = \frac{\Delta V}{R_1} \quad (5.35)$$

Combining what we know so far:

$$I_1 = \left(\frac{R_1}{R_2}\right) I_2 = \left(\frac{R_1}{R_2}\right) \frac{\Delta V}{R_1} = \frac{\Delta V}{R_2} \quad (5.36)$$

That does it for the loop rule. Next, we need to apply the junction rule. Our only junctions are at points **a** and **b**, and both give the same result:

$$I = I_1 + I_2 \quad (5.37)$$

Now we can combine this result with the equations we got from the loop rule for I_1 and I_2 :

$$I = I_1 + I_2 = \frac{\Delta V}{R_2} + \frac{\Delta V}{R_1} \quad (5.38)$$

$$I = \Delta V \left(\frac{1}{R_2} + \frac{1}{R_1} \right) \quad (5.39)$$

$$\implies R_{\text{eff}} = \frac{\Delta V}{I} = \left(\frac{1}{R_2} + \frac{1}{R_1} \right)^{-1} \quad (5.40)$$

On inspection, this is exactly our formula for adding resistors in parallel. How about the case of series resistors, Fig. 5.4? Much easier in fact. Again, we already have everything labeled, so we just start with the loop rule – made simpler by the fact that we have only one loop now. We will traverse it clockwise from point **a** once again, which means we first pass through resistors R_1 and R_2 in the direction of the current, and then through the battery in the positive direction:

$$0 = -IR_1 - IR_2 + \Delta V \quad (5.41)$$

$$\implies \Delta V = I(R_1 + R_2) \quad (5.42)$$

There is no junction rule in this case, since we have no junctions! So this is it, which you should recognize as our formula for adding resistors in series.

5.5.2 Example: analyzing a complex circuit

That was easy enough, right? Of course, that means it is time to consider a more *pathological* example, such as the circuit in Fig. 5.14a. Don't panic! If we systematically follow the rules, and our handy-dandy guide for using them (Page 133), we can make short work of this circuit too. The first step is to label all of the components and assign directions to the separate currents in each unique branch, Fig. 5.14b. Again, if we guess the direction incorrectly, it isn't a big deal – the sign of the current will just come out negative, letting us know that the direction is opposite what we expected. What is important is just to put down *something* so that the rules can be properly applied.

Just to make things more concrete, let's give all of our components some real values too:

$$\begin{array}{lll} \Delta V_1 = 19\text{ V} & \Delta V_2 = 6\text{ V} & \Delta V_3 = 2\text{ V} \\ R_1 = 6\,\Omega & R_2 = 4\,\Omega & \\ R_3 = 4\,\Omega & R_4 = 1\,\Omega & \end{array} \quad (5.43)$$

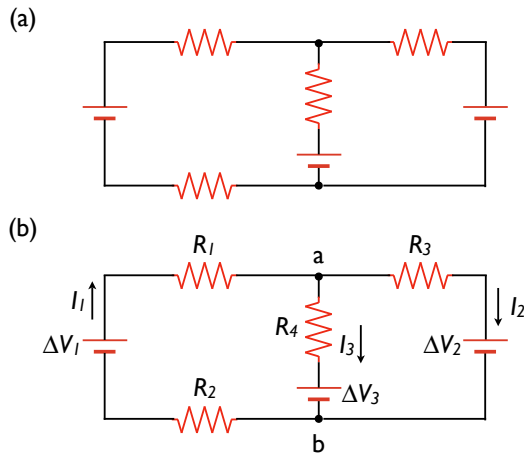


Fig. 5.14 Analyzing a complex circuit with Kirchhoff's rules. (a) Circuit as-given, (b) circuit with symbols and directions of currents assigned.

The next step is to apply the loop rule as many times as possible. In this case, we have two distinct loops: a loop on the left, including only currents I_1 and I_3 , and a loop on the right, including only currents I_2 and I_3 . We will start by applying the loop rule to the left loop, traversing clockwise from point **b** and remembering to follow the sign conventions:

$$0 = -I_1 R_2 + \Delta V_1 - I_1 R_1 - I_3 R_4 - \Delta V_3 \quad (5.44)$$

Make sure you understand why each term has the sign that it does, using Fig. 5.13 as a reference if necessary. Next, we can apply the loop rule to the right side loop, this time starting from point (a) and moving clockwise:

$$0 = -I_2 R_3 - \Delta V_2 + \Delta V_3 + I_3 R_4 \quad (5.45)$$

Still, we have too many unknowns and not enough equations. The next step is to apply the junction rule at points (a) and (b):

$$I_1 = I_3 + I_2 \quad (5.46)$$

$$I_3 = I_1 - I_2$$

In fact, we get the same result applying the rule at either point. This makes sense, based on conservation of charge and the way the circuit is set up. Sometimes applying the loop and junction rules give you duplicate results, this is not a problem *per se*. So how do we solve this mess of equations? First, let's put in the numbers we already know into Eqns. 5.44, 5.45, and 5.46, so we know what we have to find yet in the first place:

$$(5.46) \mapsto I_1 = I_2 + I_3 \quad (5.47)$$

$$(5.44) \mapsto 0 = -4I_1 + 19 - 6I_1 - I_3 - 2 \\ = 17 - 10I_1 - I_3 \quad (5.48)$$

$$(5.45) \mapsto 0 = -4I_2 - 6 + 2 + I_3 \\ = -4 - 4I_2 + I_3 \quad (5.49)$$

Now, let's rearrange and simplify Eqns. 5.47 and 5.48, and then substitute one into the other:

$$I_3 = I_1 - I_2 \quad (5.50)$$

$$10I_1 + I_3 = 17 \quad (5.51)$$

$$\implies 10I_1 + (I_1 - I_2) = 11I_1 - I_2 = 17 \quad (5.52)$$

Next, do the same thing for Eq. 5.47 and Eq. 5.49:

$$I_3 = I_1 - I_2 \quad (5.53)$$

$$4I_2 - I_3 = -4 \quad (5.54)$$

$$\implies 4I_2 - (I_1 - I_2) = -I_1 + 5I_2 = -4 \quad (5.55)$$

We're nearly done. Notice the similarity of Eq. 5.52 and 5.55 ... let's multiply Eq. 5.52 by 5, and add that to equation 5.55:

$$\begin{array}{r} 55I_1 - 5I_2 = 85 \\ + \quad -I_1 + 5I_2 = -4 \\ \hline 54I_1 = 81 \end{array} \quad (5.56)$$

This gives us $I_1 = 1.5$ A. Now that we know I_1 , we can put that into Eq. 5.55 and solve for I_2 :

$$-1.5 + 5I_2 = -4 \implies I_2 = -0.5 \text{ A} \quad (5.57)$$

Finally, we can use Eq. 5.50 to determine that $I_3 = 2.0$ A. Since both I_1 and I_3 came out positive, this means our original guess for the directions was correct. However, since I_2 came out negative, that means our initial guess was incorrect, and the I_2 actually goes the other direction. That was it!

5.6 RC Circuits

So far we have worried only about circuits with constant currents. In this section, we will start to analyze circuits whose current varies with time, though it is still in a single direction. The first example we will consider is Figure 5.15a, a resistor, capacitor, a voltage source, and switch in series. The switch S is open at first, and then suddenly closed. What happens? Before the switch S is closed, no current can flow in the circuit. We also know that if we wait for a long enough time after closing the switch, there can be no current – the capacitor will be charged to a value $Q = C\Delta V$, but nothing else will happen.

As soon as the switch is closed, the voltage source ΔV begins to charge the capacitor C . What the voltage source really wants to do is drive charges through the resistor and capacitor to create a current. It can't create a *steady-state* current in the capacitor, as we know, but the source is persistent, and keeps sending charge to the capacitor as long as it can. It will keep doing this until the capacitor is fully charged to its maximum value of $Q = C\Delta V$. The flow of charges out of the source into the capacitor is, while it is going on, a *current*. The main difference now is that we know this current can't continue indefinitely, there is only a current present between the time we close the switch S and the time when the capacitor is fully charged. In the end, the current into the capacitor driven by the source diminishes over time, until it cannot pump any more charge into the capacitor. Once the capacitor is full, we have reached our steady state of zero current.

We know the charge on the capacitor increases as a function of time, but in what fashion? There is a simple formula for this, but the math behind its derivation is a bit tedious, and we will just present the result. If we assume the capacitor is totally uncharged before we close the switch, and we call the time at which the switch is closed $t = 0$, the charge on the capacitor varies according to:

$$q(t) = Q \left(1 - e^{-t/RC} \right) \quad (5.58)$$

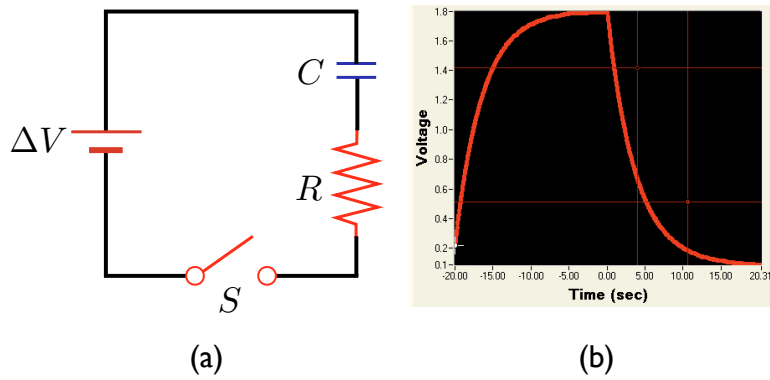


Fig. 5.15 (a) A capacitor in series with a resistor, switch, and battery. When the switch S is thrown, the capacitor begins to charge. After one time constant τ , the charge is 63% of the maximum value, $0.63C\Delta V$. (b) Measured voltage on a $2200\mu\text{F}$ capacitor in series with a 1500Ω resistor and 2.0 Volt source as a function of time. The voltage is applied at $t = -20$ sec, the capacitor charges for $t < 0$. At $t = 0$, the voltage is turned off, and the capacitor discharges for $t > 0$. In this case $\tau \approx 3.3$ sec.

where $e = 2.71828\dots$ is Euler's number, the base of natural logarithms (\ln). This is what you see in the left part of the plot in Fig. 5.15b. We can see from this equation that the charge at $t = 0$ is zero ($q(0) = 0$), and approaches its maximum value of Q as $t \rightarrow \infty$ ($q(\infty) = Q$). We can write the voltage on the capacitor as a function of time as well, since the relationship $\Delta V_C(t) = q(t)/C$ must still be true:

$$\Delta V_C(t) = \frac{Q}{C} \left(1 - e^{-t/RC}\right) = \Delta V \left(1 - e^{-t/RC}\right) \quad (5.59)$$

In principle, this equation tells us that it would take an *infinite* amount of time to fully charge the capacitor. This is just mathematics – the equation *doesn't* know that charge is quantized and comes in discrete bits of $e = 1.6 \times 10^{-19}$ C.

The term RC that appears in Equations 5.58 and 5.59 is curious. As it turns out, the units of RC end up being *time*, and the quantity RC we call the *time constant*, τ .

Time constant τ of an RC circuit:

$$\tau = RC \quad (5.60)$$

This gives τ in seconds [s] when R is in Ohms [Ω] and C is in farads [F].

What this means is that the product of the resistance and capacitance determine how long it takes to charge the capacitor! If we wait a time τ after throwing the switch, one time constant, our capacitor has charged to 63.2% ($= 1 - 1/e$) of its maximal value Q . If you substitute $t = \tau = RC$ in Equation 5.58 you can easily verify this. What is important is that the larger τ is, the longer it takes to charge a capacitor, and the smaller τ is, the more quickly it charges. At ten time constants ($t = 10\tau$), the capacitor is over 99.99% charged.

Ok. What happens if we wait a long time, the capacitor is essentially fully charged now, and we open switch S again? Well, all the charge we put on the capacitor is going to come right back out. Just before we close the switch, the voltage on the capacitor is Q/C . Once we close the switch, the charge flows back out of the capacitor into the resistor. Charges first leave the bottom plate in Figure 5.15a and enter the resistor, which lets some charges move from the top plate to the bottom plate of the capacitor. Lather, rinse, repeat, and after some time the capacitor is completely discharged.

If we close the switch at $t = 0$, the charge on the capacitor varies as:

$$q = Qe^{-t/RC} = Qe^{-t/\tau} \quad (5.61)$$

Again the time scale is in units of RC - after one time constant τ , we have now *lost* 63.2% of the charge, so $q = 0.368Q$. We can write the voltage on the capacitor down too:

$$\Delta V_C = \frac{Q}{C} e^{-t/\tau} = \Delta V e^{-t/\tau} \quad (5.62)$$

Now we can better explain Figure 5.15b. In the experimental setup used (identical to your lab hardware), $R = 1500\,\Omega$, $C = 2200\,\mu\text{F}$, and $\Delta V = 2.0\,\text{V}$. At $t = 20\,\text{sec}$ in the graph, the switch is closed, and the capacitor begins to charge. At $t = 0$, about 6 time constants later, the capacitor is about 99.8% charged and the switch is opened again. The capacitor discharges, and another 6τ later it is nearly fully discharged.

5.7 Miscellaneous

Christmas tree lights:

Imagine for a minute that the resistors in Figure 5.4 are light bulbs. What happens if one of them fails? In particular, if the bulb's filament breaks, since there is only a single current running through both bulbs, there is an open-circuit condition and *no current at all*. Both bulbs go dark, even though only one is broken. You can imagine now why it is a bad idea to wire many, many resistors or light bulbs together in series some times – a single point of failure renders everything useless.

Probably you have experienced this problem with older holiday lights. These lights are wired in series, so if any single bulb on the string becomes “open,” *no* bulbs will light. Modern equivalents have an internal “shunt” that activates when the bulb's filament burns out to avoid this problem. When the filament breaks, they actually short-circuit the bulb to keep the circuit continuous and the other bulbs in the string lit.

See http://en.wikipedia.org/wiki/Christmas_lights for a more in-depth discussion.

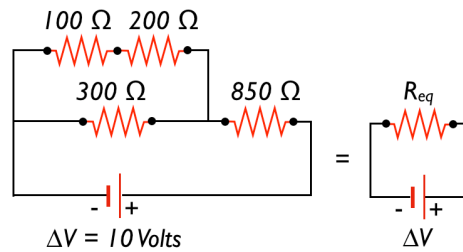
5.8 Problems

Solutions begin on page 270.

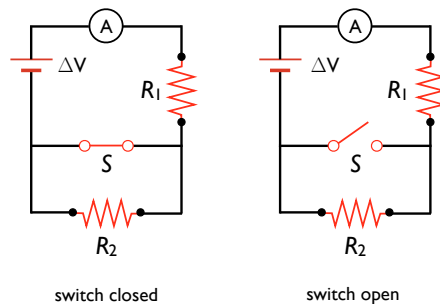
5.1. In order to maximize the percentage of the power that is delivered from a battery to a device, the internal resistance of the battery should be ...

5.2. Two resistors connected in series are measured to have an equivalent resistance of $1000\ \Omega$. The same two resistors in *parallel* are measured to have an equivalent resistance of $250\ \Omega$. What are the values of the resistors?

5.3. What is the equivalent resistance R_{eq} for the circuit below?

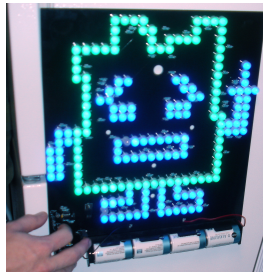


5.4. Refer to the figures below. What happens to the reading on the ammeter when the switch S is opened?

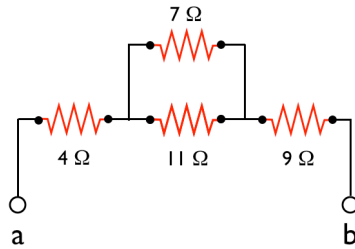


5.5. A light bulb has a resistance of $230\ \Omega$ when operated at a voltage of 120 V . What is the current in the bulb? Recall $1\text{ mA} = 10^{-3}\text{ A}$.

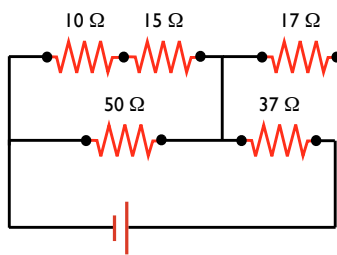
5.6. Consider the device below.[18] It takes approximately 135 light-emitting diodes (LEDs) to make up Err, second in command of the Mooninite Army. If each LED has a resistance of $200\ \Omega$ while lit, and all of the LEDs are in parallel, what is the equivalent resistance of Err?



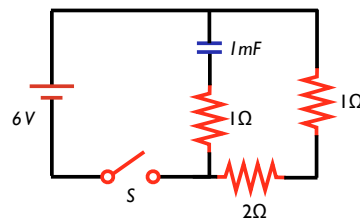
5.7. What is the equivalent resistance between points a and b ?



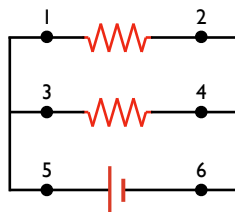
5.8. What is the equivalent resistance of the five resistors below?



5.9. The switch S is suddenly closed in the figure below. What will the steady-state current be in the $2\ \Omega$ resistor?



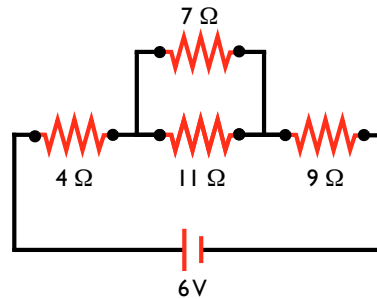
5.10. Rank the currents at points 1, 2, 3, 4, 5, and 6 from *highest to lowest*. The two resistors are identical.



5.11. Two 1.60 V batteries - with their positive terminals in the same direction - are inserted in series into the barrel of a flashlight. One battery has an internal resistance of $0.270\,\Omega$, the other has an internal resistance of $0.151\,\Omega$. When the switch is closed, a current of 0.600 A passes through the lamp. What is the lamp's resistance?

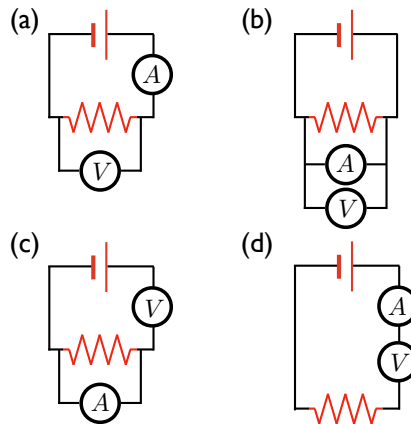
5.12. A flashlight uses a 1.5 V battery with a negligible internal resistance to light a bulb rated for a maximum power of 1 W. What is the maximum current through the bulb? Assume that the battery has more than enough capacity to drive this current, *i.e.*, it is ideal.

5.13. What is the current through the $9\,\Omega$ resistor in the figure below?



5.14. A 9 V battery with a $1\,\Omega$ internal resistance is connected to a $10\,\Omega$ resistor. What is the actual voltage across the $10\,\Omega$ resistor? Assume that the battery behaves as an ideal voltage source of 9 V in series with its internal resistance.

5.15. Refer to the figure below. Which circuit properly measures the current and voltage for the resistor? You may assume that the voltmeters and ammeters are perfect, and the battery is ideal.



5.16. A regular tetrahedron is a pyramid with a triangular base. Six $14.0\,\Omega$ resistors are placed along its six edges, with junctions at its four vertices. A 9.0 V battery is connected to any two of the vertices. **(a)** Find the equivalent resistance of the tetrahedron between these vertices. **(b)** Find the current in the battery.

5.17. A group of students on spring break manages to reach a deserted island in their wrecked sailboat. They splash ashore with fuel, a European gasoline-powered 240 V generator, a box of North

American 100 W, 120 V lightbulbs, a 500 W 120 V hot pot, lamp sockets, and some insulated wire. While waiting to be rescued they decide to use the generator to operate some bulbs.

(a) Draw a diagram of a circuit they can use, containing the minimum number of lightbulbs with 120 V across each bulb, and no higher output.

(b) One student catches a fish and wants to cook it in the hot pot. Draw a diagram of a circuit containing the hot pot and the minimum number of lightbulbs with 120 V across each device, and not more. Find the current in the generator and its power output.

5.18. You need a $45\,\Omega$ resistor, but the stockroom has only $20\,\Omega$ and $50\,\Omega$ resistors. How can the desired resistance be achieved under these circumstances?

Chapter 6

Magnetism

In science one tries to tell people, in such a way as to be understood by everyone, something that no one ever knew before. But in poetry, it's the exact opposite. – Paul Adrien Maurice Dirac

Abstract Magnetism is a crucially important areas of applied physics, more so than you may be aware. Everything from motors to loudspeakers to Magnetic Resonance Imaging (MRI) relies on magnets and magnetic fields. Though magnetism may seem like a phenomena completely distinct from electricity (and often less intuitive), in fact they are both different aspects of the unified force of “electromagnetism.” Using what we have learned from special relativity, we will be able to *prove* that electric and magnetic fields are really the same thing.

The electric fields and potentials we studied in Chapters 2 and 3 resulted from static distributions of electric charges in space. Magnetic fields, on the other hand, come from moving charges - the electric currents we studied in Chapters 4 and 5.

Magnetic fields affect moving charges, and conversely, moving charges produce their own magnetic fields. Another aspect of this symmetry between electric and magnetic fields is that time-varying magnetic fields induce electric fields, and *vice versa*. Electric and magnetic fields are fundamentally linked in their behavior in the time domain - the static aspect of one field is no more than the dynamic manifestation of the other.

However, It was not until 1820 that a formal link was established between the sciences of Electrostatics and Current Electricity and magnetism. In that year Ørsted (Fig. ??) discovered that a magnetic compass needle was deflected by an electric current – in other words, electric currents produce magnetic fields. Within a few short months, Ampère (Fig. ??) had developed a theory integrating electricity and magnetism. This theory is symbolized by the notion of equivalence of a magnetic dipole (*e.g.*, a bar magnet or a solenoid) and an electric dipole.

6.1 Magnetic Fields and Forces

While studying electric fields and forces, we described the interactions between charged objects in terms of electric fields. We said that an electric field surrounds any electric charge (or charge distribution), and that the presence of an external electric field causes electric charges (or charge distributions) to accelerate. When charged objects are stationary, knowledge of external electric fields and the object’s own electric field is sufficient to describe the static interactions between them.

The situation is different when charges are moving relative to one another or an external observer. Our first experience with moving charges was in the form of an *electric current*, the net flow of charges through some region in space. We discussed the relation between current and electric potential, but curiously neglected to discuss the electric fields around moving charges. Indeed, the interaction between moving charges is qualitatively different in many respects, and for this reason the *magnetic field* is introduced. Moving charges are said to give rise to magnetic fields, which are treated separately from (but on equal footing with) electric fields. What we should not lose sight of is that, in fact, electric and magnetic fields represent *the same fundamental force of electromagnetism*, merely in different guises.

In this picture, in addition to containing an electric field, the region of space surrounding any *moving* electric charge also contains a magnetic field. A magnetic field also surrounds a magnetic

substance making up a permanent magnet. This is because permanent magnets can be viewed in some sense as being made up microscopically of tiny current loops.¹

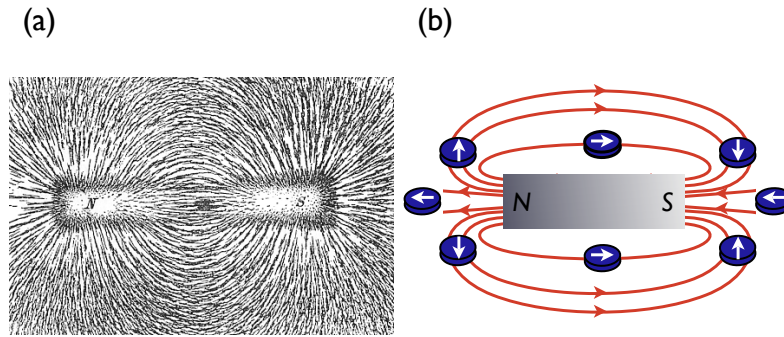


Fig. 6.1 (a) Field lines from a bar magnet, as visualized by spreading iron filings around the magnet.[19] (b) Schematic illustrating the magnetic field lines from a bar magnet.

Historically, the symbol B has been used to represent a magnetic field. The direction of the magnetic field B at any location is the direction in which a compass needle would point at that location - magnetic field is a vector \vec{B} just as the electric field \vec{E} is. As with the electric field, we can represent the magnetic field by means of drawings with magnetic field lines. Figure 6.1 shows how the magnetic field lines of a bar magnet behave. Magnetic field lines point away from north poles, and toward south poles, as electric field lines point away from positive charges and toward negative charges. The main difference between the magnetic and electric aspects of the electromagnetic force is that *there are no isolated magnetic charges, magnets always come in north-south pole combinations*. You can verify this by breaking a magnet in half - this does not separate the poles, but produces two magnets with two poles each.

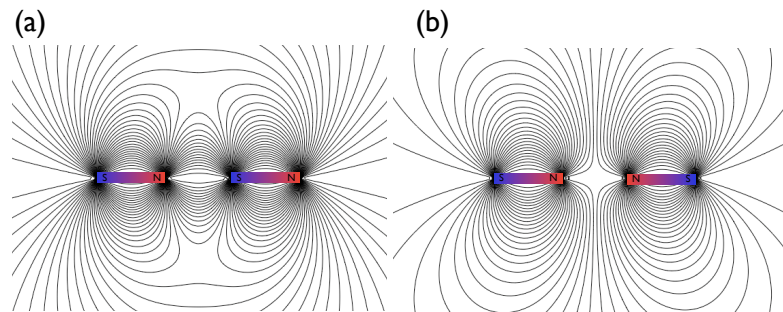


Fig. 6.2 (a) Magnetic field pattern surrounding two bar magnets aligned N-S. Note that the field is reinforced in the region between the two magnets. (b) Magnetic field pattern surrounding two bar magnets aligned N-N. Note that the field is weaker between the two magnets, and *cancels* along a vertical line equidistant between them.

Figure 6.1a displays the field lines around an ordinary bar magnet, as visualized by spreading iron filings around the magnet, while Figure 6.1b shows a schematic illustration of the field lines and direction. Figure 6.2 shows the field lines for two bar magnets brought close together, connected ‘north-north’ and ‘north-south’. In the region between two opposite poles, Fig. 6.2a, the field lines are straight lines, representing a constant magnetic field of uniform direction. This is what happens

¹ A modern quantum physics view of the problem recognizes that electrons themselves have tiny magnetic moments, called *spin*, which are the cause of most magnetism we are familiar with. This does not affect our discussion, however.

when you break a single bar magnet in half, and move the pieces apart. Between like poles, Fig. 6.2b, the magnetic field vanishes where the fields from each pole cancel, and the field lines repel each other.

Associated with the presence of a magnetic field is a certain amount of potential energy, as with an electric field. Though we will not go into detail here, the energy tied up per unit volume goes as the square of the magnetic field-line density.

6.1.1 The Magnetic Force

Also associated with a magnetic field \vec{B} at some point in space, there is a magnetic force \vec{F}_B which affects a charged particle moving in that point in space. For now, let us assume that there are no electric or gravitational fields are present, only a magnetic field \vec{B} . If test charge q moves with a velocity \vec{v} in a magnetic field \vec{B} , the magnetic force \vec{F}_B has the following properties:

Properties of magnetic fields:

1. The magnitude $|\vec{F}_B|$ of the magnetic force exerted on the particle is proportional to the charge q and to the speed $|\vec{v}|$ of the particle.
2. The magnitude and direction of \vec{F}_B depend on the velocity of the particle and on the magnitude and direction of the magnetic field \vec{B} .
3. When a charged particle moves parallel to the magnetic field vector, the magnetic force acting on the particle is zero.
4. When the particle's velocity vector makes any angle $\theta \neq 0$ with the magnetic field, the magnetic force acts in a direction perpendicular to both \vec{v} and \vec{B} . In other words, \vec{F}_B is perpendicular to the plane formed by \vec{v} and \vec{B} .
5. The magnetic force exerted on a positive charge is in the direction opposite the direction of the magnetic force exerted on a negative charge moving in the same direction.
6. The magnitude of the magnetic force exerted on a moving charged particle is proportional to $\sin\theta_{vB}$, where θ_{vB} is the angle between particles velocity \vec{v} and \vec{B} .

Figure 6.3 illustrates the direction of the magnetic force, magnetic field, and velocity vectors.

These properties can be nicely contained in a single vector equation:

Magnetic Force on a Charged Particle:

$$\vec{F}_B = q\vec{v} \times \vec{B} \quad \text{or} \quad |\vec{F}_B| = q|\vec{v}||\vec{B}|\sin\theta_{vB} \quad (6.1)$$

where θ_{vB} is the angle between the particle's velocity \vec{v} and the magnetic field \vec{B} . \vec{B} has units of teslas (T). \vec{F}_B is perpendicular to both \vec{B} and \vec{v} .

The SI unit of magnetic field strength is the tesla (T), whereas the SI unit of magnetic flux (magnetic field lines flowing through some area, like electric flux) is the weber (Wb). 1 weber = 1 tesla flowing through 1 square meter, and is a very large amount of magnetic flux. If the magnetic force is in newtons, velocity in meters per second, and magnetic field in tesla, we can see from Equation 6.1 that a charge of 1 C moving perpendicularly to a magnetic field of 1 T with a speed of 1 m/s experiences a force of 1 N. Of course, we can also see that for a stationary particle or any uncharged particle, there is no force, and there is also *no force when \vec{v} is parallel to \vec{B}* . Since the magnetic force is always perpendicular to the velocity, it never changes the energy of the charge it acts on. However, since the magnitude of the magnetic force does depend on the charge, it cannot strictly be classified as a conservative force either.

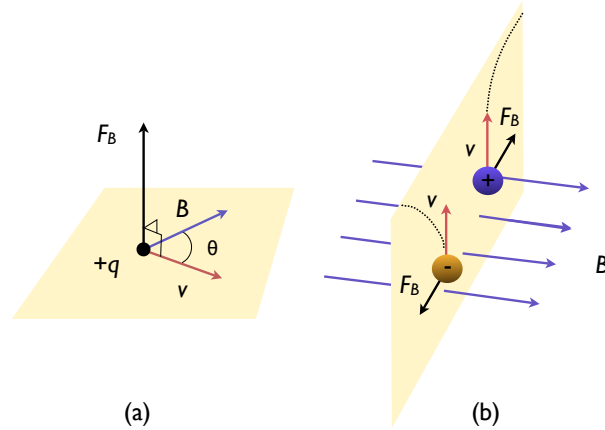


Fig. 6.3 (a) The direction of the magnetic force on a positively charged particle moving with a velocity \vec{v} in the presence of a magnetic field. When \vec{v} is at an angle θ with respect to \vec{B} , the magnetic force is perpendicular to both \vec{v} and \vec{B} . (b) Oppositely directed magnetic forces \vec{F}_B are exerted on two oppositely charged particles moving at the same velocity in a magnetic field. The dashed lines show the paths of the particles.

Figure 6.3 illustrates the vector relationship between the force \vec{F}_B , the velocity of a *positively* charged particle \vec{v} , and the magnetic field \vec{B} . The force \vec{F}_B will act in *opposite directions* on positively and negatively charged particles as the electric field does, and just as we expect from Eq. 6.1 since \vec{F}_B is proportional to q . This is important to keep in mind, particularly since electric currents are almost invariably made up of moving *negatively charged* electrons, while mass spectrometers (Sect. 6.3.1.1) often involve the motion of *positively charged* ions. In other words, you will have to deal with both positive and negative cases, so be careful about signs and directions!

6.1. An electron passes through a magnetic field without being deflected. What can you say about the angle between the magnetic field vector and the electron's velocity, if no other forces are present?

- They could be in the same direction
- They could be perpendicular
- They could be in opposite directions
- Both the first and third are possible

6.2. Consider a proton moving with a speed of $|\vec{v}| = 1 \cdot 10^5$ m/s through the earth's magnetic field ($|\vec{B}| = 55 \mu\text{T}$). When the proton moves east, the magnetic force acts straight upward. When the proton moves northward, no force acts on it. What is the direction of the magnetic field?

- North
- South
- East
- West

6.3. What is the magnitude of the magnetic force in the previous example?

6.1.2 Magnetism as a Consequence of Relativity

Thus far, we have considered the electric fields and potentials of static charge configurations, and even discussed at length the flow of charges that make up electric currents. We have curiously omitted a discussion of what the electric field of a moving charge looks like, however. As it turns out, the magnetic field we normally think of as a distinct physical phenomena is nothing more than a relativistic view of the electric field of moving charges. Before we delve deeper into magnetic phenomena, we will first demonstrate how magnetic fields are nothing more than a consequence of relativistic length contraction.

In order to see the fundamental symmetry between the electric and magnetic fields, we will conduct a hypothetical experiment using a current-carrying wire and a moving test charge, as shown

in Fig 6.4. We have a conducting wire with current flowing to the right when viewed from the laboratory reference frame (O). For simplicity, we will assume the current is due to the flow of *positive* charges, spaced evenly with an average separation l^O when viewed from the lab frame O .²

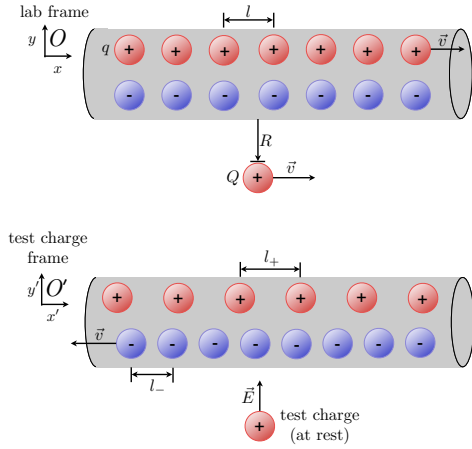


Fig. 6.4 An electric current in a wire viewed from the laboratory reference frame (O), and the reference frame of a moving test charge Q (O'). In the test charge frame, the spacing of the positive charges apparently *increases* while the spacing of the negative charges apparently *decreases*.

We know that our conducting wire must be electrically neutral in the laboratory frame, so in addition to the positive charges there must be an equal number of negative ions – the atoms making up the wire – also spaced at a distance l^O . Now (still in the laboratory frame) we place a positive test charge Q a distance R from the wire. Since the wire is electrically neutral, there is no force on the test charge. What happens if the test charge is moving? We will give the test charge Q a velocity \vec{v} parallel to the wire, the same velocity with which the positive charges in the wire are moving for simplicity.

What does the now moving test charge experience, viewed from its own reference frame (O')? Since it is moving in the same direction, with the same velocity, as the positive charges in the wire, *it sees those positive charges as at rest relative to itself*, and *the negative charges as moving to the left with velocity \vec{v}* .

When the positive charges are viewed from the laboratory frame O , they appear to have an average spacing of l^O , moving at velocity \vec{v} . Once we switch to the test charge's frame, the positive charges appear to be at rest – in switching reference frames, the velocity of the positive charges goes from \vec{v} to zero. From special relativity (Ch. 1), we know that moving objects undergo a *length contraction*. When we view the spacing l^O of the positive charges in the lab frame O , *we are viewing the contracted length*. In the test charge's frame O' , we must un-contrast the spacing l^O into the O' frame to figure out what the test charge really sees. If we call the spacing of the positive charges that the moving test charge experiences in its frame O' as $l_+^{O'}$, we can easily relate it to the spacing viewed from the lab frame O :

$$l_+^{O'} = l^O \gamma \quad (6.2)$$

$$l_+^{O'} = \frac{l^O}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (6.3)$$

Since we know $\gamma \geq 1$, it is clear that the spacing the test charge sees is *larger* than what we see in the lab frame. Meanwhile, what about the negative charges, which are stationary in the lab frame? The test charge sees from its frame the negative charges moving to the *left* with velocity \vec{v} , so their spacing must be *contracted* to figure out the spacing of the negative charges $l_-^{O'}$ the test charge sees:

² Even though we know now that negatively-charged electrons really carry the current, working with positive charges will make the discussion simpler (by avoiding a lot of pesky minus signs), and will not change the analysis in any way.

$$\gamma l_-^{O'} = l^O \quad (6.4)$$

$$l_-^{O'} = \frac{l^O}{\gamma} \quad (6.5)$$

$$l_-^{O'} = l^O \sqrt{1 - \frac{v^2}{c^2}} \quad (6.6)$$

Again, since $\gamma \geq 1$, the positive test charge sees a *reduced* spacing of the negative charges. Since the positive and negative charges now no longer appear to have the same spacing when viewed from the test charge's frame, *the test charge sees a net negative charge density*, since there are effectively more negative charges per unit length than positive charges. The presence of a net negative charge density from the test charge's point of view means that it experiences a net attractive force from the wire. From the lab frame, we would not expect any force between the test charge and the wire, but sure enough, a proper relativistic treatment leads us to deduce that a force must in fact be present.

How big is the force? First, we need to figure out the charge density in the wire that the test charge sees. Since we don't want to restrict ourselves to any particular length of wire, we will calculate the number of charges per unit length as viewed in the test charge's frame, $\lambda^{O'}$. How do we find this? We know that all charges in the wire have charge q , and we know their average spacing. Dividing q by the average spacing for each kind of charge will give us the number of charges per unit length for both positive and negative charges, and subtracting those two will give use the net charge density:

$$\lambda^{O'} = \lambda_+^{O'} - \lambda_-^{O'} \quad (6.7)$$

$$= \frac{q}{l_+^{O'}} - \frac{q}{l_-^{O'}} \quad (6.8)$$

$$= \frac{q}{l^O} \sqrt{1 - \frac{v^2}{c^2}} - \frac{q}{l^O} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (6.9)$$

$$= \frac{q}{l^O} \left(\sqrt{1 - \frac{v^2}{c^2}} - \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \right) \quad (6.10)$$

This is a bit messy. However, we know that the drift velocity of charges in a conductor is *very* small compared to c ($v_d \sim 10^{-3}$ m/s, see Sect. 4.4.1). When $v \ll c$, we can use the following approximations:³

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \approx 1 + \frac{1}{2} \frac{v^2}{c^2} \quad v \ll c \quad (6.11)$$

$$\frac{1}{\gamma} = \sqrt{1 - \frac{v^2}{c^2}} \approx 1 - \frac{1}{2} \frac{v^2}{c^2} \quad v \ll c \quad (6.12)$$

Using these approximations in Eq. 6.10, we can come up with a simple expression for $\lambda^{O'}$:

³ These approximations come from a Taylor series expansion. Don't worry if you don't know how to derive them.

$$\lambda^{O'} = \frac{q}{l^O} \left(\sqrt{1 - \frac{v^2}{c^2}} - \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \right) \quad (6.13)$$

$$= \frac{q}{l^O} \left(1 - \frac{1}{2} \frac{v^2}{c^2} - \left(1 + \frac{1}{2} \frac{v^2}{c^2} \right) \right) \quad (6.14)$$

$$= -\frac{q}{l^O} \frac{v^2}{c^2} \quad (6.15)$$

Now that we have the charge density of the wire as viewed from the test charge's frame, what is the electrostatic force? The problem is now to find the electric field at a distance R from a long, uniformly charged wire of charge density $\lambda^{O'}$, which we already did in Sec. 2.8.6. Using Eq. 2.28, we can immediately write down the electrostatic force experienced by the test charge in its reference frame:

$$|\vec{F}| = Q|\vec{E}| = Q \frac{2k_e |\lambda^{O'}|}{R} = \frac{2k_e Q q v^2}{R l c^2} \quad (6.16)$$

We can simplify this a bit. The current in the wire is the charge q divided by the time it takes the charges to move a unit length, which is $\Delta t = l/v$.⁴ Thus the current can be written as qv/l :

$$|\vec{F}| = Qv \left(\frac{2k_e I}{c^2 R} \right) \quad (6.17)$$

If we associate the quantity in parenthesis with an effective magnetic field, then we have derived Eq. 6.1:⁵

$$|\vec{F}| = Qv|\vec{B}| \quad \text{with} \quad |\vec{B}| = \frac{2k_e I}{c^2 R} \quad (6.18)$$

This is it. A test charge moving near a current-carrying wire experiences a net force proportional to its charge, velocity, and the current in the wire. We have managed to derive the existence of the magnetic field and magnetic force from nothing more than Coulomb's law⁶ and special relativity – **a magnetic field is nothing more than the field of moving charges**. Further, by analogy with Eq. 6.1, we have established that there is a magnetic field surrounding a long, straight wire. This is perhaps the most important result – **electric currents create magnetic fields**. Electricity and magnetism really are the same thing viewed from different reference frames. Amazing, isn't it?

In some sense, it is remarkable that we can measure magnetic forces due to currents at all. The drift velocity is *miniscule* compared to c , $\frac{v}{c} \sim 10^{-12}$ or so, and γ is barely different from 1, about $1.0 + 10^{-24}$. The magnetic force results from a tiny relativistic correction, certainly, but it is indeed a significant effect in the end because there are truly astronomical numbers of charges per unit length inside conductors. Even though the force per charge is miniscule, they make up for it in numbers. Before moving on, we note that if you repeat this analysis for the more complicated case that the test charge's velocity is *not* the same as the charges in the wire, and *not* parallel, you still arrive at exactly Eq. 6.1. It just takes quite a bit longer ...

6.1.3 Magnetic Field of a Long, Straight Wire

What is the direction of the magnetic field we just derived, associated with the current in the wire? So far, we only know its magnitude, but we can figure out the direction based on symmetry. The effective field \vec{B} is due to the current in the wire, which is directed along the axis of the wire (of course). The magnitude of the magnetic field is also proportional to the current and falls off as distance increases, as one might expect. If we reverse the direction of the current, the force changes

⁴ This just comes from kinematics, we know that the charge covers a distance l according to $l = v\Delta t$.

⁵ $\theta = 90^\circ$ in our case, so $\sin \theta = 1$.

⁶ Or, equivalently, Gauss' law

sign, so the direction of \vec{B} must depend on the direction of the current. Symmetry maintains that it must be radially symmetric about the wire axis as well – in other words, the magnetic field must be constant in magnitude on circles drawn around the wire.

The force itself is directed perpendicular and toward the wire, and perpendicular to the test charge's velocity. *If the force is proportional to and perpendicular to the velocity, and proportional to the magnitude of the magnetic field, the force can only result from a vector product (or “cross product”) between the velocity and magnetic field, $\vec{F} = Q\vec{v} \times \vec{B}$.* A cross product between \vec{v} and \vec{B} fulfills all the requirements – if the force and magnetic field are perpendicular, the magnetic field must be perpendicular to *both*. For this to be true *and* still have a radially symmetric field as required by symmetry, there is only one possibility: the magnetic field circulates around the wire! This is shown schematically in Figure 6.5.

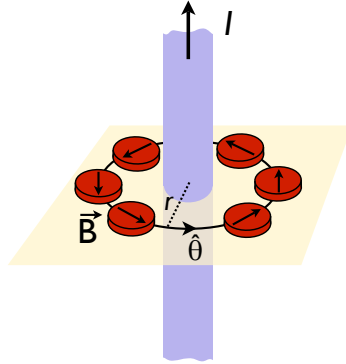


Fig. 6.5 Magnetic field around a current-carrying wire. When a current I flows, compass needles deflect in directions tangent to the circle, pointing in the direction of \vec{B} due to the current. The second right-hand rule gives the direction of \vec{B} from I , and vice-versa. The unit vector $\hat{\theta}$ is used to represent the angular direction.

Now we have a simple mathematical form of the magnetic field surrounding a current-carrying wire:

Magnetic field around a long, straight wire:

$$\vec{B} = \frac{2k_e I}{c^2 R} \hat{\theta} = \frac{\mu_0 I}{2\pi r} \hat{\theta} \quad (6.19)$$

where I is the current in the wire, r is the distance from the axis of the wire, and $\hat{\theta}$ is the angular unit vector around the wire axis.

Just like electric field and electric potential near a point charge, B *diverges* when you get infinitesimally close to the wire. The magnetic field doesn't really become *infinite*, that just means that when we get too close, we are actually *inside* the wire, and different physics must be used.

The μ_0 in Equation 6.19 is a new constant, the “permeability of free space,” and has the value

Permeability of free space:

$$\mu_0 \equiv 4\pi \times 10^{-7} \text{ T} \cdot \text{m/A} = 4\pi \times 10^{-7} \text{ N/A}^2 \quad (6.20)$$

The constants μ_0 , ϵ_0 , and the speed of light c are intimately related, so you really only have to remember two of the three:

One less constant than you think:

$$\mu_0 = \frac{1}{\epsilon_0 c^2} \quad (6.21)$$

This is the reason that μ_0 is *defined* by Equation 6.20 using “ \equiv ” instead of “ $=$ ” – the interdependence of these three constants led physicists to just define μ_0 as fixed, since c and ϵ_0 determine it uniquely anyway. If you substitute $k_e = 1/4\pi\epsilon_0$ and $\epsilon_0 c^2 = 1/\mu$ into Eq. 6.19 above, you can see that both forms are correct:

$$|\vec{B}| = \frac{2k_e I}{c^2 R} = \frac{2I}{4\pi\epsilon_0 c^2 R} = \frac{2I}{2\pi\epsilon_0 c^2 R} = \frac{\mu_0 I}{2\pi R} \quad (6.22)$$

Now there is just one more nagging point about the field surrounding the wire. Which direction does it circulate, clockwise, or counterclockwise?

6.1.4 Handedness

What we do not fully know yet is the proper sense of circulation of the magnetic field surrounding a current-carrying wire. In order to determine that, we need to think a bit deeper about three-dimensional geometry and “handedness.”

The fact that \vec{B} , \vec{v} , and \vec{F}_B are mutually perpendicular implies a unique *axis* for each, since in three dimensions there are only three mutually perpendicular axes. This fact alone does not determine a unique *direction* for all three, however. We have two possible choices for the convention of direction, corresponding to two senses of “handedness,” or two possible coordinate systems, as shown in Fig. 6.6. You may recall the same problem when learning about torque and angular momentum. Or, if you are a chemist, you know this problem as *chirality*. An object is said to be ‘chiral’ if its mirror image cannot be superimposed on the original. No amount of rotation or translation will make the mirror image look exactly like the original. Your hands are good examples - no amount of rotation or manipulation will change a left hand into a right hand, hence the name. This is clearly the case for the two coordinate systems in Fig. 6.6a and Fig. 6.6b, or the two helices in Fig. 6.6c.

The magnetic force is in some sense chiral. Looking back at Fig. 6.3, if we were to reverse the direction of \vec{v} , then we would also have to reverse the direction of \vec{B} , but not \vec{F}_B . Similarly, we could reverse both \vec{v} and \vec{F}_B and \vec{B} would be left unchanged.⁷ The diagram of \vec{B} , \vec{v} , and \vec{F}_B in Fig. 6.3 is not equivalent to its mirror image, and is hence chiral. We will not dwell on this point, further, but suffice it to say, as a convention we always choose the *right handed* coordinate system.

We can easily pick which is the right-handed coordinate system and choose the proper directions of \vec{B} , \vec{v} , and \vec{F}_B , with a simple rule, the **right-hand rule number 1**:

Right-hand rule # 1:

1. Point the fingers of your right hand along the direction of the velocity.
2. Point your thumb in the direction of the magnetic field \vec{B} .
3. The magnetic force on a positive charge points out from the back of your hand.

-OR-

1. Point your fingers in the direction of \vec{v} .

⁷ This is a result of the fact that the magnetic field is technically a *pseudovector*, not a true vector. Pseudovectors act just like real vectors, except they gain a sign flip under improper rotation.[20] An improper rotation is an inversion followed by a normal (proper) rotation, just what we are doing when we switch between right- and left-handed coordinate systems. A proper rotation has no inversion step, just rotation.

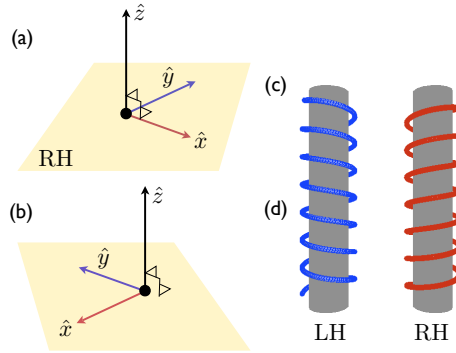


Fig. 6.6 (a) A right-handed coordinate system (b) A left-handed coordinate system. Can you see that **a** and **b** are not equivalent? Try rotating them in your head. (c) A right-handed and a left-handed helix. Normal DNA is a right-handed helix, though left-handed DNA does exist. No definitive biological significance of “left-handed”-DNA has yet been shown.

2. Curl your fingers in the direction of \vec{B} , moving through the smallest angle.
3. Your thumb now points in the direction of the magnetic force for a positive charge.

Both forms of the right-hand rule (should) give you the same result, use whichever is more intuitive for you. Note that if you replace \vec{v} with x , \vec{B} with y , and \vec{F}_B with z , the same rules let you choose a right-handed coordinate system. For a current-carrying wire, we can come up with a more specific rule, since the velocity of the charges making up the current is always along the axis of the wire. This rule is unimaginatively called the second right-hand rule:

Right-hand rule #2:

Point your thumb on your right hand along the wire in the direction of the current. Your fingers naturally curl around the direction of the magnetic field caused by the current, which circulates around the wire.

6.4. Consider a proton moving with a speed of $1 \cdot 10^5$ m/s through the earth’s magnetic field ($|\vec{B}| = 55 \mu\text{T}$). When the proton moves east, the magnetic force acts straight upward. When the proton moves northward, no force acts on it. What is the direction and magnitude of the magnetic field?

6.2 Ampère’s Law

Equation 6.19 lets us calculate the magnetic field due to a long, straight wire, but not much else. Deriving everything from electrostatics and special relativity is certainly too tedious for common usage. A more general technique is due to André-Marie Ampère, it is much in the spirit of Gauss’ law (Sect. 2.8). “Ampère’s law” relates the current flowing through a closed surface to the magnetic field tangential to the curve bounding the surface.

Take any arbitrary closed path surrounding a current, as in Figure 6.7, and break it up into infinitesimal segments Δl . Now find the component of the magnetic field parallel to the segment, B_{\parallel} , and compute the product $B_{\parallel} \Delta l$. The sum of all such products around the closed path gives the current passing through the surface bounded by the path:

Ampère’s law:

$$\sum_{\text{closed path}} B_{\parallel} \Delta l = \mu_0 I_{\text{enclosed}} \quad (6.23)$$

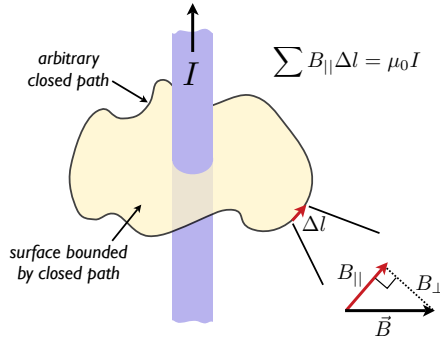


Fig. 6.7 Ampère's Law. Take any arbitrary closed path surrounding a current, and break it up into infinitesimal segments Δl , and find the component of the magnetic field parallel to the segment $B_{||}$. The sum of all such products around the closed path gives the current passing through the surface bounded by the path.

where $B_{||}$ is the field component parallel to the segment Δl , and I_{enclosed} is the current passing through the surface defined by the closed path.

Again, just like with Gauss' law, we choose particularly convenient paths around a current, such that everywhere on the path B is either perfectly parallel, or perfectly perpendicular. Unlike Gauss' law, we have to be careful about the *direction* in which we trace out the path.

Take the long, straight wire carrying a current I . We know from symmetry that the magnetic field must be radially symmetric about the wire, so we choose our Ampèrian paths to be circles centered on the wire, just like we chose spheres as our Gaussian surfaces surrounding point charges (Sec. 2.8).

By symmetry, the magnetic field has the same value everywhere on the circle, and must be tangential to the circle. That is, $B_{||} = B$ for every segment Δl on the wire. Computing the current is now easy, since $B_{||}$ can just be taken out of the sum:

$$\sum B_{||} \Delta l = B_{||} \sum \Delta l = B_{||} l_{\text{path}} = B_{||} \cdot 2\pi r = \mu_0 I \quad (6.24)$$

$$\Rightarrow B_{||} = \frac{\mu_0 I}{2\pi r} \quad (6.25)$$

This is exactly the result given by Equation 6.19, which we derived from Ampère's law and symmetry alone! While Ampère's law is very simple and elegant, it cannot easily be used for complex current configurations which lack a nice symmetry like our wire has, and *it is only valid for static cases, when the E and B fields do not vary with time*. There, however, is a slightly more complex form which is valid in general. It relates not only current, but the time variation of the electric field to $\sum B_{||} \Delta l$.

6.2.1 Ampère's and Gauss' Laws

Fundamentally, both Gauss' law and Ampère's law are manifestations of the *divergence theorem* (a.k.a. Green's theorem or the Gauss-Ostrogradsky theorem). Put simply, it states that *the sum of all sources minus the sum of all sinks gives the net flow out of a region*. The same law applies in fluid dynamics. If a fluid is flowing, and we want to know how much fluid flows out of a certain region, then we need to add up the sources inside the region and subtract the sinks. The divergence theorem is basically a conservation law - the volumetric total of all sources minus sinks equals the flow across a volume's boundary.

In the case of electric fields, this gives Gauss' law (Equation 2.9) - that the electric flux through any closed surface must relate to a *net charge* inside the volume bounded by that surface. In the case of magnetic fields, *the same law* applies, but we know there are no unpaired "magnetic charges" - magnets always come in north-south pairs. Therefore, any closed surface always encloses *pairs* of magnetic poles, and there can be no net "magnetic charge" inside. Thus the net magnetic flux

$\Phi_B = BA \cos \theta_{BA}$, defined similarly to electric flux (Equation 2.5) out of any *closed surface* bounding a volume is zero.

Definition of magnetic flux through a surface

$$\Phi_B = BA \cos \theta_{BA} \quad (6.26)$$

where θ_{BA} is the angle between the surface normal and the magnetic field.

Given *any* volume element, the net magnitude of the vector components of the magnetic field that point outward from the surface must be equal to the net magnitude of the vector components that point inward. This means that the magnetic field lines must be closed loops. Another way of putting it is that magnetic field lines cannot originate from somewhere – following the lines backwards or forward leads back to the starting position. Hence, this is a mathematical formulation of the statement that there are no single magnetic poles. Magnetic poles *always* come in north-south pairs, never alone.

By analogy, the net magnitude of the vector components of the *electric field* pointing outward must be equal to the net magnitude of the vector components pointing inward *plus* the amount of free charge inside. Electric field lines do originate from somewhere - from charges.

Gauss' laws:

The electric flux Φ_E through any **closed** surface is equal to the net charge inside the surface, Q_{inside} , divided by ϵ_0 :

$$\Phi_{E,\text{closed surface}} = \frac{Q_{\text{inside}}}{\epsilon_0} \quad (6.27)$$

The magnetic flux Φ_M through any **closed** surface bounding a volume is zero:

$$\Phi_{B,\text{closed surface}} = 0 \quad (6.28)$$

The fact that magnetic flux out of a closed surface is zero gives us Ampère's law. If there can be no net magnetic flux out of a closed region, then the tangential components of the magnetic field around any closed curve we draw on the surface must sum to zero. If they did not, then adding up all such curves to build up a closed volume would *not* lead to zero magnetic flux, which would imply the existence of single magnetic poles. For even more detail about what this means for the boundary conditions on the electric magnetic fields, see Appendix ??.

What about a version of Ampère's law for electric fields? Surely Gauss' law for electric fields must imply something about the tangential components of the electric field around a closed loop. Indeed they do, and it is this bit which explains how electric generators work. But not until next chapter!

6.3 The Magnetic Field in Various Situations

6.3.1 Motion of a Charged Particle in a Magnetic Field

We have already found that a charged particle moving parallel to a current-carrying wire experiences a force directed toward the wire. Now, we wish to consider the slightly more general case of a single charged particle $+q$ placed in a constant magnetic field, such that the particle's velocity \vec{v} is perpendicular to the magnetic field \vec{B} , Figure 6.8. We know that the magnetic force \vec{F}_B will always be perpendicular to \vec{v} and perpendicular to the field \vec{B} . What is the resulting motion of the particle?

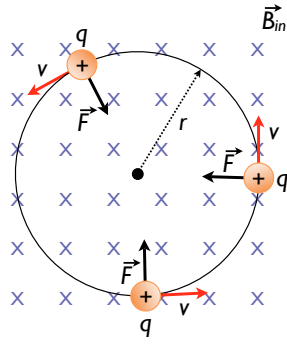


Fig. 6.8 When the velocity of a charge $+q$ is perpendicular to a uniform magnetic field, the particle moves in a circle whose plane is perpendicular to \vec{B} , which is into the page. The magnetic force \vec{F} on the charge is always directed toward the center of the circle.

Take the case of the particle at the bottom of the circle in the figure, where the particle has a velocity directed to the right. Applying the right-hand rule gives a force vertically upward. The particle curves upward as a result, and then experiences a force to the left. And so on.

More generally we might ask: what is the locus of points such that the force, velocity, and magnetic field are always perpendicular? *A circle!* **The magnetic force is always directed toward the center of a circular path**, therefore the magnetic force causes a centripetal acceleration. As we know, whenever a particle moves in a circular path, it experiences an effective centripetal force mv^2/r , which must equal the sum of all other forces. Centripetal force changes the direction of \vec{v} , but not its magnitude, so we can relate it to the magnetic force with Newton's second law:

$$F_B = F_{\text{centr.}} = qvB = \frac{mv^2}{r} \quad (6.29)$$

We can use this to find the radius of the path of a charged particle in a magnetic field:

A charged particle in a constant magnetic field moves in a circle, radius r :

$$r = \frac{mv}{qB} \quad (6.30)$$

The radius of the particle's path is proportional to its momentum mv , and inversely proportional to its charge q and the magnitude of the magnetic field B . Equivalently, we can say the radius depends on the *charge to mass ratio* of the particle, m/q .

If we know the radius of the particle's path, then Equation 6.29 says that the velocity has to be $v = \frac{qrB}{m}$. Since the particle is in uniform circular motion, we can define an *angular frequency* ω , the time it takes the particle to go around one orbit:

Angular frequency of a charged particle in a constant magnetic field

$$\omega = \frac{v}{r} = \frac{qrB}{m} \left(\frac{1}{r} \right) = \frac{qB}{m} \quad (6.31)$$

where we have used Equations 6.29 and 6.30 in the last two steps. The period of the motion can be found as well:

Period of motion of a charged particle in a constant magnetic field

$$T = \frac{2\pi r}{v} = \frac{2\pi}{\omega} = \frac{2\pi m}{qB} \quad (6.32)$$

In other words, the charged particle undergoes oscillatory motion, with a period proportional to the mass to charge ratio m/q , and inversely proportional to the magnetic field B . This is roughly the basis of one type of magnetic resonance, similar to MRI.

What happens if the initial velocity is not perfectly perpendicular to the magnetic field? The motion of the particle within the plane perpendicular to the field is still a circle, but we have to add on a constant component in the direction parallel to the field. Circular motion in one plane, and constant velocity in a perpendicular plane gives a *helix*. Think about that for a minute, and inspect Figure 6.9.

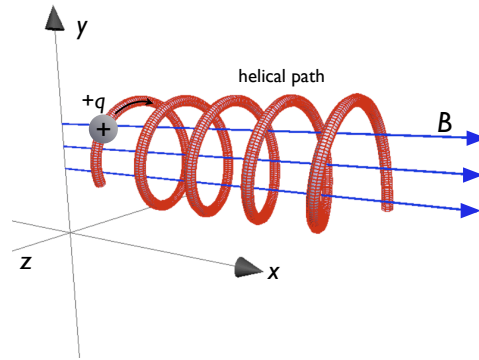


Fig. 6.9 A charged particle with an initial velocity at an angle to the magnetic field moves in a helical path.

6.3.1.1 The Mass Spectrometer

Figure 6.10 illustrates the basic operation of one type of mass spectrometer. Charged particles enter a region at left where two parallel plates create a constant electric field E in the vertical direction, while at the same time a constant magnetic field is applied into the page. The electric field causes a force qE to be exerted on the particle *upward*, while the magnetic field exerts a force qvB *downward*.

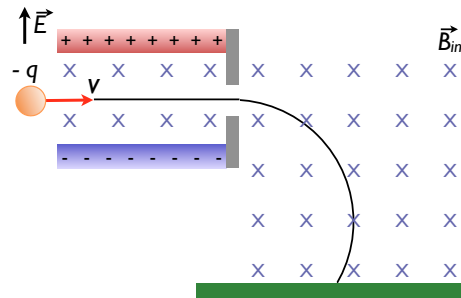


Fig. 6.10 A mass spectrometer. When the particle enters the region at left with both electric and magnetic fields, it experiences a force of $qE - qvB$ in the vertical direction. If the ratio of electric to magnetic fields equals the particle's velocity, $E/B = v$, the net force is zero, and the particle passes through an aperture. It then passes into a region with only magnetic field, where it experiences a force qvB , and follows a circular path of radius $r = mv/qB = mE/qB^2$. The heavier the particle, the farther it travels before hitting the detector plate (green). By measuring the position at which particles hit the detector plate, one can measure their mass to charge ratio m/q . A step preceding this detector ensures all particles are singly-charged, which allows precise mass (and therefore chemical) identification.

If the net electric + magnetic force is zero, the particle has no acceleration, and travels on a straight line path through a narrow aperture. For this to happen, we require:

$$qE = qvB \quad \Rightarrow \quad v = \frac{E}{B} \quad (6.33)$$

This is a “velocity selector,” which creates a stream of particles of a specific velocity based on the ratio E/B . Once the particles leave the aperture, they experience only a magnetic field, and therefore

only a force qvB directed initially downward. From the preceding section, we know that the particle's subsequent motion will be in a circular path of radius

$$r = \frac{mv}{qB} \quad (6.34)$$

If we solve Equation 6.33 for B , and substitute it in the equation above, we see that the radius of curvature can be expressed independently of its velocity:

$$r = \frac{mE}{qB^2} = \left(\frac{m}{q} \right) \frac{E}{B^2} \quad (6.35)$$

After the first part of the detector fine-tunes the particles' velocity, the second stage forces them to curve in a path that depends on their mass to charge ratio m/q . If before the detector stage we ensure that all particles are singly-charged, or at least all have the same charge, the radius of curvature is directly related to the mass of the particle. The radius of curvature, and thus the mass of the particles, can be measured by placing a position-sensitive charge detector (green box) inside the second stage of the detector. Heavier particles curve less in the magnetic field, and land farther along the green plate, while lighter ones curl in tightly and land closer to the left side of the detector plate. Mass spectrometers based on this principle can be used to identify elements or compounds (as in a mass spectrometer), or to separate isotopes of a given element.

6.3.2 Magnetic Force on a Current-Carrying Conductor

We know already how a single charged particle moves in response to a magnetic field. We also know that an electric current is nothing more than a stream of moving charges. It is easy to see, then, that a wire carrying a current should experience a force in a magnetic field. The direction of the force is perpendicular to the direction of the current and the magnetic field, in agreement with the first right-hand rule.

What is the force on a current carrying wire? Let us take a wire of length l , carrying a current I in a magnetic field of strength $|\vec{B}| \equiv B$ perpendicular to the wire's axis. If we break the current up into single charges moving at the drift velocity $|\vec{v}| \equiv v_d$, then each charge making up the current experiences a magnetic force $|\vec{F}_B| = qv_dB$. The total force on a segment of wire is the force per charge carrier times the total number of carriers in the wire. Given the number of carriers per unit volume n , and the wire's volume $A \cdot l$, we have:

$$\begin{aligned} (\text{Total force}) &= (\text{force per charge carrier}) \times (\text{number of carriers}) \\ |\vec{F}_B| &= (qv_dB)(nAl) \end{aligned} \quad (6.36)$$

We already developed Equation 4.5 relating the current to drift velocity, $I = nqv_dA$, so we are left with:

$$v_d = \frac{I}{nqA} \quad (6.37)$$

$$|\vec{F}_B| = nqv_dBAI \quad (6.38)$$

$$= nq \left(\frac{I}{nqA} \right) BAI \quad (6.39)$$

$$= IBl \quad (6.40)$$

If the wire isn't perpendicular to the magnetic field, but at some angle θ , we can repeat the analysis above with $B \sin \theta$ in place of B , and arrive at a general equation for the force experienced by current carrying wire:

Force on a current-carrying wire:

$$|\vec{F}_B| = IBl \sin \theta \quad (6.41)$$

where θ is the angle between the current and magnetic field.

Figure 6.11 shows the net force on a current-carrying wire. Note that a vector directed *into* the page is represented as a cross inside a circle, \otimes , corresponding to the tails of arrows.

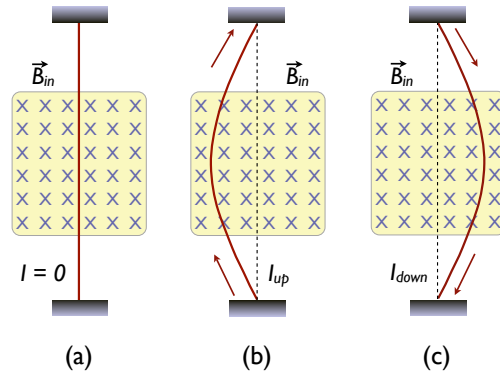


Fig. 6.11 A segment of flexible wire between the poles of a magnet, with the field (blue crosses) directed into the page. **(a)** With no current in the wire, there is no horizontal force and the wire remains horizontal. **(b)** When the current is upward, the wire deflects to the left. **(c)** When the current is downward, the force is to the right.

Conventions for drawing vectors:

A vector directed *into* the page is represented by a cross inside a circle, \otimes

A vector directed *out of* the page is represented by a dot inside a circle, \odot

6.3.3 Magnetic Force Between Two Parallel Conductors

We now know that a magnetic force acts on a current-carrying conductor when the conductor is placed in an external magnetic field. We also know that current-carrying wires create their own magnetic fields. From this it also follows that *a current-carrying wire experiences a force from another current-carrying wire*.

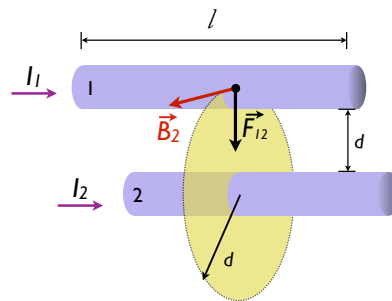


Fig. 6.12 Two parallel wires both carry steady currents and exert a force on each other. The field \vec{B}_2 at wire 1 due to wire 2 produces a force on wire 1 given by $F_{12} = B_2 I_1 l$. The force is attractive if the currents are in the same direction (shown), and repulsive if they are in opposite directions.

Figure 6.12 shows two long, straight, parallel wires carrying currents I_1 and I_2 separated by a distance d . Wire 2 produces a magnetic field \vec{B}_2 , which acts on wire 1. The direction of \vec{B}_2 is perpendicular to the wire, and must have a magnitude:

$$|\vec{B}_2| = \frac{\mu_0 I_2}{2\pi d} \quad (6.42)$$

from Equation 6.19. Equation 6.41 gives us the force \vec{F}_{12} on wire 1 due to the presence of \vec{B}_2 due to I_2 in wire 2:

$$|\vec{F}_{12}| = |\vec{B}_2| I_1 l = \left(\frac{\mu_0 I_2}{2\pi d} \right) I_1 l = \frac{\mu_0 I_1 I_2 l}{2\pi d} \quad (6.43)$$

For an arbitrary wire, we can better write this in terms of the force per unit length:

Force per unit length on wire 1 parallel to wire 2:

$$\frac{|\vec{F}_{12}|}{l} = \frac{\mu_0 I_1 I_2}{2\pi d} \quad (6.44)$$

where d is the separation between wires 1 and 2 carrying currents I_1 and I_2 , respectively.

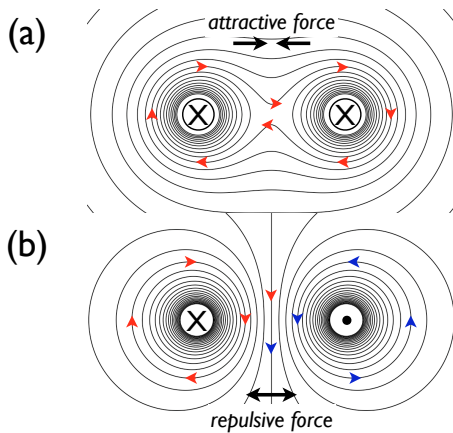


Fig. 6.13 The magnetic field lines around two parallel current-carrying wires. **(a)** When both currents are into the page, the field lines circulate clockwise for both wires, and in the region between the wires they tend to cancel one another. The net interaction is attractive. **(b)** When the currents are in opposite directions, the field is reinforced between the wires, since one has a clockwise and one a counter-clockwise circulating field. The net interaction is repulsive.

The direction of \vec{F}_{12} is downward toward wire 2, as expected from the first right-hand rule. From Newton's third law, we additionally know that $\vec{F}_{12} = -\vec{F}_{21}$, that is, the force on wire 2 is equal and opposite that on wire 1.

Two parallel wires carrying current in the same direction attract each other, and as you might expect, when the currents are in the opposite direction they repel one another. The reason for the force being attractive for currents in the same direction and repulsive for opposing currents relates to the magnetic field lines in the region between the two wires, as shown in Figure 6.13. When the currents are in the same direction, the field lines tend to cancel between the wires, which leads to an attractive force. If the currents are identical, the force is exactly zero on a line halfway between the wires. When the currents oppose, the field lines reinforce between the wires, enhancing the field and leading to a repulsive force.

6.5. What should happen to the length of a spring if a large current passes through it? *Hint: Think about the current in neighboring spring coils.*

- It shortens
- It lengthens
- Nothing

6.3.4 Torque on a Current Loop

Now we know how to find the force on a straight length of current-carrying wire in magnetic field. From there, it is no big trick to show that a *loop* of wire in a magnetic field experiences a *torque*. This result will be crucial to understanding how, *e.g.*, electric motors and generators function in the next chapter.

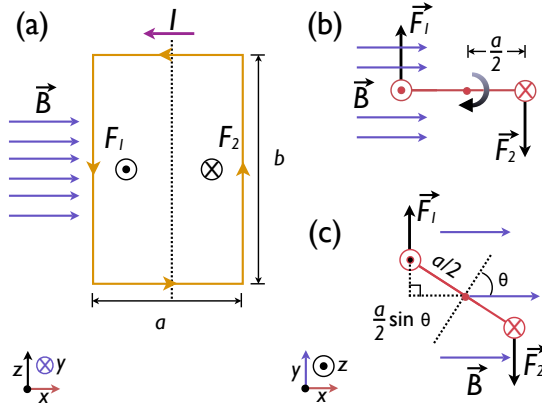


Fig. 6.14 (a) Top view of a rectangular loop carrying a current I in a magnetic field \vec{B} . No magnetic forces act on the sides of length a parallel to \vec{B} , since they are parallel to the field, but forces do act on the sides of length b , since they are at right angles to the field. (b) A side view of the loop, showing that the forces \vec{F}_1 and \vec{F}_2 on the b sides create a torque that tends to rotate the loop clockwise. (c) If \vec{B} is at an angle θ with a line perpendicular to the loop plane, the torque is $BIA \sin \theta$ where A is the area of the loop (*i.e.*, $b \cdot a$).

Take the loop of wire carrying a current I in a constant magnetic field \vec{B} in Figure 6.14a. No magnetic forces act on the sides of length a parallel to \vec{B} , since they are parallel to the field ($\sin \theta = 0$). We do expect forces to act on the sides of length b , however, since they are at right angles to the field. Further, since the sides are identical *except* for the fact that the current is in opposite directions, we expect that they experience the same magnitude of force (Equation 6.41), but in opposite directions:

$$|\vec{F}_1| = |\vec{F}_2| = BIlb \quad (6.45)$$

From right-hand rule #1, the force on the left side of the loop, \vec{F}_1 has to be out of the page, while the force on the right side of the loop, \vec{F}_2 , has to be into the page. If we fix the loop such that it pivots along a vertical axis running through the middle of the loop (the dashed line in Figure 6.14b), what will happen?

Figure 6.14b shows the loop viewed on edge. Both forces try to rotate the loop clockwise about the pivot axis. Recall that a *torque* $\vec{\tau}$ occurs when we have a force F applied some distance d from a pivot point, and $|\vec{\tau}| = Fd \sin \theta_{Fd}$, where θ_{Fd} is the angle between the force and the displacement to the pivot point. Consistent with our right-handed coordinate system, positive torque corresponds to clockwise rotation.

The forces \vec{F}_1 and \vec{F}_2 are applied at a distance $a/2$ from the loop's pivot point, and the angle between the force and displacement is 90° , so the net torque is:

$$|\vec{\tau}|_{\max} = F_1 \frac{a}{2} + F_2 \frac{a}{2} = (BIlb) \frac{a}{2} + (BIlb) \frac{a}{2} = BIlab \quad (6.46)$$

The area of the loop is $A = ab$, so we can express the torque more generally as

$$|\vec{\tau}|_{\max} = BIA \quad (6.47)$$

This simple result only holds when the field \vec{B} is parallel to the plane of the loop. We can easily repeat the for the case when the field makes an angle θ with a line perpendicular to the plane of the loop, as shown in Figure 6.14c. All we have to do is change B to $B \sin \theta$ - the torque only results from the component of \vec{B} parallel to the loop plane:

$$|\vec{\tau}| = BIA \sin \theta \quad (6.48)$$

The loop has a maximum torque BIA when the field is parallel to the plane of the loop, and is *zero when the field is perpendicular to the plane of the loop*. When placed in a magnetic field, the loop will tend to rotate to smaller values of θ , until its plane is perpendicular to the loop (or such that its area *normal* is parallel to the field), minimizing the torque it feels. What good is all this? The torque created on a current loop by a magnetic field is the basis of many electric motors!

We can further generalize our result and consider not just one loop of wire, but N loops of wire tied together – a coil. All we have to do is add together the magnitude of the N torques $|\vec{\tau}|$ from each loop, since they all act in the same direction:

Torque on a coil of N turns in a magnetic field:

$$|\vec{\tau}| = BIAN \sin \theta \quad (6.49)$$

where I is the current carried by a each loop of area A in a coil of N turns, placed in a constant magnetic field of magnitude B , and θ is the angle between a line perpendicular to the loop and B .

This simple and most general result holds for coils of arbitrary shape, not just rectangles, so long as the loop can be contained by a cartesian plane. Since the problem of current loops comes up fairly often, we often define the quantity IAN to be the magnitude of the *magnetic moment* of the coil $|\vec{\mu}|$. The magnetic moment vector $\vec{\mu}$ always points *perpendicular to the plane of the coil*, and the angle θ is now the angle between the magnetic moment and the field B . Using this definition:

$$|\vec{\tau}| = |\vec{\mu}| |\vec{B}| \sin \theta \quad (6.50)$$

As a last remark, we point out that an electron orbiting an atomic nucleus can be thought of as a current loop, which implies that atoms would experience a torque when placed in a magnetic field. In a rough manner of speaking, this is the basis for Magnetic Resonance Imaging (MRI), the actual details of which are beyond our discussion. Magnetic resonance deals with the magnetic moments of individual electrons or protons, which is actually due to their quantum-mechanical *spin*, a topic we will cover in quantum physics.

6.3.5 Magnetic Fields of Current Loops and Solenoids

The magnetic field produced by a current-carrying wire can be magnified at a point by bending the wire into a loop. Consider the loop in Figure 6.15. The small segment of the loop Δx_1 produces a magnetic field at the loop's center which is directed out of the page. The segment Δx_2 *also* produces a magnetic field directed out of the page, which adds to the field from segment Δx_1 . This occurs for every tiny segment of the whole loop, with the result that the field at the center is much larger than anywhere else.

The magnetic field at the center of a loop of radius R carrying a current I is given by

$$\vec{B}_{\text{center}} = \frac{\mu_0 I}{2R} \hat{z} \quad (6.51)$$

where \hat{z} is the direction pointing out of the page.

Deriving this result requires some calculus, so we will not reproduce it here. What we should notice is that compared to Equation 6.19 for a *straight* wire, the field is now π times larger - bending the wire into a loop enhances the field, at least at the center, by about three times. Mathematics tells us that this is as good as it gets though – no other shape will give us a bigger enhancement.

The magnetic field lines for a current loop are shown in Figure 6.16a. All of the field lines converge toward the central region of the loop, creating a much higher field there. All of the field lines enter at the bottom of the loop and exit at the top. Notice how the current loop behaves as if it has a north and a south pole, just like a bar magnet. The *magnetic moment* of a circular current loop

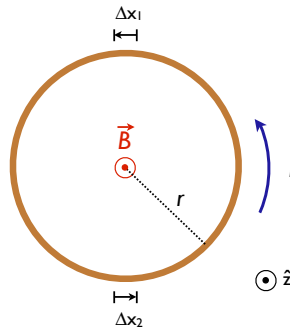


Fig. 6.15 All segments of the current loop produce a magnetic field at the center of the loop directed *out of the page*.

is the same as we found in Section 6.3.4, $|\vec{\mu}| = IA = 2\pi rI$, directed along the axis perpendicular to the loop plane and in accordance with the right-hand rule.

It is no accident that current loops look like permanent magnets - as we mentioned in Section 6.1, permanent magnetic materials can in some sense be modeled as consisting of tiny, atomic current loops. In fact, the field lines in Figure 6.16a are not just “like” a bar magnet, *a current loop creates a magnetic field nearly indistinguishable from a bar magnet*. This is our fundamental linkage between electricity and magnetism.

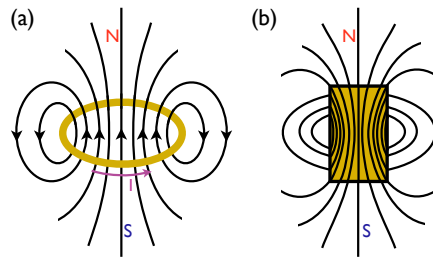


Fig. 6.16 (a) Magnetic field lines surrounding a current loop. (b) Magnetic field lines surrounding a bar magnet. Note the similarity between this line pattern and that of a current loop. Outside of the magnet itself, they are nearly indistinguishable – both are approximately magnetic dipoles.

We can make current loops into a longer “bar magnet” by adding them together. If we make a coil of N equivalent loops of wire stacked together, each carrying a current I , the field at the center is the sum of the fields from each of the N coils:

$$\vec{B}_{\text{center}} = N \left(\frac{\mu_0 I}{2R} \right) \hat{z} \quad (6.52)$$

In other words, if bending a wire into a single loop enhances the field maximally, then the next best thing is to just add more loops. The field from every loop just adds to the total, so long as the currents are all running in the same direction.

6.3.5.1 Solenoids

Instead of stacking individual loops, we can take a long straight wire and bend it into a coil. Such a coil is called a *solenoid*, a type of *electromagnet*. Solenoids are important because they act as magnets only when current is supplied (there is no remnant field, like a permanent magnet has), and create an extremely uniform field inside them. A solenoid is one form of an electromagnet – solenoids using superconducting wire are crucial for creating the large magnetic fields required for Magnetic Resonance Imaging.

Figure 6.17 shows a schematic of a solenoid and its magnetic field lines. The conductors going into and out of the page carry a current I . The field lines inside the solenoid are very nearly parallel, uniformly spaced, and close together. Subsequently, the field inside is strong - being the superposition of the field of many individual coils - and very uniform. Note how the solenoid looks just like a long bar magnet now – again, they are nearly indistinguishable.

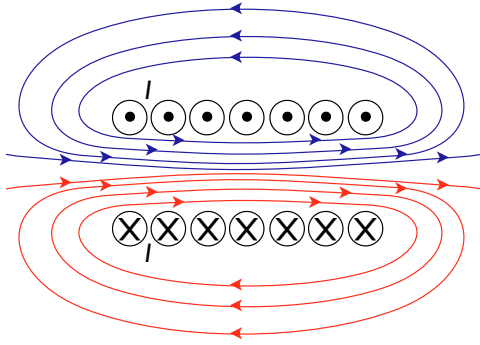


Fig. 6.17 Magnetic field lines around a solenoid. The field is nearly uniform inside if we are far from the edges, and small outside.

The field outside the solenoid is weaker, nonuniform, and in the *opposite direction*. We can make the field inside more and more uniform by adding more and more coils, making the solenoid longer. If the solenoid is long compared to its diameter, the field will be very uniform toward the middle.

What is the field inside the solenoid? We can use Ampère's law (Equation 6.23) to find out. Let us imagine that the total number of turns is N , and the length l . Take a closed loop for Ampère's law like the loop labeled "1" in Figure 6.18. We will consider the solenoid to be so long that the field outside is essentially zero. Ampère's law tells us to sum up $B_{\parallel}\Delta l$ around this loop. Since the field is constant on each side of rectangle 1 (though not the same on every side), we can just sum up $B_{\parallel}\Delta l$ for each side. The contribution from the top and bottom sides is clearly zero - the field is *perpendicular* to the length there. The contribution from the outside edge is also zero, since $B \approx 0$ there. The only non-zero contribution is from the inner side of the rectangle:

$$\sum_{\text{path 1}} B_{\parallel}\Delta l = B_z L = \mu_0 I_{\text{enclosed}} \quad (6.53)$$

The right-hand side is total current that passes through rectangle 1. If there are N loops over the length l , the current enclosed is just NI .

Field inside a long solenoid:

$$\vec{B} = \mu_0 \frac{N}{L} I \equiv \mu_0 n I \hat{z} \quad (6.54)$$

where \hat{z} is on the axis of the solenoid, N is the number of turns of wire each carrying a current I , and L is the length of the solenoid (so there are $n \equiv N/L$ turns per unit length).

For the last line, we have defined the quantity n to be the number of "turns per unit length" for convenience. Now, what if we try to apply Ampère's law to loop 2? Again, the top and bottom sides give no contribution, since the field is perpendicular. The left and right sides experience the same (parallel) field B , but on the right side the length vector is in the same direction as \vec{B} , while on the left side it is opposing. This means that one side gives a positive contribution, and the other an equivalent negative contribution. So $\sum B_{\parallel}\Delta l = 0$, which must be true since $I_{\text{enclosed}} = 0$ for this path!

Path 3 is even easier - if the solenoid is long enough to neglect the field outside, then the contribution from every side is zero. Again we have $\sum B_{\parallel}\Delta l = 0$, and $I_{\text{enclosed}} = 0$, consistent with Ampère's law.

6.4 Permanent Magnetic Materials

We know from everyday experience that permanent magnetic materials, when magnetized, are sources of strong magnetic fields. Why? We can, in a very rough sense, imagine electrons as in a circular orbit around an atomic nucleus. Electrons moving in this circular orbit constitute a current of sorts, and with that current loop is an associated magnetic moment.

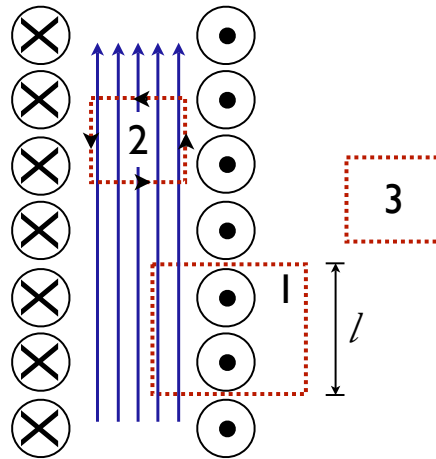


Fig. 6.18 Ampère's law paths for a solenoid.

This effect, as it turns out, is rather small. The magnetic properties of many materials can be explained by the fact that electrons not only behave as if they are orbiting the atomic nucleus, but they also behave as if they are spinning like a top. (This analogy should not be taken literally, the “true” explanation results from quantum mechanical phenomena.) This spinning motion also represents moving charge, and with it is also associated a magnetic moment.

Electrons tend to group in pairs such that their “spin” magnetic moments cancel – you might remember this as Hund’s rule from chemistry. As a result, materials with an *even* number of electrons tend not to be strongly magnetic. If there are an *odd* number of electrons, a net magnetic moment results. Each one of the N unpaired electrons in a magnetic material possesses a magnetic moment $\vec{\mu}$. If the material is a *permanent magnet*, all of the individual moments tend to line up in the same direction spontaneously, and they add together to form a very large field. If the material is magnetic, but not a permanent magnet (proceeding section), the moments do not spontaneously align, but can be forced into alignment with a small external magnetic field.

If we define the number of unpaired electrons per unit volume is $n \equiv N/V$, then the quantity $n\vec{\mu} \equiv \vec{M}$ is called the *magnetization* of the material, or the magnetic moment per unit volume. The quantity μ_r is the *relative permeability* of the material, just like μ_0 is the permeability of vacuum.

The net result of this is that *magnetic materials behave as if there is a large magnetic field present inside them*, in addition to the external field. This internal magnetic field has a maximum value when the material is fully magnetized, known as the “saturation magnetization” of the material. The saturation magnetization can be the equivalent of hundreds of teslas in common magnetic materials!

Magnetic field inside a permanent magnet, for small external fields:

$$\vec{B}_{\text{inside}} = \mu_r \vec{B}_{\text{external}} \quad (6.55)$$

What this equation tells us is that the field inside a magnetic material can be as much as μ_r times the external field. In other words, *the permeability of a material amplifies the applied magnetic field*. This is only true for relatively low fields – once the material is completely magnetized, the field inside reaches a constant value known as the “saturation magnetization.” The relative permeability can be as high as $10^5 - 10^6$, so the fields inside magnetic materials are truly colossal.

The behavior of the total magnetic field for a magnetic material, internal plus external, is shown in Figure 6.19. Important to note is that the total field is not zero when the applied field is zero –

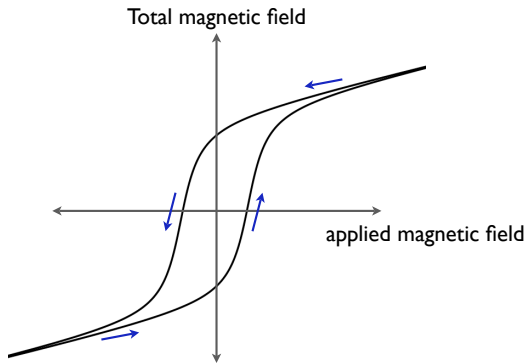


Fig. 6.19 Total magnetic field versus the applied magnetic field for a permanent magnet.

permanent magnets have a *remnant magnetic field*, which is why they are permanent magnets in the first place! Another key point is that the direction of the remnant magnetic field (positive or negative) depends on the history of the applied field – a phenomena known as hysteresis, which is the basis for magnetic information storage in hard disks.

We have necessarily left out a great many fundamental details about permanent magnetic materials. A good introductory place to learn more is: <http://hyperphysics.phy-astr.gsu.edu/hbase/solids/magperm.html>.

6.4.1 Non-permanent magnetic materials

There are a great many materials which are not permanent magnets, but can *become* magnetized by an external magnetic field. In these materials, what often occurs is that the electron spin moments do not all line up together, but are in random directions. An external field can align them, however, which will magnetize the material. In all respects it will behave like a permanent magnet, except that it has *no hysteresis* - once the external field is removed, the material is no longer magnetized.

The strength of the induced magnetic alignment in a non-permanent magnetic material is nothing more than its relative permeability μ_r . The internal magnetic field is enhanced by a factor μ_r , like in a permanent magnetic material, but non-permanent magnets retain no magnetic behavior once the external field is removed.

6.4.2 Electromagnets

Now we can understand a bit how strong electromagnets work. Figure 6.20 shows a cross-section of an electromagnet. The permanent magnetic material (Iron, for example) is in the shape of an “O” with one small notch cut out of it. Wrapped around the closed end opposite the notch is a coil of copper wire of length L (running into and out of the page) with N turns each carrying a current I .

What happens in this construction? The current in the “solenoid” coil creates a magnetic field of $\mu_0 NI/L$ in the left-to-right direction. This relatively small magnetic field serves to magnetize the iron core, such that the field inside the core is μ_r times the field from the copper coil: $B_{\text{inside}} \approx \mu_r \mu_0 NI/L$. So what is the field inside the gap? One can use Ampère’s law for that, or the boundary conditions on the magnetic field (Appendix ??) but we will only quote the result for the field inside the gap here:

$$B_{\text{gap}} \approx \mu_r \mu_0 \frac{N}{L} I \quad (6.56)$$

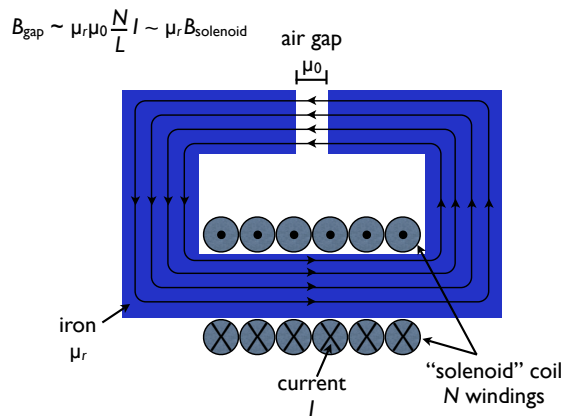


Fig. 6.20 An electromagnet with a permanent magnet core. A current in the “solenoid” coil wrapped around the iron core creates a small magnetic field. This small magnetic field magnetizes the core, creating a much larger field in the gap region. The field in the gap is larger than that of the solenoid alone by roughly a factor μ_r .

So long as the gap is very narrow compared to the size of the core itself, the field is just about the same as that inside the magnetic material. Just like inside the core itself, the field in the gap is *enhanced* by a factor μ_r , so good electromagnet cores are made from materials with very high μ_r . Table 6.1 lists the relative permeability for a few permanent magnetic materials.[21] Given that μ_r can be thousands or hundreds of thousands, the reason for having a core in an electromagnet is clear - it is a magnetic field amplifier!

Table 6.1 Relative Permeabilities and Remnant Fields of Some Magnetic Materials[21]

Material	μ_r (representative)	μ_r (maximum)	Remnant field [T]
iron	200	200,000	1.3
nickel	100	600	0.4
cobalt	70	250	0.5
permalloy ($\text{Ni}_{78.5}\text{Fe}_{21.5}$)	8,000	100,000	0.7
mumetal ($\text{Ni}_{75}\text{Cr}_2\text{Cu}_5\text{Fe}_{18}$)	20,000	$\sim 1,000,000$	0.7
316 stainless steel	~ 1	~ 1	~ 0

6.4.3 Permeability and Magnets on Your Fridge

A material that is strongly attracted to a magnet is also said to have a high *permeability* μ_r . Why is that? Why do magnets stick to a refrigerator door?⁸

A permanent magnet sticks to the side of the refrigerator because part of the refrigerator is able to *become magnetized* by a magnetic field, even though it is not a permanent magnet (*i.e.*, it has no hysteresis). The internal field in the magnetized region is proportional to the relative permeability μ_r , as noted above. The magnetized region acts just like another magnet, and the field lines from this induced magnetic alignment join with those of the inducing magnet, forming continuous field lines that link the two together, as in the case of the two permanent magnets. The opposite alignment of magnetic poles gives an attractive force, which tends to bring the magnet closer to the fridge, which increases the force ... until they are stuck together.

This is a little bit like charging by induction with electric fields - the magnetic field from a permanent magnet pole (say, N) induces a magnetic pole in the refrigerator of the opposite sign (S), just like a positively charged rod induces a negative charge on a conductor (Sect. 2.2.2). The

⁸ So long as it is not stainless steel. Austenitic stainless steels, like 310 and 316 (and 304 to a lesser extent) have extremely low permeabilities, and show almost no response to an external magnetic field.

difference is, again, that magnetic poles only come in pairs, so there is no magnetic version of charging by *conduction* (Sect. 2.2.1).

6.6. Permanent magnets sticking to a refrigerator door happens because the permanent magnet is able to induce magnetic poles in the steel of the door. This process is analogous to electrically charging objects by *induction*, which we discussed in Ch. 2, where a charged object induces opposing charges in a conductor without contact.

Can a process like *conduction* charging, where an object transfers some of its charges to another, happen with magnets?

6.5 Problems

Solutions begin on page 274.

Chapter 7

Induced Voltages and Inductance

It is a capital mistake to theorize before one has data. – Sir Arthur Conan Doyle

Abstract The net flow of charges - or current - leads to a magnetic field as we learned in Chapter 6. Magnetic fields can in turn induce currents, as it turns out, through *induction*. Historically, it was first discovered by Oersted that magnetism was produced by current-carrying wires, as we found in Chapter 6. It was not immediately clear, however, that electricity could in turn be produced by magnetism. While steady currents produced constant magnetic fields, experiments showed that steady magnetic fields could not produce currents. It was not until the experiments of Faraday¹ and Henry² that it was discovered that only *time varying* magnetic fields could produce currents.

7.1 Induced Voltages and Magnetic Flux

Both Faraday and Henry discovered that currents could be produced in a coil of wire by simply moving a magnet in and out of the coil, as shown in Figure 7.1. Simply placing the magnet in the coil of wire did nothing. Only when the magnet was moving relative to the coil was a voltage induced in the coil and a current created. Whether the magnet moves and the coil is stationary, Fig. 7.1a, or the coil moves and the magnet is fixed, Fig. 7.1b, is not important, only relative motion matters.

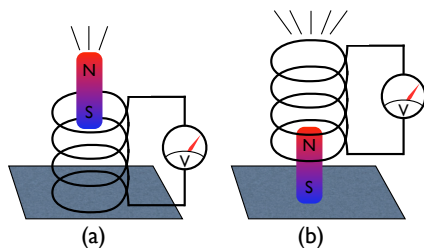


Fig. 7.1 (a) When a magnet is pushed through a coil of wire, a voltage is induced in the coil. (b) When the coil moves around the magnet, a voltage is also induced. Whether the loops of wire move or the magnet moves is immaterial, a voltage is induced so long as the magnetic flux through the coil changes in time.

Strictly speaking, it is not a current that is induced in the coil, but a voltage difference between its end points. If the coil is part of a closed electric circuit, a current will flow, but a potential difference will be induced even in a disconnected coil. If there is a voltage present, and the wire is conducting, this means that an electrical current will be induced in any closed circuit when the magnetic flux through a surface changes. Electromagnetic induction underlies the operation of generators, induction motors, transformers, and most other electrical machines.

¹ Michael Faraday (1791–1867), an English physicist and chemist who contributed significantly to the field of electromagnetism.

² Joseph Henry (1797–1878), a Scottish-American scientist who pioneered electromagnetic induction, along with Michael Faraday.

The induced voltage is produced whenever there is relative motion of the coil and magnet, and it was also discovered that the more loops of wire there are in the coil, the larger the induced voltage, Fig. 7.2. Twice as many loops gives twice as much voltage, everything else remaining the same.

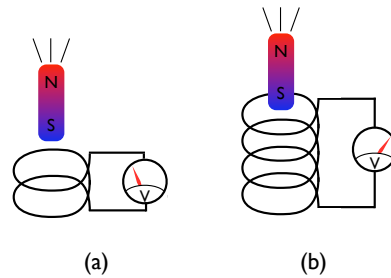


Fig. 7.2 (a) A magnet pushed through a coil induces a voltage. (b) When the magnet is pushed through a coil with twice as many loops, the induced voltage is twice as much.

Finally, it was discovered that the induced voltage depends on how fast the magnetic field through the coil changes - which in this simple example just means how fast the magnet moves relative to the coil.

Induced Voltages:

The induced voltage in a coil is proportional to the number of loops, and the rate at which the magnetic field through the loop changes.

7.2 Faraday's Law of Induction

More precisely, Faraday found that the *induced voltage produced between the ends of a loop of wire is proportional rate of change of the magnetic flux passing through the surface of the loop*. Magnetic flux, Φ_B , is defined similarly to electric flux (Fig. 7.3):

Definition of magnetic flux through a loop

$$\Phi_B = B_{\perp} A = BA \cos \theta_{BA} \quad (7.1)$$

where θ_{BA} is the angle between the surface normal and the magnetic field. The flux through a closed surface bounding a volume is still zero.

Magnetic flux is the product of area of the loop and the perpendicular component of the magnetic field through it, as shown in Fig. 7.3, and the induced voltage in the coil depends on the rate that the flux changes with time. What this means is that *either* the field can be changing in time, *or* the area facing the magnetic field can be changing in time, or both, and a voltage will be induced. For example, we could either move the magnet back and forth into the loop, or rotate the coil in a constant magnetic field.

The result of all of this is Faraday's law of electromagnetic induction, which relates the change in magnetic flux through a loop per unit time to the induced voltage in the loop:

Faraday's law of electromagnetic induction If a circuit contains N closely wound loops and the magnetic flux changes by $\Delta\Phi_B$ in a time Δt , the *average* voltage induced in the loop is

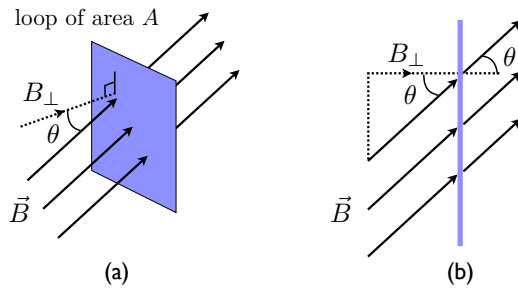


Fig. 7.3 Magnetic flux through an area A . **(a)** A uniform magnetic field \vec{B} is incident on an area A at an angle θ with the direction *normal* to the area. **(b)** Edge-on view of the area. The flux through the loop is $\Phi_B = B_{\perp}A = BA \cos \theta$

given by:

$$\Delta V = -N \frac{\Delta \Phi_B}{\Delta t} \quad (7.2)$$

This law covers all the basic phenomena we just discussed - the induced voltage depends on the number of turns in the coil, and how fast the magnetic flux through the coil changes. What about the minus sign though? What the minus sign says is that *the induced voltage will try to create a current that opposes the change in magnetic flux*. If a current is induced in the coil, it will circulate in such a way to try and stop the change in flux, by creating a magnetic field of its own.

For a minute, think about what would happen if the minus sign weren't there. In this case, a time-varying flux would create a current in a loop of wire, which would create a field that changes in the *same way as the field causing the flux*. This field would then *add* to the field causing the flux, which would increase the current even more, and then further add to the original field. This positive feedback would quickly run amok! Any infinitesimally small change in magnetic field with time would get amplified, and cause a runaway current in the coil (at least until it melted). Since this situation is clearly absurd, it makes some sense that the induced current must *oppose* the change in flux, rather than add to it. It is precisely this negative feedback of coils which makes them useful circuit elements, which we will come to in following sections.

Incidentally, this does not mean that the magnetic field created by the induced current is always opposite that of the field causing the flux in the first place - it is trying to stop the *change* in flux, not cancel the flux completely. For example, if the magnetic field causing the flux is *increasing*, the induced current will create a field in the opposite direction to oppose the increasing flux, but if the flux is *decreasing*, the induced current will create a field in the *same* direction to “shore up” the flux and stop it from decreasing. This principle is known as *Lenz's law*, and we will return to its implications in later sections.

7.3 Inductance

7.3.1 Mutual Inductance

As a concrete example, consider the two solenoids in Fig. 7.4. The top solenoid is powered with a time-varying current $I(t) = I_0 \cos \omega t$, which produces a time-varying magnetic field $|\vec{B}(t)| = B_0 \cos \omega t$. This time-varying magnetic field creates a time-varying *flux* in the lower “pickup” solenoid, which in turn leads to an induced voltage. The current, magnetic field, and induced voltage all vary sinusoidally, though not all with the same phase as we shall see.

What is the phase relationship between the current in the source coil and the voltage in the “pickup” coil? First, we know that the magnetic field created by the source coil is just proportional to the current in the coil, so it will be in phase with the current. When the current is at a maximum, so is the magnetic field.

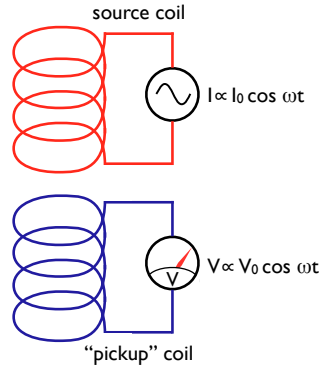


Fig. 7.4 Mutual induction of two solenoids. The top solenoid is powered with a time-varying current $I(t) = I_0 \cos \omega t$, which produces a time-varying magnetic field $|\vec{B}(t)| = B_0 \cos \omega t$. This time-varying magnetic field creates a time-varying flux in the lower solenoid, which in turn leads to an induced voltage.

This in turn means that when the current and magnetic field are maximal, then the flux in the pickup coil is maximal - only the magnet field is changing, the area is constant in this case. The induced voltage in the pickup coil, however, depends on the *time rate of change* of the flux, not the flux itself. When is the rate of change maximal? The time rate of change $\Delta\Phi_B/\Delta t$ is nothing more than the *slope* of the flux versus time curve. Since the current and magnetic field are sinusoidal in time, $\cos \omega t$, so too is the flux in the pickup coil. The maximum slope for a sinusoidal curve is where it crosses zero on the y axis, and it has zero slope at peaks and troughs.

What this means is that $\Delta\Phi_B/\Delta t$ for the pickup coil is maximum whenever the field from the source coil, and therefore the current in the source coil, is zero. Whenever the current in the source coil is maximal, the induced voltage is zero, since $\Delta\Phi_B/\Delta t$ is zero. In short, the induced voltage in the pickup coil is still sinusoidal, but a quarter cycle (90°) out of phase with the current in the source coil.

What this setup essentially does is wirelessly transmit power from the source to the pickup coil through the time-varying magnetic field. This is known as *mutual inductance*, the basis for electrical transformers. The key thing in designing a transformer is to somehow focus as much of the flux from the source coil as possible and guide it into the pickup coil, such that as much electrical energy from the source coil is transferred into the pickup coil as possible. One relatively easy way to do this is to use a high permeability magnetic material to guide the flux, as in an electromagnet (Section 6.4.2).

7.3.2 Self Inductance

What happens if we have only one coil? A single coil powered by a time-varying voltage creates a time-varying magnetic field around it. Can a coil be affected by its *own magnetic field*? Yes. The coil doesn't know where the field comes from, or how it was created, all it sees is a time-varying field, which it will try to oppose. When we power a single coil, some of the flux lines emanating from the coil will pass through the coil itself, and as the current and field change in time, Faraday's law tells us that there must be an induced voltage, just as if the time-varying field were caused by a second coil. This is known as *self inductance*.

We know that the induced voltage ΔV must be given by Faraday's law:

$$\Delta V = -N \frac{\Delta\Phi_B}{\Delta t} \quad (7.3)$$

We also know that the change in flux is just due to the current in the coil itself, so:

$$\frac{\Delta\Phi_B}{\Delta t} \propto \frac{\Delta I}{\Delta t} \quad (7.4)$$

Combining these two facts, we see that the induced voltage in the coil ΔV must be proportional to the change in current with time:

Self-inductance:

$$\Delta V \propto \frac{\Delta I}{\Delta t} \quad \text{or} \quad \Delta V = -L \frac{\Delta I}{\Delta t} \quad (7.5)$$

where the constant of proportionality L is called the **inductance** of the coil.

We can also use our two proportionality equations, Eq. 7.3 and 7.4, to find an expression for L itself:

$$L = N \frac{\Delta \Phi_B}{\Delta I} = \frac{N \Phi_B}{I} \quad (7.6)$$


The unit of inductance L are volt-seconds per ampere [V·s/A], or **henries** [H].

So the inductance depends on the number of turns in the coil, the current in the coil, and the flux of course. The fact that inductance depends on flux means that it is a function of the coil geometry, and in general is difficult to calculate. For a simple solenoid, we know all of these quantities, however, and one can show $L = \mu_0 N^2 A / l = \mu_0 n^2 V$, where A , l , and V are the cross-sectional area, length, and volume of the coil, respectively, and $n = N/l$ as usual.

The inductance of a coil tells us how dramatically a coil responds to changes in its own current. From Lenz's law, we know that the induced voltage in the coil will try to stop any changes its flux, which means opposing changes in current in the coil itself. Inductance is therefore a sort of a "resistance to change in current," which makes inductors are useful in circuits with time-varying signals. This is nothing more than the "negative feedback" implied by Lenz's law we referred to above, and the negative feedback of inductors can be used in, *e.g.*, audio amplifiers and many other circuits to "smooth out" rapid changes or fluctuations in signals, as we will explore further below.

7.3.2.1 Inductors as Circuit Elements

The reluctance of a coil of wire to change current rapidly due to its self inductance can actually be a useful thing in an electronic circuit. Undesired rapid changes in current (due to a power spike, for example) can be smoothed out by putting an inductive element in a circuit, and some GFI outlets are based on this idea. Filters for high-frequency filters (such as those in audio amplifiers) can be built from inductors due to their reluctance to allow rapidly-varying currents through them – a rapidly-varying current is just another description of a high-frequency signal. Combined with capacitors, which like to let rapidly-varying signals through but not slowly-varying (or dc) signals, one can tailor the frequency or time response of all sorts of circuits. A circuit element used primarily for its self-inductance is simply called an *inductor*

Circuit diagram symbol for an inductor L : 

Owing to the fact that inductors also store energy in their magnetic field, which we will discuss in the following section, inductors can also be used to temporarily store energy, just like capacitors. In fact, capacitors and inductors are closely linked conceptually:

Inductors and Capacitors:

An inductor's behavior toward current is the same as a capacitor's behavior toward voltage.

This rule of thumb will become more clear when we discuss the inductive version of the RC circuit, the RL circuit.

7.3.2.2 RL Circuits

Before we discuss circuits with inductors in them, we should first think about what role inductance might play in some of the simple circuits we have constructed so far. Consider the simple resistive circuit in Fig. 7.5a. This is a circuit we have seen many times before. We know that once the switch S is closed, there will be a current of $I = \Delta V / R$ in the resistor, so the voltmeter across the resistor will read ΔV (provided the wires have negligible resistance). Now, what happens *just after we close the switch S* ? Is there immediately a current $I = \Delta V / R$ in the resistor, or does it take a little while to build up? In either case, why?

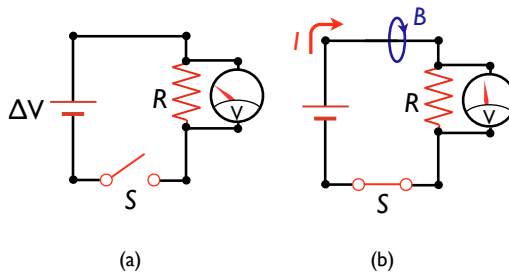


Fig. 7.5 Instantaneous current in a resistive circuit. At the instant the switch S is thrown, what is the current in the resistor?

First, remember that *any* working electric circuit has to create a closed loop – current has to go from the source, around a circuit, and back into the source. Next, remember from the preceeding section that any closed loop of wire has a **self inductance** L , which tends to resist rapid changes in current. Putting this together, *the closed loop of the circuit itself acts as an inductor, and tries to resist changes in current*. Since we have to have a closed loop to make any kind of circuit, this means that *all* of our other circuits already behave as if they have inductors present!

As soon as the switch S is thrown, I doesn't immediately change from 0 to $\Delta V / R$. The current beginning to flow in the circuit creates a magnetic field \vec{B} circulating around the wires in the circuit, which in turn increases the flux inside the loop, Fig. 7.5b. Eventually, a steady-state is reached, and the current is constant. The constant current will lead to a voltage drop across the resistor of $\Delta V_R = IR$. The voltage drop across the resistor represents an opposition to the current – the larger the current, the larger the voltage drop across the resistor.

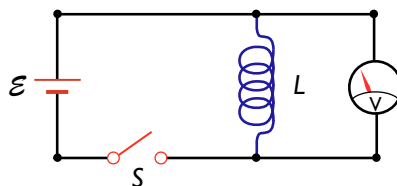


Fig. 7.6 Instantaneous current in an inductive circuit. At the instant the switch S is thrown, what is the current in the inductor?

Now, for our first real inductive circuit, an inductor L connected in series with a voltage source ΔV , Fig. 7.6. What happens when we close the switch S ? At the instant the switch is thrown, the current tries to flow into the inductor. We know from our discussion of self inductance that when the current is changing in time, a voltage is induced in our inductor. Using the loop rule, which says that

the sum of voltage drops and sources around a closed loop must sum to zero, we can readily find the induced voltage in the inductor, ΔV_L :

$$\Delta V_L = -L \frac{\Delta I}{\Delta t} \quad (7.7)$$

This looks a lot like the voltage drop across a resistor, and by analogy **we interpret the inductance L as an opposition to the change in current**. The faster the current changes, the larger $\Delta I/\Delta t$, and the larger the voltage built up in the inductor – the faster you try to change the current in an inductor, the more readily it “soaks up” the available voltage to confound your efforts. For this reason, inductors can be useful for *preventing* rapid surges in current. Connecting a point in a circuit to ground *via* an inductor effectively shunts away rapid current variations to protect sensitive equipment.

We are finally ready for a more useful circuit, the RL series circuit shown in Fig. 7.7. Suppose the switch is closed at $t=0$. As soon as this happens, the current begins to increase, but the inductor tries to prevent it from increasing too quickly – the maximum voltage drop across the inductor occurs when the current is changing most rapidly, right when the switch is closed. By “stealing” as much of the voltage from the source as possible, the inductor prevents the resistor from taking part of the voltage drop and thereby inhibits current from flowing.

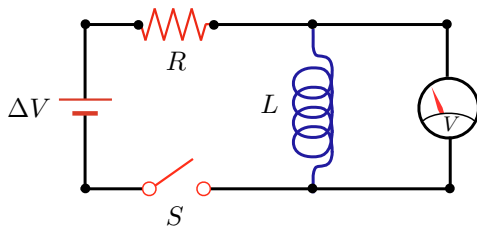


Fig. 7.7 An “ RL ” circuit. The inductor prevents the current from changing too quickly in the circuit – the current through an inductor behaves like the voltage across a capacitor.

As the current approaches its steady-state value, the *changes* in current become less and less, and the inductor has a smaller and smaller voltage drop. When the current is finally stabilized at a constant value, an ideal inductor actually has *no* voltage drop, since the current isn’t changing at all, $\Delta I/\Delta t = 0$. This is reminiscent of our RC circuits of Sect. 5.6. Only while the capacitor was charging or discharging did a current flow, not in the steady state. Inductors behave toward voltage as capacitors behave toward current – a voltage only develops across an ideal inductor when the current is changing just like a current only flows in a capacitor when the voltage is changing.

In the case of RC circuits, we found that the time it took to charge or discharge the circuit depended on a *time constant* $\tau = RC$. In the case of an RL circuit, we can define a similar time constant which gives the time required for the voltage to get within $1/e^3$ of its steady-state value:

Time constant τ of an RL circuit:

$$\tau = \frac{L}{R} \quad (7.8)$$

This gives τ in seconds [s] when R is in Ohms [Ω] and L is in Henries [H].

The equation for the current as a function of time for an RL circuit is also just like the *voltage* as a function of time for an RC circuit:

$$I(t) = \frac{\Delta V}{R} \left(1 - e^{-t/\tau}\right) \quad (7.9)$$

Just like a capacitor takes time to charge up, an inductor takes time to let a current flow. The larger the inductance L , the longer it takes for the current to reach its steady state value, just like varying C in an RC circuit. In contrast to the RC case, however, increasing R *decreases* the waiting time. In the

³ Here again we mean e the base of the natural logarithms, not e the unit of charge.

RL circuit, a larger resistance is able to “steal” more of the voltage from the inductor, lessening its ability to impede current flow. In the RC case, increasing the resistance also “steals” more voltage from the source, which leaves a smaller voltage available to charge the capacitor – hence it takes longer.

We can even take the analogy between inductors and capacitors one step further. Capacitors store electrical energy by separating charges. The induced voltage across an inductor prevents the voltage source from immediately producing a current, which means that the source must do work to achieve current flow. If the source must do work against the inductor, then there must be some source of stored energy inside the inductor. As it turns out, the presence of a magnetic field in the inductor is the source of energy, just like the presence of the electric field between the plates of a capacitor is a source of energy. Following the same derivation as Sect. 3.6.2, we can relate the potential energy stored in an inductor to the current in the inductor:

Energy stored in an inductor:

$$PE = \frac{1}{2}LI^2 \quad (7.10)$$

Again, notice that if you replace L with C and I with V , you have exactly the expression for potential energy stored in a capacitor. Now we can make our glib rule of thumb even more succinct:

Inductors and Capacitors:

Current in inductors is just like voltage on capacitors.

7.4 Transformers

The dual coil setup Section 7.3.1 is the most basic form of a transformer. If our source solenoid has N_1 turns, and is powered by a voltage ΔV_1 , then the magnetic field created by it is proportional to $\Delta V_1 N_1$, since I and ΔV are proportional by Ohm’s law, and B , I , and N_1 are proportional. Faraday’s law tells us that the induced voltage in the pickup solenoid ΔV_2 is proportional to the rate of change of that field, as well as the number of turns in the pickup coil N_2 :

$$\Delta V_2 = -N_2 \frac{\Delta \Phi_{B1}}{\Delta t} \quad (7.11)$$

On the other hand, we now know that we can relate the change in Φ_{B1} to the voltage in coil 1 through its self inductance (Eq. 7.6):

$$\Delta V_1 = -N_1 \frac{\Delta \Phi_{B1}}{\Delta t} \quad (7.12)$$

Combining these two equations, we can eliminate $\Delta \Phi_{B1}/\Delta t$ completely:

Voltage relationship between source (1) and pickup (2) coils in a transformer:

$$\Delta V_2 = \frac{N_2}{N_1} \Delta V_1 \quad (7.13)$$

here $N_{1(2)}$ is the number of turns in the source (pickup) coil, and $V_{1(2)}$ is the voltage on the source (pickup) coil.

What this tells us is that when N_2 is greater than N_1 the pickup coil voltage is actually *larger* than that of the source coil, and we call this configuration a “step-up” transformer. Step-up transformers

take a given time-varying voltage, and amplify it by a factor N_2/N_1 . When N_2 is smaller than N_1 we have a “step-down” transformer, which takes a given time-varying voltage and reduces it by a factor N_2/N_1 .

Of course, there is no free lunch, and we can’t get power from nowhere. The total power input to the source coil has to equal the total power at the pickup coil, or

$$I_1 \Delta V_1 = I_2 \Delta V_2 \quad (7.14)$$

This also gives us the relationship between the currents in the source and pickup coil:

Current relationship between source (1) and pickup (2) coils in a transformer:

$$I_2 = \frac{N_1}{N_2} I_1 \quad (7.15)$$

here $N_{1(2)}$ is the number of turns in the source (pickup) coil, and $I_{1(2)}$ is the current in the source (pickup) coil.

This tells us that if we step up the voltage, we have to step down the current, and vice versa, in order to conserve energy.

7.5 Voltage Induced by the Motion of a Conductor in a Field

For now, back to moving charges. What happens when we take a conducting bar of length l , and move it in a magnetic field, as shown in Fig. 7.8? Our straight conductor is moving to the right at a velocity \vec{v} perpendicular to a constant magnetic field \vec{B} directed out of the page. Every electron in the conductor is moving at a velocity \vec{v} , and therefore experiences a magnetic force $|\vec{F}_m| = q|\vec{v}||\vec{B}|$ directed downward. As a result, electrons tend to “pile up” at the bottom of the conductor, leaving a net charge imbalance in the bar.

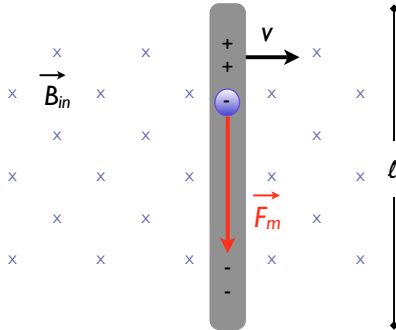


Fig. 7.8 A conducting bar of length l moving with velocity \vec{v} through a uniform magnetic field \vec{B} perpendicular to both \vec{v} and the axis of the conductor. A magnetic force \vec{F}_m acts on electrons in the conductor, giving rise to a voltage of $\Delta V = |\vec{B}|l|\vec{v}|$

This charge imbalance has to give rise to a uniform electric field inside the conductor, \vec{E} , directed downward. Of course, the presence of this charge imbalance and electric field also means the electrons experience an electric force $q\vec{E}$ upward, opposite the magnetic force! The charge imbalance will continue to grow until this electric force balances the magnetic force:

$$\Sigma F = q|\vec{E}| - q|\vec{v}||\vec{B}| = 0 \quad (7.16)$$

When the forces balance, we have equilibrium, and $|\vec{E}| = q|\vec{v}||\vec{B}|$. A uniform electric field \vec{E} over the length l of the bar is nothing more than a potential difference, $\Delta V = El$. Putting this all together,

the movement of the conducting bar in a magnetic field leads to a potential difference across the length of the bar:

Motional Voltage on a Conducting Bar:

$$\Delta V = |\vec{v}||\vec{B}|l = |\vec{E}|l \quad (7.17)$$

where l is the length of the bar, and \vec{v} and \vec{B} are at right angles.

By itself, this is not so useful, but we can make the moving bar part of an electric circuit, as shown in Fig. 7.9a. The bar now slides on conducting rails, and the motional voltage produced in the bar induces a current in the rails. An equivalent circuit is shown in Fig. 7.9b.

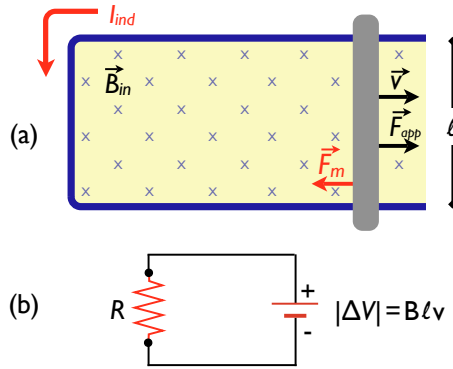


Fig. 7.9 Motional Voltage. **(a)** A conducting bar sliding across two rails creates an induced current. The magnetic force opposes the motion to try and reduce the change in flux through the loop. **(b)** The equivalent circuit corresponding to (a). The resistor R represents the rails and the conducting bar, ΔV represents the induced voltage.

In which direction is the induced current, and how big is it? The flux through the closed loop defined by the rails and the moving bar is just $\Phi_B = |\vec{B}_{in}|A_{\text{loop}}$, where A_{loop} is the area of the loop. The area of the loop is changing with time, of course, since the bar is moving, but the magnetic field is not. We can easily write down the magnitude of the induced voltage, ΔV , which along with Ohm's law will give the current $I = \Delta V/R$:

$$\Delta V = -\frac{\Delta\Phi_B}{\Delta t} = -B\frac{\Delta A}{\Delta t} \quad (7.18)$$

The movement of the bar at a constant velocity \vec{v} implies that it covers a distance Δx in a time Δt . The area of the loop is just $l\Delta x$ at any particular time, so the rate of change of the area can be found easily:

$$\Delta V = -B\frac{\Delta A}{\Delta t} = -B\frac{l\Delta x}{\Delta t} = -Blv \quad (7.19)$$

As we might have expected, the induced voltage ΔV , and the current $I = \Delta V/R$ depend on how fast the bar moves, how big the field is, and how long the bar is. Further, we see how a constant magnetic field can still give rise to a time-varying magnetic flux - if the field is constant, we have to change the area for there to be an induced voltage. This sort of voltage is called a "motional voltage," or "motional EMF" since it results from a conductor moving in a magnetic field.⁴

But. What about the direction of the current? When the bar moves to the right, due to some external force \vec{F}_{appl} , the flux is *increasing* with time. The induced current wants to *oppose the change in flux*, which in this case means *it wants to slow the motion of the bar*. This is consistent with the magnetic force on the bar being to the left, opposing the external force. The induced current wants

⁴ EMF stands for "electromotive force," a somewhat antiquated term for a source of voltage, originating from earlier times when physicists did not make a hard distinction between force and energy. We have avoided this term wherever possible to avoid confusion, as substituting "voltage" changes no essential physics.

to stop the increase in flux, so it will circulate in a direction that *opposes* the constant field in the loop, *i.e.*, counterclockwise.

What if we reverse the direction of the bar's velocity, as shown in Fig. 7.10b? If the bar were moving to the left instead, the flux would be decreasing. The induced current would circulate in such a way to stop this decrease - it would try to *increase* the flux in the loop, and would therefore circulate clockwise. Induction always acts in such a way to reduce $\Delta\Phi/\Delta t$, whether this means increasing Φ or decreasing it.

The most important thing in either case is that the *magnetic force and induced current are opposing the motion of the bar*, which is causing the change in flux in the first place. And, we have a nice symmetry between electricity and magnetism now.

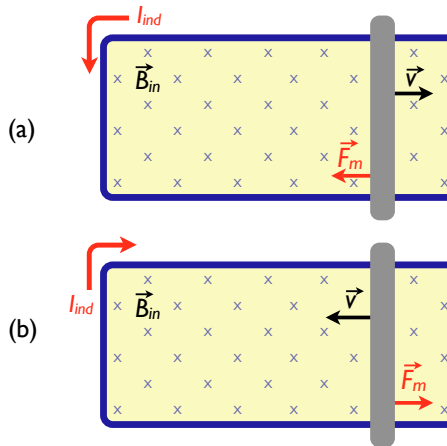


Fig. 7.10 (a) As the conducting bar slides to the right, the flux through the loop increases with time. Lenz's law states that the induced current must be counterclockwise, so that it produces a counteracting magnetic flux out of the page. (b) When the bar moves to the left, the flux *decreases* with time, so the induced current will be *clockwise*. Induction always acts in such a way to reduce $\Delta\Phi/\Delta t$, whether this means increasing Φ or decreasing it.

7.5.1 Eddy current brakes

In the end, there is nothing unique about our conducting bar from the previous section. Any time we have a moving conductor intersecting a magnetic field, or vice-versa, there is an induced current and a retarding force. The relative motion of a conductor and a magnetic field causes a circulating current within conductor. These induced currents are also known as “eddy currents,” since they are somewhat analogous to the swirling currents created when you move an oar through the water, for instance.

As we known, these induced eddies of current create magnetic fields that oppose the change in flux through their diameter. As a concrete example, consider the pendulum in Fig. 7.11a, which consists of a conducting plate swinging through a region of constant magnetic field, perpendicular to the plane of the pendulum's motion. If we pull the plate back to an angle θ and release it, we once again have a moving conductor in a magnetic field, Fig. 7.11b, and we expect again a retarding force.

As the conducting pendulum moves through the magnetic field, circulating currents form, which generate a magnetic field opposing the change in flux through the conducting plate. The only way to change the flux through the plate in this case is to slow the pendulum down, so induction results in a strong braking force on the pendulum. In contrast, a non-conducting pendulum will experience no additional force.

In short, a conducting pendulum in a (perpendicular) magnetic field will be dramatically slowed and stopped, hence the name “eddy current brake.” If the field is sufficiently strong, the pendulum will not even complete one cycle, which (hopefully) you have seen demonstrated in class. The stronger the magnetic field the pendulum moves through, or greater the electrical conductivity of the conductor, the greater the currents developed and the greater the opposing force.

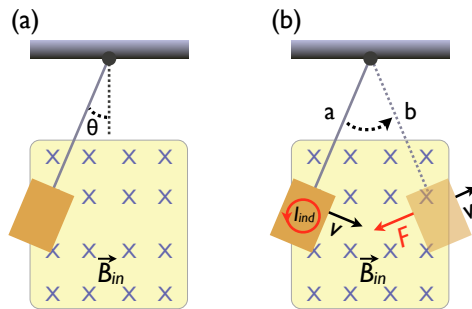


Fig. 7.11 Eddy current braking of a pendulum. **(a)** A conducting plate is released from an angle θ in the presence of a magnetic field perpendicular to the plane of motion. **(b)** The motion of the conducting pendulum through the magnetic field creates circulating currents which try to oppose the change in flux through the plate. These currents manifest themselves in a braking force on the pendulum (i.e., $v > v'$).

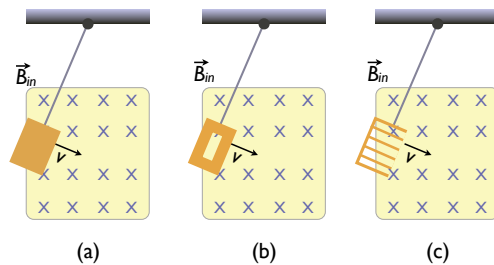


Fig. 7.12 Which pendulum experiences the largest (magnetic) force? (*Hint*: in which case can eddy currents be more easily created?)

Eddy current brakes can be quite useful, and are actually used in the braking mechanism of some (metallic) train wheels. One advantage is that the eddy current braking effect is *stronger* when the wheels spin faster, so as the train slows, the braking gradually lets up on its own, and produces a smooth stopping motion.

Eddy currents are also useful for traffic detection systems, detection of coins in vending machines, and metal detectors – in all three of these cases, one can make use of the induced currents and forces when conductors move through magnetic fields. Can you imagine how eddy currents could be used in each case?

Demonstrating Eddy Currents: Find a small bar magnet (a strong cylindrical refrigerator magnet should work) and drop it vertically through a length of pipe. Now drop a piece of non-magnetic material of about the same size through the tube. The magnet should take much longer to fall through the tube.

Repeat the experiment with a piece of plastic (e.g., PVC) pipe. Now there should be no difference. *Why?*

7.6 Generators

We found in Chapter 6 that we could make a simple electric motor by utilizing the torque on a current loop in a magnetic field. Electromagnetic induction allows the opposite, a *generator* – we can create an electric current by spinning a loop of wire in a magnetic field. In fact, motors and generators rely on the same underlying principles: moving charges experience a force perpendicular to their motion and the magnetic field present. A motor is more or less a generator run in reverse, and vice versa.

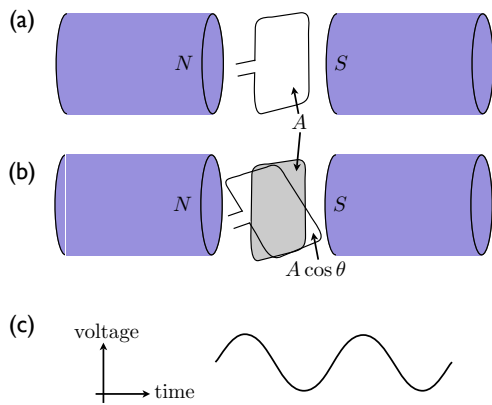


Fig. 7.13 Basis of an electric generator. (a) A loop of wire is rotated inside a permanent magnet. The mechanical input to rotate the loop of wire is converted into electrical energy through induction – voltage is induced in the loop as it rotates. Mechanical input can be supplied by, *e.g.*, steam or falling water. (b) As the loop rotates, the area of the loop perpendicular to the magnetic field changes with respect to the magnetic field. Since magnetic flux is the product of the area perpendicular to the field, this varies from a maximum when the loop is vertical to a minimum of zero when the loop is horizontal as the loop is rotated. (c) This results in an induced voltage which varies sinusoidally with time when the loop is rotated at constant angular velocity. One complete revolution of the loop corresponds to one complete cycle of voltage and current.

Figure 7.13a-c illustrates the basic operation of an electrical generator, which is nothing more than a device to convert mechanical energy into electrical energy. A loop of wire is rotated at constant angular velocity inside a permanent magnet. As the loop rotates, the area it exposes to the magnetic field changes, Fig. 7.13b. Remember magnetic flux is the product of area of the loop, and the perpendicular component of the magnetic field through. In this case, the field does not change, but the area exposed to the field *does*. When the loop is lying parallel to the magnetic field, there is no flux, and when the loop is perfectly flush with the magnet pole faces the flux is maximal.

The rotation of the loop then creates a time-changing magnetic flux through the loop, which varies from maximum to zero and back to maximum. This results in an induced voltage which varies sinusoidally when the loop is rotated at constant angular velocity. One complete revolution of the loop corresponds to one complete cycle of voltage and current, as shown in Fig. 7.13c. Since the current and voltage in the loop varies in time, we call this “alternating current,” which we will cover in slightly more depth in the next chapter.

7.7 A summary of sorts

In the end, we have a nice symmetry between induced electric and magnetic fields. Time changing fields of one sort induce time changing fields of the other sort. First, we have induced *electric* fields from time-varying magnetic fields:

Induced electric fields:

An electric field is induced in any region of space which has a time-changing magnetic field. The induced electric field is proportional to the rate at which the magnetic field changes, and is directed at a right angle to the magnetic field at any instant.

Induced magnetic fields:

A magnetic field is induced in any region of space which has a time-changing electric field. The induced magnetic field is proportional to the rate at which the electric field changes, and is directed at right angles to the electric field at any instant.

Chapter 8

ac Circuits and Electromagnetic waves


Abstract Alternating current (ac) is nothing more than current that varies (sinusoidally) in time, and in Section 7.6 we learned how to produce alternating current with a simple rotating loop generator. As it turns out, nearly all appliances around us run on alternating current - the “wall current” you get from an outlet is alternating current at a frequency of 60 Hz. Not only is ac current important for everyday life, even simple circuits behave differently when powered by time-varying currents and voltages. In this chapter, we will very briefly discuss ac circuits, and move on to electromagnetic waves, which will lead the way to optics and modern physics.

8.1 Resistors in an ac Circuit

An ac circuit is nothing more than various combinations of the components we already know about connected to a sinusoidally varying voltage source, which varies with time as:

$$\Delta V(t) = \Delta V_{\max} \sin \omega t = \Delta V_{\max} \sin 2\pi f t \quad (8.1)$$

here $\Delta V(t)$ is the voltage at any instant in time t , ΔV_{\max} is the peak voltage, ω is the angular frequency, and f the frequency in Hz. The magnitude of an ac not only varies with time, it actually changes sign as well.

Circuit diagram symbol an ac voltage source: 

What happens when we connect components to an ac source? As the simplest example, we will just hook up a single resistor to an ac voltage source, as shown in Fig. 8.1a. We know the voltage across the resistor varies according to Eq. 8.1. Just because the voltage is changing doesn't mean that Ohm's law is not valid, however, so we can immediately find the current through the resistor as a function of time as well:

$$I_R(t) = \frac{\Delta V(t)}{R} = \frac{\Delta V_{\max}}{R} \sin 2\pi f t \quad (8.2)$$

Well, big deal. The voltage goes up and down, and so does the current. This should not be a surprise. The *power* in the resistor is more interesting though. We can readily calculate that from Eq. 4.27:

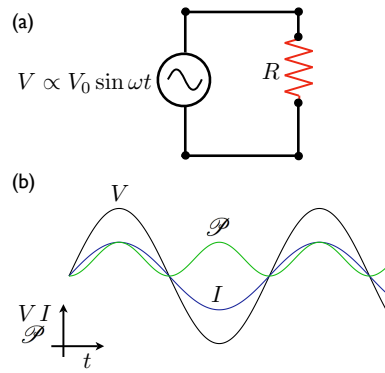


Fig. 8.1 A purely resistive ac circuit. **(a)** A single resistor powered by an ac voltage source, $V(t) = V_0 \cos \omega t$. **(b)** Power $\mathcal{P} = IV$, current I , and voltage V in the resistor. Current and voltage are in phase for a purely resistive circuit.

$$\mathcal{P}_R(t) = I\Delta V = \frac{\Delta V_{\max}}{R} \sin 2\pi ft \cdot \Delta V_{\max} \sin 2\pi ft \quad (8.3)$$

$$= \frac{\Delta V_{\max}^2}{R} \sin^2 2\pi ft \quad (8.4)$$

The power dissipated in the resistor also varies with time. Moreover, its period is only *half* that of the current and voltage. Since the power dissipation in the resistor depends on the product of voltage and current (or the square of voltage *or* current), it doesn't matter if the voltage and current are negative, their *product* is always positive.

Even more interestingly, the power dissipated is actually *zero* whenever the current and voltage go through zero. While the *average* voltage or current over any integer number of periods is zero, the average *power* is not. Further, the dissipation produced by a sinusoidal voltage is *not* the same as just applying a constant dc voltage of ΔV_{\max} , since the alternating voltage is only at its maximum value for an instant. In ac circuits, it is common to use a special kind of average, the **root mean square** or **rms**.

The rms average of a collection of n numbers x_1, x_2, \dots, x_n is defined like this:

$$x_{\text{rms}} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}} \quad (8.5)$$

Basically, the rms average takes the average of the *squares* of the numbers, and then takes the square root of that. The rms average is useful when dealing with periodic functions, since it does *not* average to zero over a full cycle, but gives a sort of averaged amplitude independent of whether the function changes sign. We can find the rms value of the current, voltage, and power with a bit of algebra, but it is tedious. We will merely quote the results:

rms Voltage, Current, and Power in ac Circuits

$$V_{R,\text{rms}} = \frac{\Delta V_{\max}}{\sqrt{2}} \approx 0.707V_{\max} \quad (8.6)$$

$$I_{R,\text{rms}} = \frac{\Delta I_{\max}}{\sqrt{2}} \approx 0.707I_{\max} \quad (8.7)$$

$$\mathcal{P}_{R,\text{av}} = I_{\text{rms}}^2 R \quad (8.8)$$

The average power is just calculated from the rms current or voltage as you would expect. To be concrete: this means that an alternating current of 5 A produces the same dissipation in a resistor as a dc current of $5/\sqrt{2}$ A, about 30% less. The rms voltage, current, and resistance obey Ohm's law just as the maximum values do:

Ohm's law for rms and maximum voltages:

$$\Delta V_{R,\text{rms}} = I_{\text{rms}} R \quad (8.9)$$

$$\Delta V_{R,\text{max}} = I_{\text{max}} R \quad (8.10)$$

With these relationships, we can also relate the power dissipated to the maximum current, just for completeness:

$$\mathcal{P}_{\text{av}} = I_{\text{rms}}^2 R = \frac{1}{2} I_{\text{max}}^2 R \quad (8.11)$$

When you plug an electrical device into the wall, you are connecting it to an ac voltage source. Normal power in the US uses an rms voltage of 120 V, which means that the actual *peak* voltage at the wall outlet is $120 \cdot \sqrt{2}$ V, or about 170 V. It is typically rms values of current and voltage that are quoted for ac circuits, and for the remainder of the chapter, that is what we will quote. One can easily convert between rms and maximum values if desired – it is just a factor $\sqrt{2}$ in the end.

8.2 Capacitors in ac Circuits

Resistors in ac circuits offered only a few surprises. What about capacitors? Understanding how a capacitor responds to a sinusoidally-varying voltage requires reminding ourselves how a capacitor responds to any sort of changing voltage. If we connect a capacitor to a constant voltage source, as soon as the switch to the voltage source is closed the capacitor begins to charge. A large current flows initially as the capacitor charges. As the capacitor gains more and more charge, the voltage drop across it increases, which opposes the change in current. After several time constant's worth of waiting, the capacitor is fully charged, and current no longer flows. If we turn off the voltage source, a current again flows while the capacitor discharges, but again the current goes to zero after a short time. A capacitor therefore restricts current flow to very short time intervals, depending on its time constant $\tau = RC$.

If we connect a single capacitor to an ac voltage source, Fig. 8.2, what will happen? At $t = 0$ on the graph, the voltage (blue curve) starts from zero and quickly increases. Ramping up the voltage on the capacitor means that a large current will flow (black curve), attempting to charge the capacitor. The faster the voltage increases – the larger the slope of the $V(t)$ curve – the larger the current will be. When $V(t)$ reaches its plateau one quarter of the way through the cycle, the voltage is nearly constant, and no current flows through the capacitor. Shortly thereafter, the voltage *decreases*, and the capacitor responds by discharging, again at a rate proportional to the slope of the $V(t)$ curve. Once the voltage changes sign, the capacitor begins charging up again with the opposite polarity, and the whole cycle repeats itself. What is important to realize is that in ac circuits, current *does* flow through capacitors – it is just like the RC circuits we studied earlier, except that now we are effectively turning the voltage on and off continuously.

The current on the capacitor reaches its maximum positive and negative values whenever the voltage is zero. Similarly, the current goes to zero whenever the voltage is at a maximum, as at those points the voltage is momentarily essentially constant. *In the end, this leads to the current through the capacitor also being sinusoidal, but with a quarter cycle (90°) phase shift.* The usual way of stating this is that the “voltage lags the current by 90° ,” a reference to the fact that the current reaches its maximum a quarter cycle *after* the voltage does. More mathematically, the current response has a $+90^\circ$ phase shift with respect to the driving voltage.

How much current flows through the capacitor? We can qualitatively figure out what it depends on already. As the voltage varies, the capacitor only allows the most current to pass when the voltage is changing the most rapidly. We expect, then, that the current goes *up* as the frequency of the voltage goes *up* and the voltage changes faster and faster. As the capacitance gets larger, more and more charge is required, so we should also expect *larger* current for a *larger* capacitor.

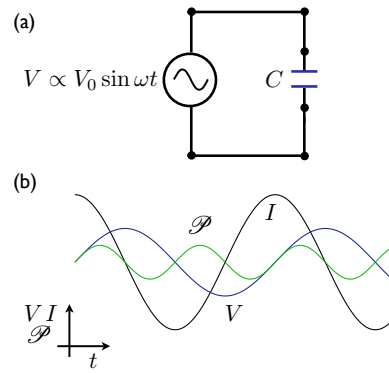


Fig. 8.2 A purely capacitive ac circuit. **(a)** A single capacitor powered by an ac voltage source, $V(t) = V_0 \cos \omega t$. **(b)** Power $\mathcal{P} = IV$, current I , and voltage V in the resistor. Current and voltage are also 90° out of phase (the voltage “lags” the current”) for a purely capacitive circuit.

Making this quantitative involves generalizing Ohm’s law to ac circuits. For resistive elements, this is not necessary, Ohm’s law works just fine. What we need is a way to relate current and voltage for *reactive* elements, like capacitors and inductors, that react to changes in current and voltage. Instead of resistance, reactive elements like capacitors and inductors have what is called a **reactance** X :

“Ohm’s Law” for Reactive Elements:

$$\Delta V_{\text{rms}} = I_{\text{rms}} X \quad (8.12)$$

where X is the *reactance* of the circuit element. Capacitors and inductors are reactive elements, resistors are not.

Units of Reactance X : if C is in farads [F] and f in hertz [Hz], reactance is in Ohms [Ω]

For capacitors, the reactance has just about the form we would expect: inversely proportional to frequency and capacitance:

Reactance for a Capacitor:

$$X_C = \frac{1}{2\pi f C} \quad (8.13)$$

where f is the frequency of the ac voltage, and C is the capacitance.

As the frequency of the voltage increases, the reactance decreases, and the current increases. Similarly, as capacitance increases, the current increases:

$$I_{\text{rms}} = \frac{\Delta V_{\text{rms}}}{X_C} = 2\pi f C \Delta V_{\text{rms}} \quad (8.14)$$

What about the power in a capacitive ac circuit? Figure 8.2 shows the current, voltage, and power for a capacitor connected to an ac voltage source. Since the voltage and current are now 90° out of phase, the maximum power now occurs halfway between the maximum current and voltage. Further, now the power can become *negative*. What does that mean? Simple, During the charging cycle, the power is positive as it is for a resistor, meaning that the source is supplying energy to the capacitor. During the discharging cycle, the capacitor is pushing charges *back* to the source, and effectively, the source is draining energy from the capacitor. Charge gets pushed back and forth between the source and capacitor, and the power swings from positive to negative. We get nothing for free, however –

during the discharge cycle, the capacitor is just pushing back the charges it stored during the charging cycle, energy is still conserved. In this light, a capacitor is useful as either temporary energy storage device, or as a way of generating a time-delayed response.

8.3 Inductors in ac Circuits

We know now that inductors respond to current the same way capacitors respond to voltage. Indeed, this is no different for ac circuits, and inductors are also **reactive elements** just like capacitors. When an inductor is connected to an ac source, the alternating voltage attempts to push an alternating current through the inductor. The inductor responds by developing a voltage across its terminals to impede the current flow – the more rapidly the source tries to force a current through the inductor, the larger $\Delta I/\Delta t$, the larger the voltage developed on the inductor and the larger the resistance to current flow. Hence, we would expect that current through the inductor would *decrease* as the frequency *increases*. Further, the voltage across the inductor is proportional to the value of the inductance L , so for larger L we expect even less current. The reactance of an inductor in fact behaves in just this way:

Reactance for an Inductor:

$$X_L = 2\pi fL \quad (8.15)$$

where f is the frequency of the ac voltage, and L is the inductance.

Now consider an inductor L connected to an ac voltage source, Fig. 8.3. It is a bit easier to begin describing the inductor's behavior starting at one quarter cycle, when the voltage is at its maximum. At this point, the voltage is momentarily constant, and so is the current in the inductor, so it offers no resistance. As the voltage begins to decrease, its time variation (slope) increases, and the inductor offers more and more resistance to current flow. The voltage across the inductor is *opposite* that of the source, and it tries to push current *back* into the source. When the change in voltage with time is maximum, when the voltage crosses zero, the inductor is pushing a maximum current back to the source.

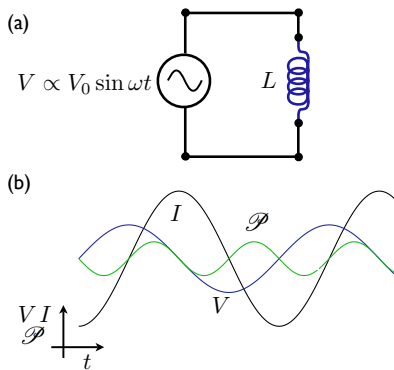


Fig. 8.3 A purely inductive ac circuit. (a) A single inductor powered by an ac voltage source, $V(t) = V_0 \cos \omega t$. (b) Power $\mathcal{P} = IV$, current I , and voltage V in the resistor. Current and voltage are 90° out of phase (the voltage “leads” the current”) for a purely inductive circuit.

As the source voltage becomes negative and its variation slows, the inductor current decreases, and when the voltage reaches its minimum, the inductor current is zero. Just like in a capacitor, the maximum current and voltage are one quarter cycle apart, but now the current is increasing *ahead* of the voltage, and we say that the “voltage across the inductor leads the current by 90° .” Again, more mathematically, the current response has a -90° phase shift with respect to the driving voltage.

The power in an inductive circuit behaves similarly to that in a capacitive circuit – for half of the cycle, while the voltage is decreasing, the inductor is absorbing energy from the source and storing

it in its magnetic field, and for the other half, while the voltage is increasing, it is pushing energy back to the source (Fig. 8.3b).

8.4 Filters

What neat things can we do with ac circuits? Already, we know enough to build simple signal filters. Consider the circuit in Fig. 8.4, which we have drawn in a manner closer to what electrical engineers typically use. A voltage V_{in} is sent in from the left, defined relative to ground. That is, a voltage V_{in} is applied between a “positive” signal wire (the upper wire), and a ground wire. A resistor R connects the signal wire to the output V_{out} , and a capacitor connects the input to ground. Basically, the resistor and capacitor are in series, and the output voltage is taken across the capacitor. What happens in this circuit?

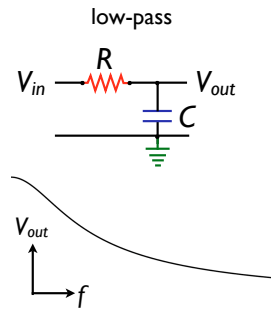


Fig. 8.4 An RC low-pass filter. Capacitors present a low reactance to high frequency signals, so they are selectively returned to ground before the output.

Resistors present an equal resistance to signals of any frequency, but capacitors present a *lower* reactance to high frequency signals. High-frequency signals entering from the input see the capacitor as a low reactance path to ground, and thus most of the high-frequency signal takes this path to the ground and never reaches the output. Low-frequency signals see the capacitor as a high reactance and avoid this path, so most of the low-frequency signal reaches the output. What this circuit really does is selectively filter out the high-frequency portions of a mixed frequency signal, and let the low frequency signals pass through – the ratio between the input voltage V_{in} and the output V_{out} depends on frequency. For this reason, this circuit is known as a “low-pass” filter. A circuit like this could be used to direct the low-frequency portions of an audio signal to a “woofer” speaker, for instance. The frequency response of this type of filter is also shown in Fig. 8.4.

Why is the resistor there, and what is the range of filtered frequencies? What this circuit can also be thought of is a *generalization of the resistive voltage divider* (series resistors), where the voltage division factor depends on frequency. When the reactance of the capacitor is equal to the resistance, half of the input *power* goes through the resistor, and half through the capacitor. Since power goes as voltage *squared*, when the reactance equals the resistance the output will be reduced by a factor $1/\sqrt{2}$ relative to the input, about 70%. The reactance and resistance will be equal at one particular frequency, the *cutoff frequency*, whose value is given by:

$$X_C = \frac{1}{2\pi f_{\text{cutoff}} C} = R \quad (8.16)$$

$$\Rightarrow 2\pi f_{\text{cutoff}} = \frac{1}{RC} = \frac{1}{\tau} \quad (8.17)$$

$$\text{and } \frac{V_{out}}{V_{in}} = \frac{1}{\sqrt{2}} \quad \text{at the cutoff frequency} \quad (8.18)$$

It should not be too surprising at this point that the cutoff frequency is just the same as one over the time constant. After all, this is precisely the same RC circuit we studied earlier, just viewed in the frequency domain instead of the time domain!

What about the circuit in Fig. 8.5? In this case, an inductor and resistor are in series. The inductor presents a low reactance to low-frequency signals, and they will be preferentially sent to ground before reaching the output. High-frequency signals will avoid the inductor, and pass easily to the output. Thus, this circuit is a “high-pass” filter, selectively filtering out the low-frequency portions of a mixed frequency signal, and letting high-frequency portions pass through. A high pass filter like this one could be used to send high frequency audio signals to a “tweeter” speaker, and block lower frequency bass signals that may damage it. More complicated (but still recognizable) high-pass and low-pass filters are used in audio equipment in exactly this way.

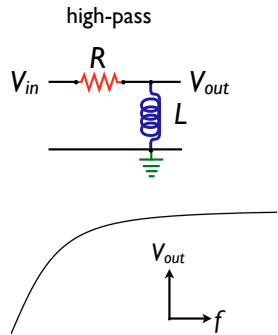


Fig. 8.5 An RL high-pass filter. Inductors present a low reactance to low frequency signals, so they are selectively returned to ground before the output.

The cutoff frequency of the RL filter can be determined just like we did above for the RC filter – when the reactance of the inductor equals the resistance, half the power goes to each component.

$$X_L = 2\pi f_{\text{cutoff}} L = R \quad (8.19)$$

$$\Rightarrow 2\pi f_{\text{cutoff}} = \frac{R}{L} = \frac{1}{\tau} \quad (8.20)$$

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{\sqrt{2}} \quad \text{at the cutoff frequency} \quad (8.21)$$

The frequency response of this filter is shown in Fig. 8.5. Once again, the cutoff frequency is just the inverse of the time constant, since the frequency- and time-domain descriptions are inverse points of view.

Question: What if we switch the position of the R and C in Fig. 8.4?

The circuit becomes a high-pass filter instead of a low-pass filter. The cutoff frequency is the same.

Question: What if we switch the position of the R and L in Fig. 8.5?

The circuit becomes a low-pass filter instead of a high-pass filter. The cutoff frequency is the same.

There is much, much more we can do with ac circuits, we have only just scratched the surface. Now it is time to move on once again, and work our way away from electricity and magnetism toward optics. Of course, optics is *also* electricity and magnetism, as we shall see!

8.5 Electromagnetic Waves

8.5.1 Electromagnetic fields of accelerating charges

We have so far discussed charges moving at constant velocity (electric currents), and stationary charges. The former gave rise to magnetic fields, while the latter give rise to electric fields. Oscillating charges (or more generally *accelerating charges*), on the other hand, give rise to *both electric and magnetic fields*.

To see how this might be, think of a charge moving in a sinusoidal pattern, like a charged mass hanging from a spring. While the charge is at the center of its motion, it is moving at constant velocity, and it creates a magnetic field. When the charge is at the top or bottom of its motion, it is stationary for an instant, and gives rise to an electric field. What happens in between? When charges accelerate, and Maxwell first predicted that *both* electric and magnetic fields are created.

When the electric and magnetic fields change in time, and create electromagnetic disturbances that travel through space as waves, like ripples in a pond. The waves created by accelerating charges are spatially and temporally fluctuating magnetic and electric fields, and are called *electromagnetic waves*, or EM waves.

Electromagnetic waves travel at the speed of light, $c = 3 \times 10^8$ m/s - in fact, electromagnetic waves *are* light, and vice versa. Whenever a charged particle accelerates, it radiates EM waves. Since electric and magnetic fields in a volume of space represent energy, *whenever a charged particle accelerates, it radiates energy*.

8.5.2 Production of Electromagnetic Waves by an Antenna

In order to get an idea of how electromagnetic waves work, we will consider a simple antenna connected to an alternating voltage source, Fig. 8.6. The alternating (sinusoidal) voltage source applied to the two antenna wires causes electric charges in the wires to oscillate (this is basically how a broadcast antenna works). For the sake of argument, the alternating voltage source has a period T , and a frequency of $f = 1/T$.

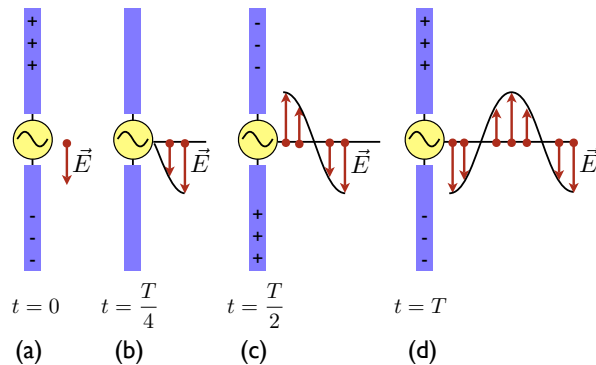


Fig. 8.6 An electric field set up by oscillating charges in an antenna. The field moves away from the antenna at the speed of light. At a given point in space, the electric field intensity oscillates as a function of time, and at a given time, the electric field intensity oscillates spatially. Note that this is true whether the antenna is *producing* the radiation or *receiving* it!

At time $t=0$, Fig. 8.6a, the upper rod is at a maximum positive voltage, and the lower a maximum negative voltage. Thus the upper rod is given a maximal positive charge, and the lower rod a maximal negative charge. The electric field at this instant is pointing downward. As the voltage source oscillates, the voltage and amount of charge on each rod decreases, reaching zero at one quarter of the source's period T ($t = T/4$) as shown in Fig. 8.6b. At this point, $E = 0$ at the antenna. The maximum E field created $T/4$ seconds earlier, however, has not disappeared! It has travelled at a

velocity c for $T/4$ seconds, so it is $cT/4$ meters away from the antenna. Remember, time-varying E field travels from the antenna at the speed of light c .

At a still later time $T/2$, Fig. 8.6c, the voltage source has completed one half cycle, and has reversed polarity. Now the E field at the antenna is reversed in direction, and again at a maximal value. This continues on, Fig. 8.6d, and the E field at the antenna oscillates in phase with the induced charge distribution as we would expect. At any instant, the E field at the antenna depends on the charge on the rods at that instant, and therefore the voltage applied by the source.

Basically, we have set up a charge distribution which oscillates in time, just like a mass on a spring oscillates. Since the motion is oscillatory, we know that the charges are accelerating, and therefore, radiating energy. One part of the radiation is just the E field traveling out from the source at a velocity c .

While the charges oscillate, they also constitute a time-varying *current* in the rods. The current is maximal when the voltage and E field are maximal, since that is the moment when the most charge is moving in the smallest amount of time. Likewise, when $E = 0$, the current is zero. The presence of a current means there is also a magnetic field created, as shown in Fig. 8.7.

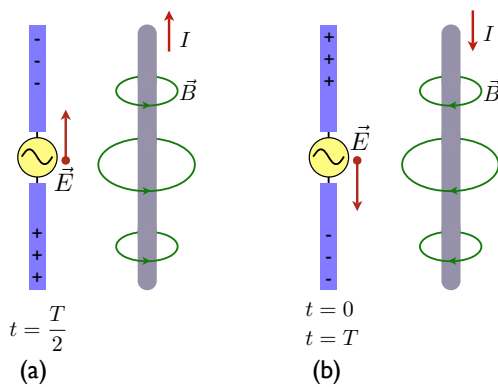


Fig. 8.7 Magnetic fields around an antenna carrying an alternating current for two different times in the current cycle: **(a)** $t = T/2$, and **(b)** $t = 0, T$. The current is in phase with the voltage source. Note that this is true whether the antenna is *producing* the radiation or *receiving* it!

The magnetic field oscillates in time, in phase with the current and the E field. By the right hand rule, \vec{B} is always perpendicular to \vec{E} . This is the other part of the radiation of the accelerating charges, the B field traveling out from the source at a velocity c .

The basic result of this is that changing magnetic fields produce an electric field, and changing electric fields produce a magnetic field. These induced electric and magnetic fields are always in phase (they reach maximum and minimum values at the same point), and the fields are at right angles.

8.5.3 Properties of Electromagnetic Waves

What more can we say about EM waves? First, many everyday EM can be described as **plane waves**, in particular when the EM wave is far from its original source. Figure 8.8 shows a plane wave at an instant in time traveling along the x -axis. The oscillations of the E and B fields occur in planes perpendicular to the x -axis, or perpendicular to the direction the wave is traveling. Even though E and B oscillate spatially, they are always perpendicular, and along with the direction of travel, obey a “handedness” rule ... which leads us to yet another right-hand rule.

Right-hand rule #3 for plane EM waves:

1. Point your right-hand fingers along \vec{E} .
2. Curl them along the direction of \vec{B} .

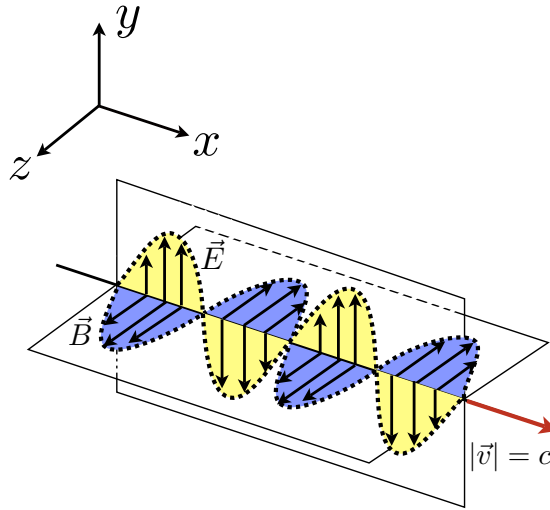


Fig. 8.8 An electromagnetic wave at one instant of time, moving in the positive x -direction with speed c . The electric field points along the y axis, and is perpendicular to the magnetic field at every point. Both \vec{E} and \vec{B} are perpendicular to the direction of wave propagation.

3. Your right thumb points along the direction the wave is traveling.

Electromagnetic waves travel with the speed of light, which in fact relates the permeability ϵ of a medium ($1/\epsilon$ relates to the strength of E), the permittivity μ of a medium (μ relates to the strength of B) and the speed of light c :

Speed of light in free space:

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}} = 2.99792 \times 10^8 \text{ m/s} \quad (8.22)$$

As it turns out, this also implies a relation between E and B themselves:

Relationship between $|\vec{E}|$ and $|\vec{B}|$ in an EM wave:

$$c = \frac{|\vec{E}|}{|\vec{B}|} \quad (8.23)$$

Just like in the mass spectrometer (Sect. 6.3.1.1), perpendicular \vec{E} and \vec{B} fields imply a particular velocity, and given that EM waves travel at c , this implies a fixed ratio $|\vec{E}|/|\vec{B}|$.

8.5.4 Energy transferred by EM waves

Accelerating charges radiate EM waves, which really means they radiate energy. How much energy? For a given EM wave, we can define an intensity of radiation \mathcal{I} which is the amount of energy absorbed per unit time, per unit (surface) area. Since the intensity of E and B ($|\vec{E}|$ and $|\vec{B}|$) vary in time, and clearly the amount of radiation absorbed depends on how big the surface area is, this is

the best we can do. Of course, energy per unit time is just power, so really \mathcal{I} is power per unit area [W/m^2].

Intensity of EM radiation

$$\mathcal{I} = \frac{\text{energy}}{\text{time} \cdot \text{area}} = \frac{E_{\max} B_{\max}}{2\mu_0} = \frac{\text{power}}{\text{area}} = \frac{E_{\max}^2}{2\mu_0 c} = \frac{c B_{\max}^2}{2\mu_0} \quad (8.24)$$

Here E_{\max} is the maximum of the E field in the wave (the amplitude), and B_{\max} is the amplitude for B . The units of \mathcal{I} are then Watts per square meter [W/m^2].

The energy U transferred to an area A in a time interval Δt is then just

Energy transferred by an EM wave in a time Δt :

$$U = (\text{energy per unit time per unit area}) \cdot (\text{area}) \cdot (\text{time}) = \mathcal{I} \cdot A \cdot \Delta t \quad (8.25)$$

So the larger the area, and the longer the time of exposure, the more energy that is transmitted by radiation. This much we already know first-hand during the Alabama summer.

If energy is transferred, then a *momentum* p must also be transferred. Though this may not seem intuitive, incident EM radiation (light) imparts momentum on anything it strikes and transfers energy to. Clearly this is a small effect, since we do not notice it being harder to walk toward or away from the sun! If we take a perfectly black surface, which absorbs *all* incident energy, it turns out the momentum transfer is:

$$|\vec{p}| = \frac{U}{c} = \frac{\mathcal{I} \cdot A \cdot \Delta t}{c} \quad \text{complete absorption} \quad (8.26)$$

This is the analogy of a perfectly inelastic collision (like one mass striking another and sticking to it). If all radiation is reflected, the analogy of a perfectly elastic collision, then the momentum transfer is twice as big. Rather than the incident EM wave simply being absorbed, which changes its velocity from c to zero, now it is reversing direction, which changes its velocity from $+c$ to $-c$.

$$|\vec{p}| = \frac{2U}{c} = \frac{2\mathcal{I} \cdot A \cdot \Delta t}{c} \quad \text{complete reflection} \quad (8.27)$$

We will come back to the subject of radiation pressure and the momentum of light in later chapters, and we will be able to more carefully explain *why* these formulas must be the way they are.

The momentum imparted by EM radiation is known as *radiation pressure*. If we remember that force can be defined as a change of momentum ($F = \Delta p / \Delta t$), and pressure is force per unit area ($P = F/A$):

EM Radiation Pressure:

$$\begin{aligned} P_{\text{radiation}} &= \frac{\mathcal{I}}{c} = \frac{E_{\max}^2}{2\mu_0 c^2} = \frac{B_{\max}^2}{2\mu_0} && \text{complete absorption} \\ P_{\text{radiation}} &= \frac{2\mathcal{I}}{c} = \frac{E_{\max}^2}{\mu_0 c^2} = \frac{B_{\max}^2}{\mu_0} && \text{complete reflection} \end{aligned} \quad (8.28)$$

Direct sunlight on Earth only imparts a momentum of about $5\mu\text{N/m}^2$, so these effects are very small, but measurable. In the solar system, radiation pressure is an important effect, it tends to push particles smaller than $\sim 0.1\mu\text{m}$ outward from the sun.

8.5.5 The EM spectra

All electromagnetic waves travel in vacuum at the speed of light c . As with any other waves, the velocity, frequency, and wavelength are related:

Frequency f , wavelength λ , and the speed of light c in vacuum

$$c = \lambda f = 2.99792 \times 10^8 \text{ m/s} \quad (8.29)$$

Electromagnetic waves cover many orders of magnitude in frequency and wavelength, but always obey this relationship. Since the velocity of light in vacuum c is fixed, this means *if you know either f or λ , you automatically know the other*. Figure 8.9 shows the frequency and wavelength ranges for some types of EM waves. Note that our common definitions of wave types are not precise, and overlap (e.g., X-rays and UV). Section 21.12 in Serway has a nice discussion of different sorts of EM waves.

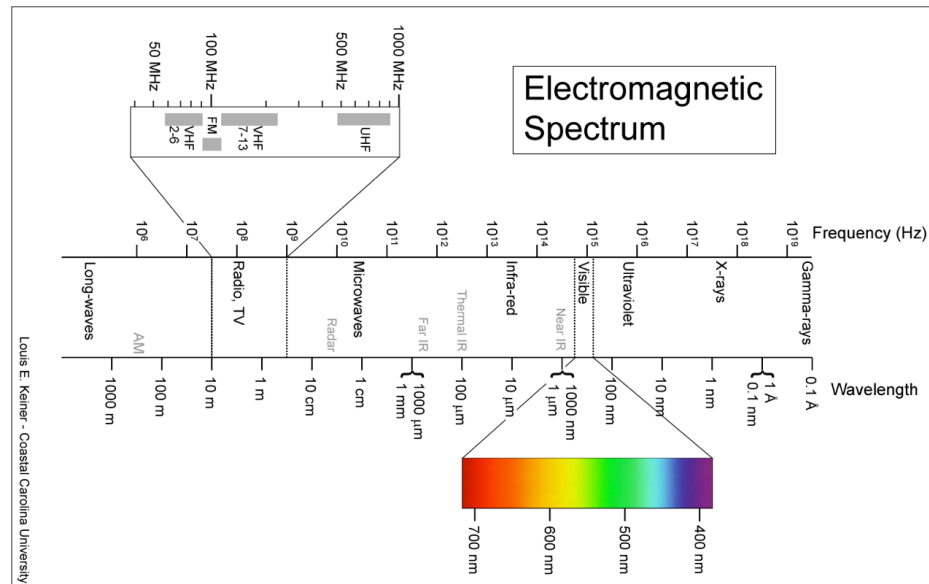


Fig. 8.9 The electromagnetic spectrum. Note that our common definitions of wave types are not precise, and overlap (e.g., X-rays and UV). Note the expanded views of the visible spectrum and common communications frequencies. At the smallest wavelengths, Ångström units are commonly used, $1\text{\AA}=10^{-10}\text{ m}$. Image from L. Keiner, <http://www.keiner.us/>.^[22]

Part III

Optics

Chapter 9

Reflection and Refraction of Light

Abstract Light as we know it is nothing more than a particular sort of electromagnetic wave, defined by a rough range in frequency or frequency. Visible light covers wavelengths of $\sim 400\text{--}700\text{nm}$, while ultraviolet (UV) and infrared push this definition to $\sim 10\text{nm--}10\mu\text{m}$. In the next chapters when we discuss optics, we will focus on applications to visible light, though everything we discuss will be applicable to the whole electromagnetic spectrum, from radio waves to gamma-rays.

9.1 The Nature of Light

Until early in the 19th century, light was modeled as a steady stream of particles, which entered and stimulated the eye in analogy to the sense of touch. This model (primarily due to Newton) was very successful in describing most everyday properties of light, like reflection and refraction. As an added advantage, it seemed to explain the sense of sight by rough analogy with the sense of touch.

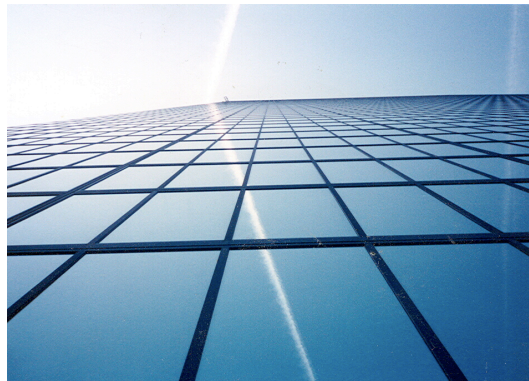


Fig. 9.1 Reflection of a jet contrail off of the John Hancock tower in Boston, Ma. Photo taken by the author.

Now we know that light is just a form of electromagnetic radiation, a *wave* phenomena. Despite clear experimental evidence for the wave behavior of light, such as interference of light waves and diffraction, this point of view was slow to gain acceptance.

As it turns out, light behaves as *both a wave and a stream of particles*, depending on how you observe it. In this and the following chapters, we will try to understand when to use each point of view, and why they are both valid. For example, classical EM wave theory describes interference very well, whereas photoelectric effects (key to solar cells) require the particle point of view. In general, when we worry about light interacting with individual electrons or other particles, the “stream of particles” picture is useful. In the end, light is neither one nor the other: light has a number of demonstrable properties, some of which are best *modeled* as waves, others which are better modeled as particles. The reality of the situation is strange, we model it as best we can!

For the moment, however, we will treat light as a steady stream of particles, as Newton did, and explore the phenomena that fall under this description. Einstein provided the first definitive theory of light as being made up of individual particles, known as *photons*. According to Einstein's theory¹, light particles, or photons each carry an energy proportional to their frequency:

Photon energy:

$$\mathcal{E} = hf = \frac{hc}{\lambda} \quad (9.1)$$

where $h = 6.63 \times 10^{-34}$ J·s is *Planck's constant*.

Here we have used Eq. 8.29, the relationship between frequency, wavelength, and speed for light waves:

Frequency f , wavelength λ , and the speed of light c in vacuum

$$c = \lambda f \quad (9.2)$$

9.1.1 Wave Packets and Wave-Particle Duality

Now wait a minute: we are modeling light as *particles*, but just used the *wave* relationship between c , λ , and f ! Are we getting away with something? No, both wave and particle viewpoints are valid, light really behaves as *both*. One way you can view this duality is to imagine light as a *modulated* wave, or “wave packet” as shown below:

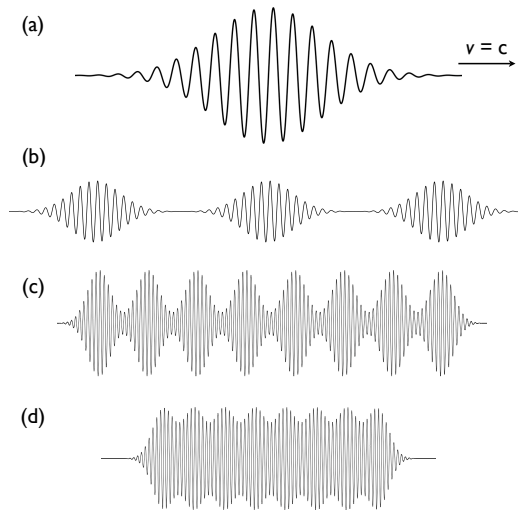


Fig. 9.2 A “wave packet” in some sense combines particle and wave properties. **(a)** A Gaussian wave packet, $y = e^{-x^2} \sin x$, illustrating how a modulated wave can “look like a particle.” The localized bundle of high intensity can behave like a particle. **(b-d)** A series of wave packets more and more closely spaced together, mimicking the transition from a stream of photons to a ray of light.

Almost all of the intensity is within the central region, hence the term wave packet. Light waves of this sort mimics the behavior of single particles. If you used a light detector to measure such a wave, you would observe discrete “ticks” corresponding to the wave packets, and nothing in between. So this isn't so weird and mysterious at all - light is a wave, and it can behave like particles just because the waves *aren't simple sin's and cos's!* The waves have regions of localized intensity, which can

¹ He won his Nobel prize for this, not relativity.

be thought of as particles. As it turns out, when we discuss Quantum Mechanics, this is also the appropriate point of view for any other particle, such as electrons.

For the rest of this chapter, we will view light as a steady stream of particles.

9.2 Reflection of Light

When light traveling in one medium encounters a boundary leading into another medium, reflection and refraction can result. *Reflection* means that part of the light encountering the second medium bounces off of it, and *refraction* means that part of the light enters the second medium, but bends during the process. More often than not, both processes occur when light travels between two media.

9.2.1 The Ray Approximation

If we view light as a steady stream of particles, we can consider these streams of light particles to be “rays” which travel in straight-line paths - the so-called **ray approximation**. Since the speed of light is constant, light particles can have no acceleration, and therefore (by Newton’s first law) must travel in straight-line paths. Our “light rays” are the paths of individual photons (or wave packets) if you like, or the “wave front” connecting the points of all EM waves with the same amplitude and phase. But more on that later.

In the ‘ray approximation’, light travels in a straight-line in a (homogeneous) medium, until it encounters a boundary between two different materials. At this boundary, light is either reflected from it, passes into the second medium on the other side of the boundary, or does a bit of both.

9.2.2 The Law of Reflection

When a light ray traveling in a (transparent) medium encounters a boundary into a second medium, part of the ray is reflected back into the first medium. Figure 9.3a illustrates several light rays “bouncing” off of a perfectly flat surface (or an interface between two media). The rays are all parallel as they are incident on the surface, they all reflect at the same angle known, and leave parallel. This is known as **specular reflection**.

On the other hand, if the surface (or the interface between two media) is rough, the reflected light comes out in many directions. This is known as **diffuse reflection**, and is shown in Fig. 9.3b. A surface is considered “smooth” and behaves as such so long as the roughness is small compared to the wavelength of the light. If the roughness is small compared to the wavelength, the light cannot “see” it.

Law of Reflection

When a light ray is reflected off of a surface, the angle of incidence θ_i is equal to the angle of reflection θ_r .

For a single ray, the angle of reflection equals the angle of incidence, as illustrated in Fig. 9.4. Experimentally, this is true, and it also follows from the boundary conditions on the E and B fields (Appendix ??). Reflective optics is pure geometry - no matter how many media you consider in sequence, it all boils down to geometry.

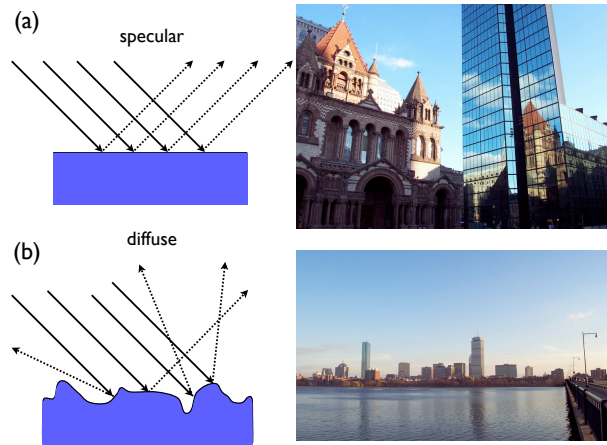


Fig. 9.3 An illustration and example of (a) specular reflection, where the reflected rays are all parallel to each other, and (b) diffuse reflection, where roughness scatters the reflected rays. The picture in (a) is the Trinity Church reflecting off of the Hancock Tower in Boston, Ma, the picture in (b) is a view of the Boston skyline reflecting off of the Charles River, taken from Cambridge, Ma. Photos by the author.

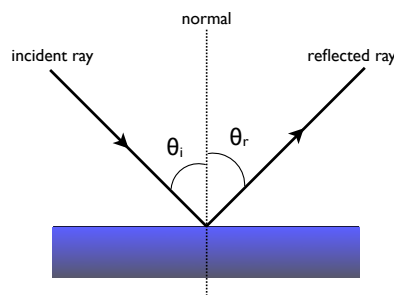


Fig. 9.4 Law of reflection: the angle of incidence is the same as the angle of reflection. (Both angles are measured with respect to a surface normal).

9.3 Refraction of Light

What happens when light is not reflected, and actually goes from one medium to another, say from bare vacuum into a piece of glass? In a vacuum, there is no matter for the light to interact with or scatter off of. The same is roughly true for air. When light enters some dense medium, however, this is no longer true. Physically, what happens is that the light, being electromagnetic radiation, interacts with the electrons and nuclei that make up the medium. The \vec{E} and \vec{B} fields making up the EM wave are affected by the electrons and nuclei, just like field lines from a point charge are affected by another point charge, with the result that *its velocity is reduced*. In some sense, having to interact with atoms retards the light.

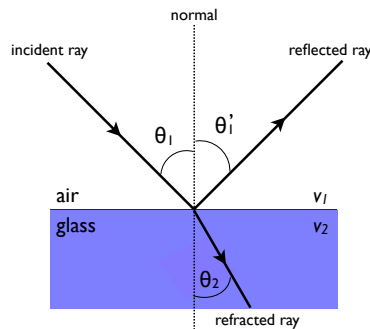


Fig. 9.5 Light rays incident on an air-glass interface. The refracted ray is bent toward the interface normal because $v_2 < v_1$.

When light encounters a new medium in which its velocity is changed, adjusts its direction in such a way to *spend more time in the medium with the higher velocity*. The result of this is that when

a light ray encounters a boundary between two transparent media, part of the ray is reflected, and part of it is bent as it enters the second medium, as shown in Fig. 9.5. The “bent” beam is said to be *refracted*. The incident, reflected, and refracted beams all lie in the same geometric plane along with the interface normal at the point of incidence.

If we first consider the case $v_2 < v_1$, as shown in Fig. 9.5, the light ray is going slower in the second medium, so it would like to minimize the distance through that media it has to cover. That is accomplished by traveling through the second media more closely to normal incidence than before - the light ray bends toward the surface normal to get out of the second medium more quickly.

A slightly more formal way to state this is through *Fermat's principle*, or the “principle of least time”:

Fermat's principle of least time:

The path taken between two points by a ray of light is the path that can be traversed in the least time.

This principle is sometimes taken as the *definition* of a ray of light. Using calculus, one can show that the angle θ_2 “chosen” by the ray of light depends on the angle of incidence and the velocities of light in the two media:

Refraction and velocity of light in media:

$$\frac{\sin \theta_2}{\sin \theta_1} = \frac{v_2}{v_1} = \text{constant} \quad (9.3)$$

The slower the velocity of light in the second medium, the more sharply the light ray bends toward the normal - the light minimizes its time spent in the slower medium by shortening its path.

On the other hand, what if $v_1 > v_2$? We could say that the light ray is going more quickly in the second medium, and can afford to spend a bit more time after going so slowly in medium 1. An easier way to think about it is that *for refraction we can run the light rays forwards or backwards, and we have to get the same result*. That is, the path of a light ray through a refracting surface is reversible. If light travels through the air and bends into the glass as shown, then light coming from the glass into the air has to behave the same way.

In any case, the light rays bend closer to the normal if they enter a region where they have lower velocity, and away from the normal in a region where they have higher velocity, as shown below. Light wants to get out of regions of low velocity by traveling more normal, and spend more time in regions of high velocity by making a more shallow angle.

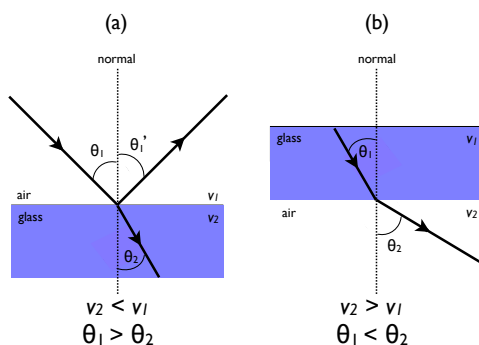


Fig. 9.6 (a) When light moves from air into glass, its path is bent toward the normal since the velocity of light is reduced in glass compared to air. (b) When light moves from glass into air, its path is bent away from the normal, since it has now entered a region of lower velocity.

9.3.1 Snell's Law

For convenience, we often define a material constant which represents the ratio between the speed of light in vacuum and in a given material, the **index of refraction** n :

Index of refraction n :

$$n = \frac{\text{speed of light in vacuum}}{\text{speed of light in a medium}} = \frac{c}{v} = \sqrt{\epsilon_r \epsilon_0 \mu_r \mu_0} = \sqrt{\kappa \epsilon_0 \mu_r \mu_0} \quad (9.4)$$

The last two relationships come from Eq. 8.22, relating the speed of light to the permeability (μ_r) and permittivity (ϵ_r) or dielectric constant (κ) in a material. The index of refraction is dimensionless (it has no units), and of course $n = 1$ for vacuum itself. If we use this definition, we can rewrite Eq. 9.3:

$$\frac{\sin \theta_2}{\sin \theta_1} = \frac{v_2}{v_1} = \frac{n_1}{n_2} \quad (9.5)$$

The index of refraction for many common conducting materials is listed in Table 9.1, compiled from several sources.[23]

Table 9.1 Refractive indices at $f = 5.09 \times 10^{14}$ Hz (yellow/orange)

Material	n	Material	n
Vacuum	1 (exactly)	Helium	1.000036
Air (STP)	1.0002926	Carbon Dioxide	1.00045
Water ice	1.31	Liquid water (20°)	1.333
Acetone	1.36	Teflon	1.35-1.38
Glycerol	1.4729	Acrylic glass	1.490-1.492
Rock salt	1.516	Crown glass (pure)	1.50-1.54
Salt (NaCl)	1.544	Polycarbonate	1.584-1.586
Flint glass (pure)	1.60-1.62	Crown glass (impure)	1.485-1.755
Bromine	1.661	Flint glass (impure)	1.523-1.925
Cubic Zirconia	2.15-2.18	Diamond	2.419

When light travels from one medium to another, its *speed* changes, but its frequency does not. Examine Fig. 9.7. When a wave passes from material 1 to material 2, the frequency at which waves arrive at the boundary from 1 must equal the rate at which waves leave the boundary into 2. If they were not equal, since EM waves carry energy, we would have to create or destroy energy at the boundary. This is clearly not OK. If the energy has to be conserved across the boundary, then by Eq. 9.1, the frequency must be conserved too. So $f_1 = f_2 \equiv f$.

On the other hand, we know from Eq. 8.29 that the speed of light must be related to frequency and wavelength:

$$v_1 = f\lambda_1 \quad \text{and} \quad v_2 = f\lambda_2 \quad (9.6)$$

Since we know $v_1 \neq v_2$, the only way out is if $\lambda_1 \neq \lambda_2$!

Light passing from one medium into another different medium:

1. Velocity changes.
2. Frequency does not change.

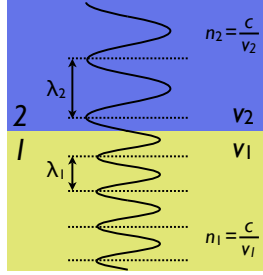


Fig. 9.7 When a wave moves between medium 1 and medium 2, its wavelength changes, but its frequency does not.

3. Wavelength changes.

Putting together what we know, we can relate the changes in frequency and speed to the index of refraction in the two media:

Ratio of frequencies, velocities, and refractive indices in media:

$$\frac{\lambda_1}{\lambda_2} = \frac{v_1}{v_2} = \frac{c/n_1}{c/n_2} = \frac{n_2}{n_1} \quad \Rightarrow \quad \lambda_1 n_1 = \lambda_2 n_2 \quad (9.7)$$

Put another way, λ times n is a *conserved quantity* for light. Really, this is just a restatement of Eq. 9.1 plus conservation of energy - $\mathcal{E} = hc/\lambda = h\nu/\lambda$. If we say medium 1 is vacuum, such that $n_1 = 1$, we can relate the index of refraction to the *change in wavelength when entering a medium*:

$$n = \frac{\lambda_0}{\lambda_n} \quad (9.8)$$

where λ_0 is the wavelength of light in vacuum, and λ_n is the wavelength of light in the medium whose refractive index is n . In the end, our most important conclusion is the following. When light leaves one medium, of refractive index n_1 , and enters another, of refractive index n_2 , then:

Law of Refraction (Snell's Law):

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (9.9)$$

where the angles $\theta_{1,2}$ are measured with respect to a line normal to the boundary.

9.3.2 Dispersion and Prisms

Table 9.1 listed the values of the index of refraction n for various materials, at a particular frequency $f = 5.09 \times 10^{14}$ Hz. Why just at $f = 5.09 \times 10^{14}$ Hz? As it turns out, the index of refraction for most anything other than vacuum depends on frequency, as shown in Fig. 9.8. This phenomena is called **dispersion**.

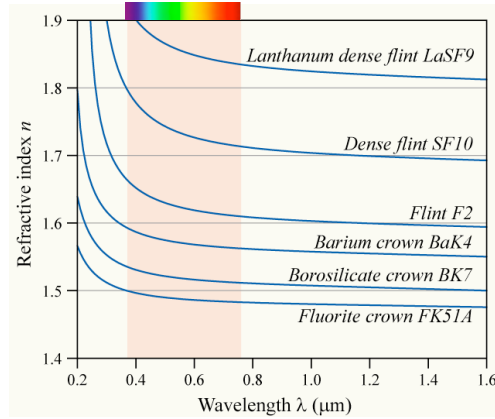


Fig. 9.8 Wavelength dependence of the index of refraction for different types of common glass. The shaded region is the visible spectrum - violet at the smallest wavelengths ($\sim 0.4\mu\text{m}$), and red at the largest ($\sim 0.75\mu\text{m}$).[23]

Dispersion leads to a number of interesting effects. Since n is actually a function of wavelength, $n(\lambda)$, the Law of Refraction (Eq. 9.5) tells us that the angle of refraction depends on the wavelength of the light. For example, violet light ($\lambda \approx 400\text{nm}$) bends much more than red light ($\lambda \approx 650\text{nm}$) when going from air into glass.

The phenomena of dispersion is nicely illustrated by prism, as shown in Fig. 9.9. A light ray of a single wavelength will pass through the prism, but it will be slightly bent on entering and leaving the prism, Fig. 9.9a. The angle of the exit ray compared to the incident ray is known as the **angle of deviation** δ .

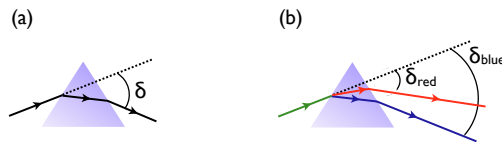


Fig. 9.9 Dispersion of light by a prism. (a) A prism refracts an entering light ray both on entering and leaving. (b) Due to the wavelength variation of n , blue light is bent more than red. (c) White light entering a prism is therefore dispersed by wavelength, yielding a spectrum. Violet light is bent the most, red light the least.

If we shine a ray of *white light* on the prism, something interesting happens. White light is nothing more than a combination of all the visible colors of light in equal proportions. When the white light passes through the prism, the blue light will be bent more than the red ones, and the colors of light become spatially separated, Fig. 9.9b. The result is a display of all the colors of the visible spectrum, Fig. 9.9c. The colors, in order of decreasing wavelength, are red, orange, yellow, green, blue, and violet (Roy G. Biv in mnemonic form). Violet light deviates the most, red the least, and the rest fall in between. Of course, other non-visible wavelengths of light are bent too - ultraviolet (UV) rays would be bent still more than violet, and infrared even less than red - we just can't see them with the naked eye.

9.3.2.1 White Light

Before Newton's studies of optics, most scientists believed that white was the true color of light, and other colors were formed only by adding something to it. Newton demonstrated this was not true by the use of prisms. His experiment was to pass white light through a prism, then direct the individual colored beams through another prism. If light were really white, and the colors were just added by the prism, the second prism should have added further colors to the single-colored beams. Since the single-colored beam remained a single color, Newton concluded that the prism actually separated

the colors already present in the light. White light is the effect of combining the visible colors of light in equal proportions.²

9.3.3 Rainbows

Rainbows are a natural form of light dispersion, in which water droplets in the atmosphere act as tiny prisms. A ray of white light in the atmosphere strikes a (quasi-circular) water droplet, and is refracted and reflected as shown in Fig. 9.10.

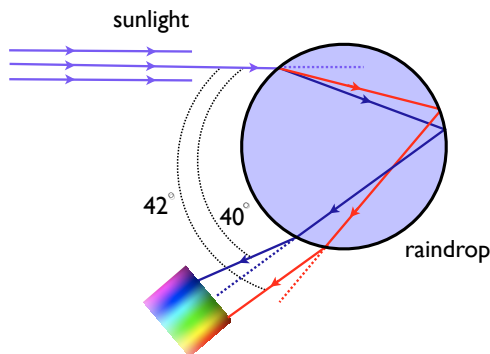


Fig. 9.10 Refraction of sunlight by quasi-spherical raindrops disperses light by wavelength, and results in a rainbow. Note that the drawing is not to scale.

When the light ray reaches the drop, it is first refracted at the front surface of the drop, which causes dispersion. Violet light deviates the most, red light the least. The separated rays then reflect off of the back surface (all wavelengths *reflect* at the same angle), and again reach the back surface of the drop. The individual rays undergo refraction as they leave the drop, and overall the red rays are bent by about 42° relative to the incident rays, and the violet are bent by about 40° . The dispersion is small, but this is what results in a rainbow - incoming sunlight is reflected back over a range of angles.

Incidentally, the light at the back of the raindrop does not undergo total internal reflection, as you might think, and some light does emerge from the back. This transmitted light doesn't create a rainbow, all the

How to we end up observing a rainbow from this small dispersion? Under the right conditions, rain drops are present in the atmosphere. A raindrop high in the sky appears red, because red light is deviated the most and actually reaches the observer. Other colors of light pass over the observer's head. For slightly lower drops, only the yellow rays are deflected at just the right angle to reach the observer. Finally, the lowest observable drops direct violet light to the observer, and disperse other wavelengths below the observer - the red light would just strike the ground and not be observed. So when we observe a rainbow, we are really seeing the $\approx 2^\circ$ angular dispersion created by tiny water drops in the sky.

As it turns out, the rainbow formation process is independent of how big the drops are, but does depend on the refractive index of the drops. For instance, seawater has a higher refractive index than rain water, so the radius of a rainbow in sea spray is smaller. If you observe a rainbow on a rainy near sea spray, this effect is visible as a misalignment of the 'rain' and 'sea' bows. A good example of this, and many other interesting atmospheric optical phenomena, can be found here: <http://www.atoptics.co.uk/>.

Of course, rainbows don't actually exist at some point in the sky, they are an optical phenomena whose apparent position depends on the observer's location and the sun's position in the sky. All

² en.wikipedia.org/wiki/White

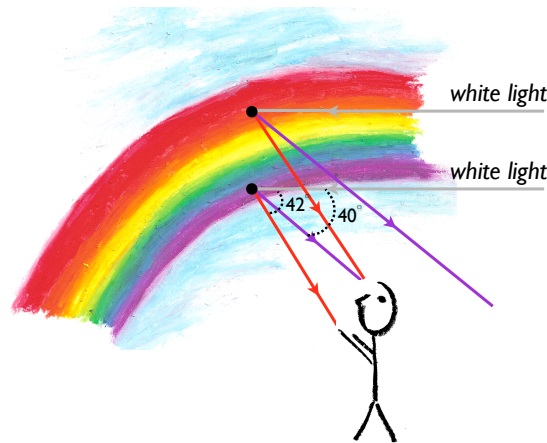


Fig. 9.11 Illustration of rainbow formation. Drawings by Christine LeClair.

raindrops reflect and refract light in the same way, but the only under the right circumstances to the dispersed rays reach the observer's eye. The position of a rainbow in the sky is always in the opposite direction of the Sun for the observer - that is, you need the sun at your back. The bow will be centered on the shadow of your head, and appears at an angle of $40\text{--}42^\circ$ above the line between your head and its shadow. As a result, if the sun is higher than 42° in the sky, the rainbow would be formed below the horizon, and would not be visible. This is not strictly true if you are high above the ground, however.

Figure 9.12 shows a “double” rainbow - a primary rainbow, along with a weaker secondary with its colors reversed. Secondary rainbows are caused by sunlight reflecting *twice* inside the raindrops, instead of just once, and the dispersion is roughly twice as large, appearing at $\approx 50\text{--}53^\circ$. As a result of the second reflection, the colors of a secondary rainbow are reversed compared to the primary. The dark area of unlit sky between the primary and secondary bows is called Alexander's band, after Alexander of Aphrodisias who first described it.



Fig. 9.12 This photograph shows a clear secondary rainbow, observed near Santa Fe, NM in 2006. Note that the colors are reversed in the secondary rainbow. Photo: Donna Wilson, used with permission.

9.4 Total Internal Reflection

When light encounters a boundary between a medium of *higher* index of refraction and one with a *lower* index of refraction, a phenomena called “**total internal reflection**” can occur. For certain angles of incidence greater than some critical angle θ_c , the refracted beam actually does not enter the second medium - it is entirely reflected at the boundary. This is shown schematically in Fig. 9.13.

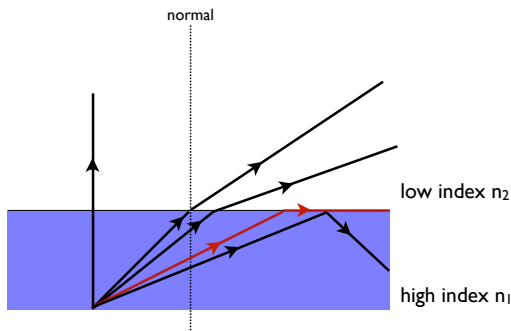


Fig. 9.13 Rays from a medium of index of refraction n_1 incident on the boundary with a medium of lower index of refraction n_2 . As the angle of incidence increases, the refraction angle increases until a critical point (red rays). At that critical incidence angle θ_c , the angle of refraction is 90° , and the light does not enter medium 2. For larger angles, total internal reflection occurs.

The critical incidence angle is when the refracted beam would make an angle of 90° with the normal, which we can find using Eq. 9.5:

$$n_1 \sin \theta_c = n_2 \sin 90^\circ = n_2 \quad (9.10)$$

We can easily solve for the *critical angle of incidence*, above which total internal reflection occurs:

Critical angle for total internal reflection:

$$\sin \theta_c = \frac{n_2}{n_1} \quad \text{for} \quad n_1 \geq n_2 \quad (9.11)$$

This is only valid when $n_1 > n_2$, because *total internal reflection only occurs when light tries to move from a medium of higher refractive index to one of lower refractive index*. If $n_1 < n_2$, the formula would give $\sin \theta_c > 1$, which is impossible. In that case, the math tells us that what we are proposing is physically not possible.

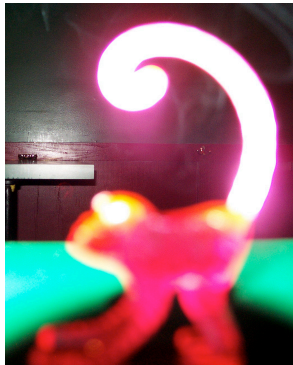


Fig. 9.14 Total internal reflection in the tail of a plastic monkey. Photo by the author.

Internal reflection in prisms is a very useful way to “guide” light to where you want it, as in a periscope. Figure 9.15 shows a few uses of prisms.

9.4.1 Fiber optics

Total internal reflection is the principle on which fiber optic technology is based. An optical fiber is a glass or plastic fiber designed to guide light along its length by total internal reflection. The fiber

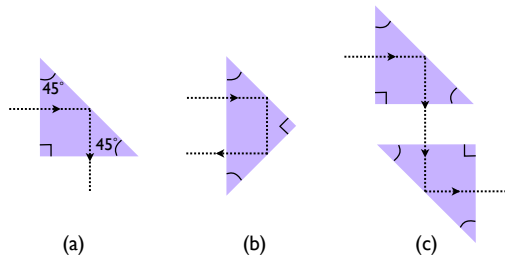


Fig. 9.15 Internal reflection in 45° - 45° - 90° prisms. (a) Changing a ray's direction by 90° , (b) reversing the ray's direction, (c) translating the ray - a periscope.

consists of a core material, surrounded by a cladding layer, as shown in Fig. 9.16. The light (say, an optical signal) is confined in the core due to the fact that the refractive index of the core is higher than that of the cladding. This is why some refer to optical fibers as “light pipes.”

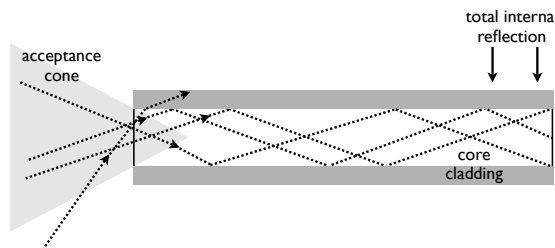


Fig. 9.16 Light propagating through an optical fiber. The cladding material has a lower refractive index than the core material, so within a range of angles (acceptance cone) incident light is confined within the fiber by total internal reflection.

Optical fibers are widely used in communications, as they allow digital data transmission over longer distances and at higher rates than most other forms of wired and wireless communications. They are also used to form sensors, and in a variety of other applications (including those tacky table-top Christmas trees that light up).

9.4.2 Multi-Touch screens

“Frustrated” total internal reflection allows multiple-touch sensing for advanced displays. By frustrated, we mean that the condition for total internal reflection is broken by the presence of a third medium, as shown in Fig. 9.17.

In a (simplified) touch-screen based on total internal reflection, a light emitting diode shines light into an acrylic plane. The light is confined by total internal reflection, due to the differing values of n for acrylic ($n \sim 1.5$) and air ($n \sim 1$). When a user touches the screen with a non-transparent finger, there is no longer total internal reflection. Light is scattered away at non-critical angles by the finger, and can be detected.

The advantages of this scheme over others (e.g., capacitive, resistive, infrared ...) is that it easily senses more than one touch at a time in touch screens or touch tablets / touchpads. One can recognize multiple simultaneous touch points, including the pressure or degree of each, as well as position

of each touch point. This allows gestures and interaction with multiple fingers or hands - such as zooming or chording.

This is the technology will power the display of the impending iPhone.TM Some interesting images and movies can be found here: <http://cs.nyu.edu/~jhan/ftirtouch/>.

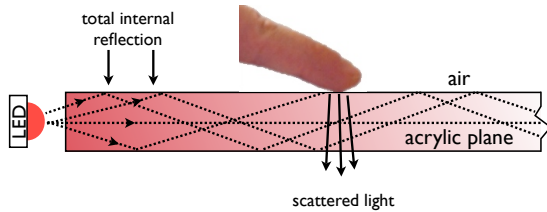


Fig. 9.17 Frustrated total internal reflection allows multiple-touch sensing for advanced displays. A light emitting diode shines light into an acrylic plane, where it is confined by total internal reflection due to the differing values of n for acrylic ($n \sim 1.5$) and air ($n \sim 1$). When a user touches the screen with a non-transparent finger, there is no longer total internal reflection. Light is scattered away at non-critical angles by the finger, and can be detected.

Chapter 10

Mirrors

Abstract The behavior of reflected light within the ray approximation follows from one simple principle – the angle of incidence is equal to the angle of reflection. Everything else we need to know about reflected light just boils down to plane geometry – so far as the physics goes, reflection is from our point of view a solved problem! Nonetheless, we can use the law of reflection along with some carefully applied geometry to derive the behavior of reflected light for a number of important and often-encountered cases.

In this chapter, we will deal with the perfect reflection of light from mirrors. Given an object and a particular sort of mirror, we will learn how to deduce what the nature of the image formed by the mirror will be. If we can first learn how to do this for a single point source of light, we can then build up any more complicated object out many point sources. Our most important example mirrors will be a simple flat mirror, a convex spherical mirror, and a concave spherical mirror. In passing, we will also investigate other technologically important geometries, such as the parabolic reflectors used in satellite dishes.

More broadly, by treating the problem of reflection in various specific geometries, we will begin to learn about the projection, focusing, and manipulation of light. Combined with what we will learn about refraction in lenses in the next chapter, we will be able to understand in detail a great number of optical instruments, such as microscopes, telescopes, and projectors.

10.1 Flat Mirrors

The most simple reflecting object is just a flat mirror, as shown in Fig. 10.1. What happens if we take a point source of light at position O , a distance p in front of the mirror? A point source of light is just what it sounds like – a single point from which light rays leave radially in straight lines. When the light rays exiting the source (blue) reach the surface of the mirror, we apply the law of reflection to determine where the reflected rays go (orange). Only a few of the rays leaving the source are drawn here.

10.1.1 Image formation

Some rays leaving the point source source are reflected off of the surface of the mirror, and reach an observer. The rays reflected off of the mirror in this case appear to come from a point I behind the mirror, if we extrapolate where these diverging rays *appear* to come from (dotted orange lines).¹ Any time we have an intersection of light rays, or a point where light rays appear to originate from, an **image** of the object which was the source of the rays is formed. From the observer's point of

¹ Since this is not a real light ray anyway, we do not worry about refraction in the glass making up the mirror. We further assume the mirrors to be negligibly thin in any case.

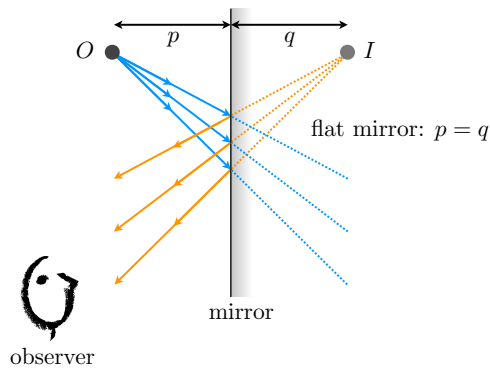


Fig. 10.1 Reflection from a flat mirror. An image is formed by light rays from an object reflecting off of the mirror's surface. The object is located at O , a distance p from the mirror, while the image location I is behind the mirror at a distance q . For a *flat* mirror, $p = q$. Solid lines indicate actual light rays, dotted lines indicate 'virtual' rays, whose apparent point of convergence determines where the observer sees the image.

view, the rays reflected off of the source object at O appear to come from a point I behind the mirror, so we would say that the view sees an *image* of the object at point I , a distance q behind the mirror.

Remember, for reflection and refraction, we have to be able to run the rays forwards or backwards and get the same result. If we trace the light rays from the object to the observers eyes, this is of course the real path the rays take. Tracing the orange rays backward through the mirror to find their point of convergence tells us *where we would need a second point source to reproduce the image observed*. All real and virtual light rays fall into two categories – ones that converge onto a point (either the image or the object), and ones that diverge.

Image formation:

Images are formed where light rays converge to a point (intersect), or where they *appear* to originate from.

If the original point source is a distance p from the mirror, straightforward geometry tells us that the image distance q must be the same, $p = q$. The image observed is exactly as far behind the mirror as the object is in front of it. The image in this case is what is known as a *virtual image* – light doesn't actually pass through the point where the image is created, but only *appears* to come from that point. A *real image* is formed when light actually passes through some point. Real images can be projected onto a screen, for example, since they result from real light sources, while virtual images cannot (hence the term "virtual").

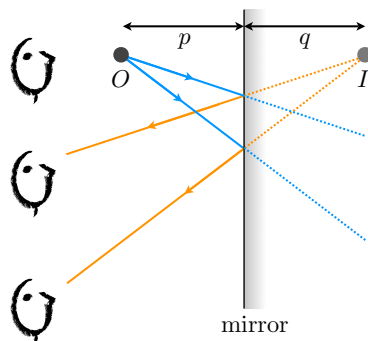


Fig. 10.2 Different observers see the same image from a flat mirror. Even though the three observers are at different positions, geometry tells us that they will all see the image formed at the same location.

Virtual image: Light rays don't actually pass through an image point, but appear to originate from there.

Real image: Light rays actually pass through a point. Only real images can be projected onto a screen.

Our flat mirror forms a virtual image, since the image an observer sees is behind the mirror, and does not result from real light rays coming from the point of the image. The virtual image is just where the actual object *appears* to be after the mirror reflects light rays coming from it. Images from flat mirrors are *always* virtual. Can we determine anything else about the image? Is the image of the same size and shape as the object? Can we more rigorously prove our assertion that $p = q$. Sure. How do we deal with more complicated objects, as opposed to simple point sources?

10.1.2 Ray Diagrams

If we know how to handle single light rays and point sources, we can handle any more complicated object by *building it out of point sources*. We can consider any object to be made up of a series of points (or pixels, if you like), and trace the light rays from each point on the object. Usually it is not necessary to trace rays from *every* point on the object, it is enough to trace rays from a few crucial points and fill in the blanks by symmetry and common sense. As an example, consider the upright blue arrow in front of a flat mirror in Fig. 10.3. Our usual example object will be an arrow, since it is a simple shape that lets us easily determine whether images formed are inverted or magnified. As we shall see, another advantage is that all we need to do is trace out the rays from the very tip of the arrow, and the rest fills in naturally.

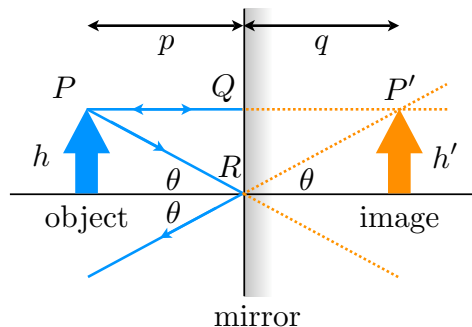


Fig. 10.3 The location and size of a reflected image from a flat mirror can be found with a simple geometric construction. Trace one ray from the object perpendicular to the mirror's surface and one ray from the object through the origin. Real light rays reflect off of the mirror, virtual light rays continue on through the mirror. The convergence of virtual rays behind the mirror gives the image location. Since triangles PQR and $P'QR$ are identical, the image and object heights are equal, $h = h'$, as are the image and object distances, $p = |q|$.

We place the arrow of height h at point P , a distance p from the mirror. Simple geometric techniques will let us figure out exactly what the image is like. First, we trace a ray outward from the tip of the arrow which intersects the mirror at a perfect 90° , intersecting the mirror at point Q . This ray will just be reflected right back – if the angle of incidence is 90° , then so must be the angle of reflection. For an observer sitting directly behind the object, this ray would *appear* to come from behind the mirror, so we continue tracing a virtual ray (dotted orange line) behind the mirror.

Now, we need to trace at least one more ray to uniquely determine what the image looks like. We need to find an intersection of real or virtual rays in order to have an image, so we have to have at least two, and in general three is safer. For the second ray, we will trace a line from the tip of the arrow to a point on the mirror at the same vertical position as the bottom of the arrow. The use of two extremal rays gives us more confidence in the position of the resulting image – if two such extreme rays find an intersecting point, we are fairly sure we have found the image location. If we chose two rays at similar angles, small inaccuracies in our drawing become more important, and we have a harder time discerning the image position and size with any accuracy. Try tracing some ray diagrams for yourself, you will quickly find this to be true.

This second ray is reflected downward from point R on the mirror at the same angle θ at which it impinges on the mirror. Extrapolating the reflected ray back through the mirror as a virtual ray (dotted orange line), we see that it converges with the first virtual ray at point P' . This point of convergence, then, must be the location of the image. Furthermore, since we are tracing out rays from the *tip* of the arrow, this must be the tip of the *image's* arrow. Symmetry alone tells us that the image arrow must be upright, like the real one. If you are not convinced, trace out the same two types of rays from the *bottom* of the arrow, and you will see!

We have established, then, that the image is virtual, and upright (not inverted). What about its size? The virtual ray from R to P' , $\overrightarrow{RP'}$ clearly must make an angle θ with the horizontal axis, since it is just a continuation of the reflected ray at point R . The lines \overline{PQ} and $\overline{QP'}$ are horizontal, so the angles $\angle RPQ$ and $\angle RP'Q$ must also be θ , since they are alternate interior angles to the θ drawn in the figure. The triangles $\triangle RPQ$ and $\triangle RP'Q$ must therefore be equivalent, since they share RQ as a side. If these two triangles are equivalent, it clear that $h = h'$, and $p = q$. Now we have proved our assertion that **the image formed by an object placed in front of a flat mirror is as far behind the mirror as the object is in front of it**. We have further proved that **the image is the same size as the object**. The images formed by flat mirrors faithfully reproduce objects.

Flat Mirrors:

1. The image is as far behind the mirror as the object is in front of it.
2. The image is the same size as the object.
3. The image is upright and virtual.

10.1.3 Conventions for Ray Diagrams

For flat mirrors, we now know almost everything we need to. Other types of mirrors will not always give images that are the same size as the object, however, and will not always be the same distance away. If the image is not the same size as the object, we say that it is *magnified*. Magnified can mean either larger *or smaller*. The degree of magnification is nothing more than the ratio of the image height to the object height – how much larger or smaller is the image compared to the object?

Lateral Magnification of a Mirror:

$$M \equiv \frac{\text{image height}}{\text{object height}} \equiv \frac{h'}{h} \quad (10.1)$$

where h is the object height and h' the image height. For a flat mirror, $M = 1$.

For future convenience, we should also lay down some conventions for our ray diagrams. First, we will always treat the mirror as the ‘zero’ for our horizontal axis. Distance is positive in front of the mirror, and negative behind it. Real images are formed in front of the mirror, while virtual images are formed behind the mirror (since no light goes through the mirror). The distance from the mirror to the object will always be p , the distance to the image always q . The height of the image will be h , the height of the object h' .

Conventions for Mirror Ray Diagrams:

1. The distance between the object and the mirror is p .
2. The distance between the image and the mirror is q .

3. The object's height is h , the image's height is h' .
4. In front of the mirror, p and q are **positive**.
5. The front of the mirror is where real rays propagate, the back is where virtual rays are formed.
6. Behind the mirror, p and q are **negative**.
7. Real light rays are solid lines, virtual rays are dotted.

10.1.4 Handedness

Before we move on to different mirror geometries, one last word about mirrors and handedness. You may remember that we discussed the difference between left- and right-handed coordinate systems in Sect. 6.1.4. You already know of course that when you look in a mirror your sense of left and right are reversed. If you wave your right hand in the mirror, the image seems to wave its left. Similarly, a mirror reflection is what relates left-handed and right-handed coordinate systems, or right-handed and left-handed corkscrews. Examine Fig. 10.4, and convince yourself once again that there is an intrinsic *handedness* or *chirality* to certain things. Only a mirror reflection can change a left-handed to a right-handed coordinate system, no number of simple rotations will do it.

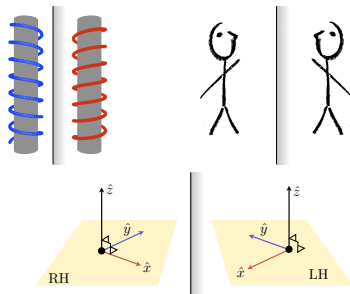


Fig. 10.4 Reflected images have reversed handedness. Clockwise, from upper left: a right-handed corkscrew becomes a left-handed one in reflection, a left hand becomes a right, and more generally a right-handed coordinate system transforms to a left-handed one.

10.2 Spherical Mirrors

Spherical mirrors are just what they sound like: the reflective surface has the shape of an arc of a circle. Spherical mirrors can be uniquely described by the radius of the circle R making up the arc, and whether they are *concave* or *convex*. **Concave** mirrors are made by putting a reflective coating on the *inside* surface of the circle, while **convex** mirrors are made by putting a reflective coating on the *outside* surface of the circle.

10.2.1 Concave Mirrors

An example of a concave mirror is shown in Fig. 10.5a. The point C is the center of curvature of the mirror (the center of the circular arc), and is a distance R from any point on the mirror's surface. The line drawn through the center of curvature C and a point V at the center of the arc defines the **principle axis** of the mirror. How do we figure out what images look like using such a mirror? Just like before, we trace light rays and apply the law of reflection and geometry.

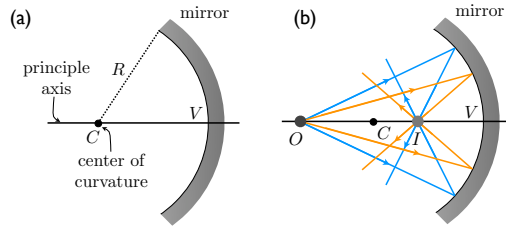


Fig. 10.5 (a) Reflection from a concave spherical mirror. The center of curvature C is the center of the spherical arc of radius R making up the mirror. The principle axis passes through the center of curvature as well as the middle of the mirror, V . (b) If we place an object O anywhere on the principle axis farther away from the mirror than C , a real image is formed at I . If the distance from O to the mirror is relatively large compared to R (such that the rays come off of the principle axis at small angles), all rays reflect through the same point.

Figure 10.5b shows a point source O placed relatively far from a spherical mirror, outside the center of curvature. Rays leaving point O with a sufficiently small angle intersect the mirror, and are all reflected back through a common convergence point I . The point I is the *image point*, and the convergence of rays indicates that an image will form there, as though there were a copy of the source at that point. Since real light rays are passing through the point I , the image formed is *real*.

For spherical mirrors in particular, we will usually assume that the light rays from the source make a small angle with the principle axis. When this condition is met, all incident rays will reflect back through the image point. On the other hand, when some rays reaching the mirror make a relatively large angle with the principle axis – when the object is relatively close to the spherical mirror – this is no longer true, as shown in Fig. 10.6. When the object is too close to the mirror, some of the rays making a large angle with the principle axis no longer reflect back through the image point, and no single point of convergence exists. This means that the image formed is not clearly focused on one point, but spread out – the image is blurry. This phenomena is known as *spherical aberration*. It is quite important for, *e.g.*, telescopes and cameras – since spherical shapes the easiest to produce, most lenses have spherical shapes and will suffer from this phenomena, as we will see in more detail in the following chapter.

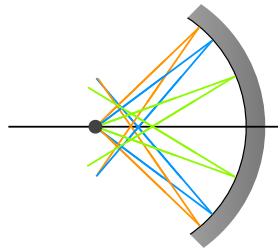


Fig. 10.6 Rays at large angles from the principle axis do not all reflect back to intersect the principle axis at the same point. As a result, when objects are too close to a spherical mirror, the image formed is “fuzzy” since the convergence of rays is now spread out. This effect is known as spherical aberration.

If we ensure that the object is sufficiently far from the mirror to avoid spherical aberration, what will the image look like? Just like with flat mirrors, we will trace the rays coming from the tip of an arrow placed in front of the mirror, as shown in Fig. 10.7. Again the arrow of height h is placed a distance p from the mirror, at point O . The center of curvature for the mirror is C , and the center of the mirror is at V .

First, we trace a ray from the tip of the arrow through the center of curvature at C . Since the mirror is the arc of a circle, any line passing through the center of curvature must be normal to the surface of the arc – that is, it must intersect the surface of the arc at a 90° angle. Therefore, the ray drawn through the center of curvature reflects back along the same path. We will call the angle this ray makes with the principle axis α .

Next, we draw a second ray from the tip of the arrow through the center of the mirror at V . This ray makes an angle θ with the principle axis, and will reflect off the mirror at V with the same angle. This ray intersects the first at the point I , and defines the tip of the image arrow. Since the intersection point lies below the principle axis, *the image is inverted*. Further, we can already see that it is not the same size as the original arrow, so the image is also *magnified*. Finally, it is real light rays that are intersecting in front of the mirror, so the image formed is real.

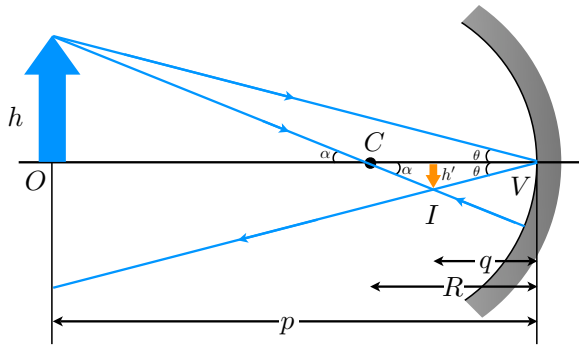


Fig. 10.7 The image formed by a spherical concave mirror for objects placed outside of the center of curvature C. The image is real, magnified, and inverted.

Concave spherical mirrors:

Images are *real*, *inverted*, and *magnified*.

Still, it would be nice to know *exactly* how big the image is, and where it is. This much we can figure out with a bit of geometry. First, we can use the two θ angles and relate the object height h and the image height h' . From the triangle formed by the object arrow and the uppermost ray:

$$\tan \theta = \frac{h}{p} \quad (10.2)$$

Similarly, from the triangle formed by the reflection of that ray and the image arrow:

$$\tan \theta = \frac{-h'}{q} \quad (10.3)$$

Note that since the image arrow points downward below the principle axis, the height of the image is negative. Some simple algebra yields the magnification of the mirror:

$$\tan \theta = \frac{h}{p} = \frac{h'}{q} \quad (10.4)$$

$$\Rightarrow M = \frac{h'}{h} = -\frac{q}{p} \quad (10.5)$$

Magnification for a concave spherical mirror:

$$M = \frac{h'}{h} = -\frac{q}{p} \quad (10.6)$$

Here h is the height of the object, h' is the height of the image, p is the object distance, q is the image distance. Negative M means the image is *inverted*.

Assuming we know h and p to begin with, we still need one more equation in order to uniquely determine h' and q , the height and position of the image. For that, we can use the α angles. From the triangle defined by the left-most α and the object,

$$\tan \alpha = \frac{h}{p - R} \quad (10.7)$$

Using the triangle defined by the right-most α and the image,

$$\tan \alpha = -\frac{h'}{R-q} \quad (10.8)$$

We can now use the above equations for $\tan \alpha$ along with Eq. 10.6 to find another useful equation relating p and q alone:

$$\begin{aligned} \tan \alpha &= \frac{h}{p-R} = -\frac{h'}{R-q} \\ \frac{h'}{h} &= -\frac{R-q}{p-R} = -\frac{q}{p} \quad (\text{using Eq. 10.6}) \\ p(R-q) &= q(p-R) \\ pR - pq &= qp - qR \\ pR + qR &= 2qp \\ R(p+q) &= 2qp \\ \frac{R}{2} &= \frac{qp}{p+q} = \frac{1}{\frac{1}{q} + \frac{1}{p}} \\ \frac{2}{R} &= \frac{1}{p} + \frac{1}{q} \end{aligned}$$

This last expression is known as the *mirror equation*, relates the image and object distances to the physical radius of curvature of the mirror alone. As we shall find out shortly, this equation is far more general than our simple derivation of it would imply. Coupled with the expression for magnification, we can now deduce the behavior of any object with any concave spherical mirror ... so long as the object isn't too close to the mirror.

Mirror equation:

$$\frac{2}{R} = \frac{1}{p} + \frac{1}{q} \quad (10.9)$$

where p is the object distance, q is the image distance, and R is the radius of curvature of the mirror.

10.2.1.1 Concave spherical mirrors and distant objects

We have already seen that forming sharp images from a concave spherical mirror requires the object to be relatively far from the mirror (at least outside the radius of curvature). What happens if the object is *really, really* far away? Say, far enough compared to R that p is essentially infinite? When the object is very, very far away, the incident rays are all very nearly parallel to the principle axis. For very distant sources, any small angle away from the principle axis will result in the rays diverging too far to hit the mirror, only those rays at tiny angles relative to the principle axis will hit the mirror. For all intents and purposes, we can assume all rays from a very distant object impinge on the mirror parallel to the principle axis, as shown in Fig. 10.8.

The mirror equation gives us yet more insight. If we let p tend toward infinity, then $1/p$ tends toward zero. In this case, $q \approx R/2$ – the image is formed exactly half way between the center of curvature and the mirror when the object is very far away compared to R . In this special case of a distant object, all the incident rays converge at the same point F (Fig. 10.8), which we call the *focal point* of the mirror. The **focal length** f of a mirror is just the distance between the mirror and the focal point on the principle axis where light from a distant object would converge. Put another way, it is the image distance q when we allow p to tend toward infinity. Thus, for our concave spherical mirror, $f = \frac{R}{2}$.

Though the focal length and radius of curvature are simply related, it is the former that you will hear more often in optics. The focal length of a mirror is where light would focus if we had a point

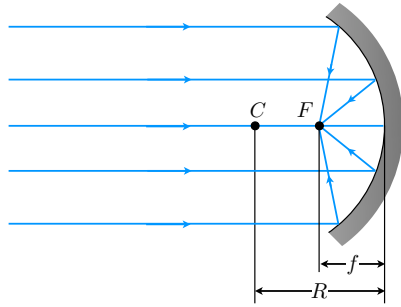


Fig. 10.8 For very distant objects ($p \leftarrow \infty$), incident light rays are essentially parallel, and all reflect through the focal point of the mirror F . For very distant objects, the image distance is $q \approx f \approx R/2$, where f is the focal length of the mirror (the position of F).

source infinitely far away, and is one way of comparing the properties of different mirrors (or lenses, as we shall see). Even though we can't actually realize this situation, we can get far enough away from a mirror to approximate it, and in fact, this is the regime in which we try to operate most optical instruments. If you have any experience with photography, you are no doubt already familiar with focal lengths. In any case: the focal length is a characteristic of a spherical mirror, just half its radius of curvature, and it allows us to re-write the mirror equation in an ostensibly more useful way:

Mirror equation in terms of the focal length:

$$\frac{1}{f} = \frac{1}{p} + \frac{1}{q} \quad (10.10)$$

where p is the image distance, q is the object distance, and f is the focal length. For a concave spherical mirror, the focal length is half the radius of curvature, $2f = R$.

The fact that spherical mirrors focus all distant light onto a single point makes them potentially useful for, *e.g.*, solar heating or focusing antennas. As we shall see in subsequent sections, however, there is a still more clever geometry which is much better for light harvesting applications.

10.2.2 Convex Spherical Mirrors

A convex spherical mirror is shown in Fig. 10.9, in which the *outer* surface of the spherical arc has a reflective coating. While a concave spherical mirror tends to focus distant light on to a single point, a convex spherical mirror tends to *diverge* incident rays. Nearly all incident rays on the surface of the convex spherical mirror diverge after reflection, as if they are coming from *behind* the mirror itself. Analyzing the image formed by this type of mirror is not much more difficult than the other cases we have dealt with, we just have to construct a ray diagram.

For the moment, two rays are enough to grasp the nature of image formation for a convex mirror. First, we draw a ray horizontally from the tip of our object arrow in Fig. 10.9. This ray is reflected upward away from the object and mirror. If we trace the reflected ray backward through the mirror, it intersects the principle axis exactly at the focal point of the mirror. Next, we draw a ray from the tip of the arrow through the center of curvature of the mirror. In front of the mirror, it is a real ray, while in back of the mirror it is a virtual ray. The intersection of our two virtual rays behind the mirror gives the image location.

In this case, we can see that the image is *upright*, *virtual*, and *magnified*. What is the actual image position and magnification factor? As it turns out, if we work through the geometry, **the same mirror equation is valid for convex spherical mirrors, if we keep in mind that p and q are negative when we are behind the mirror.** In this particular case for convex spherical mirrors, h and h' are positive, p is positive, and q is negative. Table 10.1 is a reminder of the sign conventions we

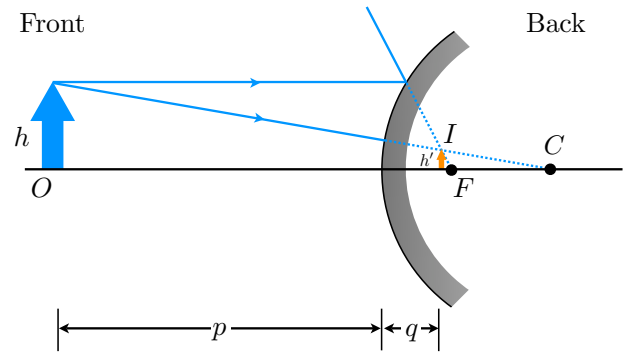


Fig. 10.9 The image formed by a spherical convex mirror is virtual, magnified, and upright.

Table 10.1 Sign Conventions for Mirrors

Quantity	Symbol	Front	Back	Upright	Inverted
Object location	p	+	−		
Image location	q	+	−		
Focal length	f	+	−		
Object height	h			+	−
Image height	h'			+	−
Magnification	M			+	−

use for mirrors. Parenthetically, we note that the mirror equation also works for flat mirrors! The radius of curvature of a flat plane is infinite, and applying this to Eq. 10.9 readily gives $p = q$.

10.3 Ray Diagrams for Mirrors

So far, we have constructed *ad hoc* ray diagrams for the different mirrors under consideration. The ray diagrams are nothing more than graphical constructions to give us an overall impression of the image formed. We tried to choose rays that gave extremal cases, in the hopes that this would give a more accurate image. In fact, we can come up with a set of general rules for constructing a ray diagram for any simple mirror, so long as we know the object location and the mirror’s center of curvature. In the end, we need only three rays. So far we have used only two, and that has worked fine. In some sense the third ray is a ‘sanity check.’ With only two rays, it is almost certain that we will have an intersection *somewhere*, even if make some small mistakes in our ray tracing. The odds of a third ray spuriously intersecting the other two at the same point is *tiny*, so if all three rays intersect at the same point, we can be sure that our diagram is reasonably correct.

How to construct ray diagrams:

- Ray 1** is drawn parallel to the principle axis, and reflects back through the focal point.
- Ray 2** is drawn through the focal point, and reflects back parallel to the principle axis.
- Ray 3** is drawn through the center of curvature, and reflects back on itself.

In using these rules and analyzing different situations for spherical mirrors, we can make the some generalizations to serve as rules-of-thumb:

Images from Spherical Mirrors:

1. Concave Mirrors (Fig. 10.10):

- a. $p > R$: object *outside* center of curvature, gives a real, inverted, and reduced image
 - b. $R > p > f$: object *outside* focal point and *inside* center of curvature, gives a real, inverted, enlarged image
 - c. $p < f$: object inside focal length gives virtual, upright image
2. Convex Mirrors:
- a. image is always virtual and upright

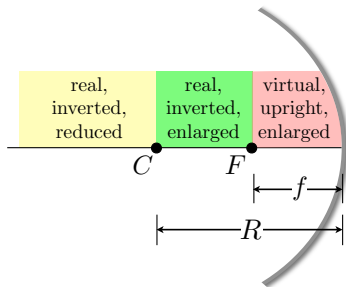


Fig. 10.10 The type of image formed by a spherical mirror depends on the location of the object relative to the center of curvature and the focus of the mirror. For objects outside the center of curvature, the image is real, inverted, and reduced. For objects between the center of curvature and focus, the image is real, inverted, and enlarged. For objects inside the focus, the images are virtual, upright, and enlarged.

Figure 10.11 shows these three rules applied to concave and convex spherical mirrors. The first rule just follows from our discussion of very distant rays incident on a spherical mirror – the *definition* of the focal point is the point at which rays parallel to the principle axis reflect through (virtual rays in the case of convex mirrors). The second rule follows in the same way. The third rule is essentially the definition of the radius of curvature – any line passing through the radius of curvature is incident normal on the surface of the mirror, and must reflect back on itself.

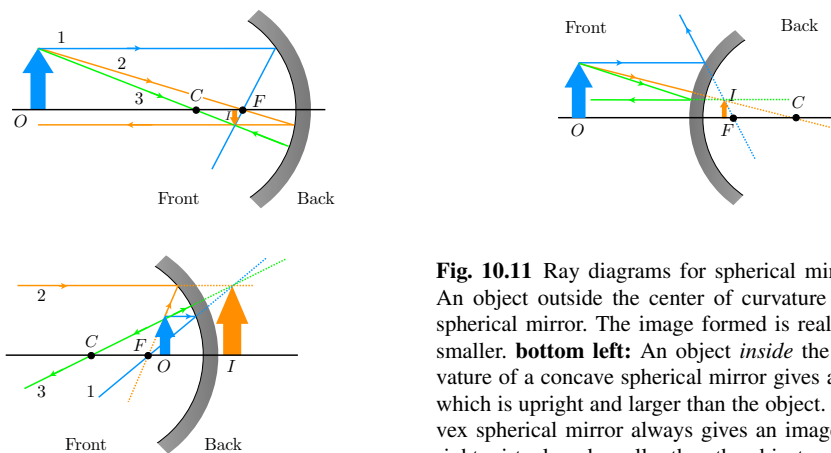


Fig. 10.11 Ray diagrams for spherical mirrors. **top left:** An object outside the center of curvature for a concave spherical mirror. The image formed is real, inverted, and smaller. **bottom left:** An object *inside* the center of curvature of a concave spherical mirror gives a virtual image which is upright and larger than the object. **above:** A convex spherical mirror always gives an image which is upright, virtual, and smaller than the object.

10.4 Parabolic Mirrors

Circular mirrors are just fine, but isn't there something more efficient? Is there a shape of mirror we could make such that *all* distant rays are focused onto a single point, not just those close to the

central axis? Indeed, there is just such a curve, and you are already familiar with it: the parabola. In fact, the parabola is unique in this regard. It is the only curve such that all incident parallel rays will be reflected and focused on to a single point, the *focus* of the parabola.

This is illustrated in Fig. 10.12. If a series of parallel rays is incident downward on the parabola, they will all converge at the focus F . Equivalently, since we can always run our ray diagrams ‘forward’ or ‘backward,’ a point source of light placed at F will produce a parallel beam of light. Incidentally, this works in three dimensions too. A circular paraboloid, made by rotating a parabola about its axis, is the only 3D surface for which all rays parallel to a given ray pass through the same point after reflection by the surface. What good is this property? Well, this is how modern car headlights use a single bulb to produce a beam of light, and it is how satellite antennas (‘dishes’) manage to focus an extremely tiny amount of radiation into a usable signal. Make the parabola as large as possible, collecting radiation from as large an area as possible, and it all gets focused to a single point, enormously amplifying the intensity. The same principle is used for radio astronomy and solar ovens.

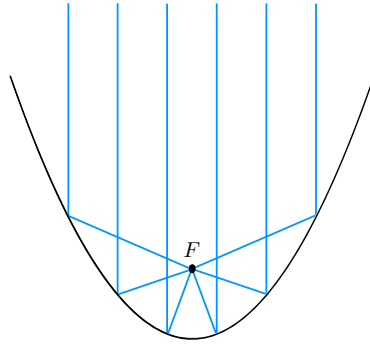


Fig. 10.12 Focusing of light by a parabolic mirror. A distant light source providing incident rays which are parallel will be reflected by the parabola and focused onto a single point F . Conversely, a point source located at the focus F will produce a beam of parallel rays. Parabolic mirrors offer some advantages over spherical mirrors for focusing – the parallel rays can come in at an angle to the parabola and still be focused, and spherical aberrations can be significantly reduced.

How does this work? Geometrically, a parabola is a conic section defined as the locus of points equidistant from a single point (the focus) and a straight line (the directrix). This is shown in Fig. 10.13. Without loss of generality, we will take the parabola centered on the origin of an $x - y$ coordinate system. Let the focus F be at the point $(0, f)$, and the directrix be the line $y = -f$. This is still perfectly general - an arbitrary point and line, since we can make f whatever we want. Our parabola is ‘between’ the focus and directrix.

Construct a line connecting F with an arbitrary point $P(x_0, y_0)$ on the parabola, and a vertical line intersecting the directrix at point $D(x_0, -f)$. A parabola is, as stated above, geometrically defined as the locus of all points for which $\overline{FP} = \overline{PD}$. If we didn’t already know that, could we figure out what curve satisfies this relationship? We can, simply calculate the lengths \overline{FP} and \overline{PD} with the distance formula:

$$\begin{aligned}\overline{FP} &= \overline{PD} \\ \sqrt{(x_0 - 0)^2 + (y_0 - f)^2} &= \sqrt{(x_0 - x_0)^2 + (y_0 + f)^2} \\ x_0^2 + y_0^2 - 2fy_0 + f^2 &= y_0^2 + 2fy_0 + f^2 \\ x_0^2 &= 4fy_0 \\ y_0 &= \frac{1}{4f}x_0^2\end{aligned}$$

Lo and behold, the curve is a parabola. One can easily repeat this calculation for a parabola centered on an arbitrary point, the same conclusion holds: a parabola is the only curve for which all points are equidistant from a single line and a single point. For a parabola centered on (x_0, y_0) symmetric about the y axis (*i.e.*, pointing upward or downward), one finds $(y - y_0) = \frac{1}{4f}(x - x_0)^2$.

So what? Now we can sketch a proof of the unique focal property of the parabola as well, using the second portion of Fig. 10.13. If we can prove that a tangent line to the parabola at point P will

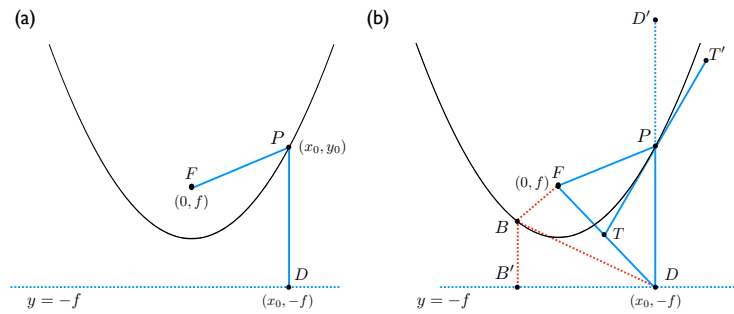


Fig. 10.13 *left*: Construction of a parabola. A parabola is the locus of points equidistant from the focus $F(0, f)$ and the directrix line $y = -f$. *right*: Any ray directed along the parabola's axis of symmetry is reflected and passes through the focus.

make equal angles with \overline{PF} and \overline{PD} , this is enough to prove the focal property. First, we must figure out how to construct a tangent to the parabola at any point.²

By definition, triangle $\triangle FPD$ is isosceles - for a parabola, \overline{PF} and \overline{PD} are equal. Let point T be the midpoint of the line connecting F and D , \overline{FD} . Now the triangles $\triangle FPT$ and $\triangle TPD$ have two equal sides, since $\overline{FP} = \overline{PD}$ and by construction $\overline{FT} = \overline{TD}$. The perpendicular bisector \overline{FD} divides the $x - y$ plane into two sections: all points which are nearer to F than to D , and all points that are nearer to D than to F . Except for point P , every point on the parabola itself lies closer to F than to D by virtue of being above the line \overline{PT} .

Let B be any other point on the parabola, and B' the point nearest to it lying on the directrix. The line segment $\overline{BB'}$ is the shortest possible segment connecting the point B on the parabola to the directrix. The segment $\overline{BB'}$ must be vertical and perpendicular to the directrix for this to be true. By construction, then, $\overline{BB'} = \overline{FB} < \overline{BD}$ - a vertical line segment from B to the directrix must be the same length as the line segment from B to F . Since $\overline{BB'}$ is the shortest distance from B to the directrix, it must be shorter than \overline{BD} . If this is true, then \overline{PT} can not pass through B , or it would be closer to the directrix than the focus, a contradiction. Thus P is the only point of intersection of the line \overline{PT} and the parabola. Thus, \overline{PT} must be tangent to the parabola at point P .

Whew! Now, if \overline{PT} is tangent to the parabola at P , the angles $\angle FPT$ and $\angle TPD$ must be equal. Further, $\angle TPD$ is equal to angle $\angle D'PT'$. If we imagine $\overline{D'P}$ to be a light ray incident on a parabolic surface reflected toward F , this establishes that the incident and reflected angles are equal. Since the point P was completely arbitrary, this means that *any* incident vertical ray must be reflected through the focus F , and that any light originating at F will be reflected as a vertical ray.

Other conic sections have reflective properties similar to the parabola. For instance, if a light source is placed at one focus of an ellipse, the rays will converge onto the other focus after being reflected. Any wave, including sound waves, may be substituted for light. A nice trick is to make an elliptically-shaped room, known as a 'whispering gallery.' If a sound is created at one focus - even a very quiet one - it will be heard clearly at the second focus. It is a dramatic demonstration. You can stand at one focus and whisper so quietly someone standing next to you cannot hear, and yet be clearly heard at the other focus. Some famous examples of rooms like this are listed in the Wikipedia: http://en.wikipedia.org/wiki/Whispering_gallery.

² Many of you probably realize how much easier this task would be with a bit of calculus - in fact, it is a trivial problem if we use calculus. The geometric problem is not trivial, but worth working through if for no other reason to emphasize the fact that parabolas are simple *geometric* constructions, not just abstract quadratic equations. In our studies of optics, good geometrical insight will serve you well.

Chapter 11

Lenses

Abstract Lenses ...

11.1 Spherical refracting surfaces

In order to start discussing lenses *quantitatively*, it is useful to consider a simple spherical surface, as shown in Fig. 11.1. Our ‘lens’ is a semi-infinite rod with one spherical surface, made of a material of refractive index n_2 greater than the surrounding material ($n_1 < n_2$). Qualitatively, we know what will happen based on the law of refraction. Rays emanating from a distant object placed at O will impinge on the spherical surface, bend toward the principle axis (toward the surface normal), and converge at a point inside the rod, forming a real image. But where?

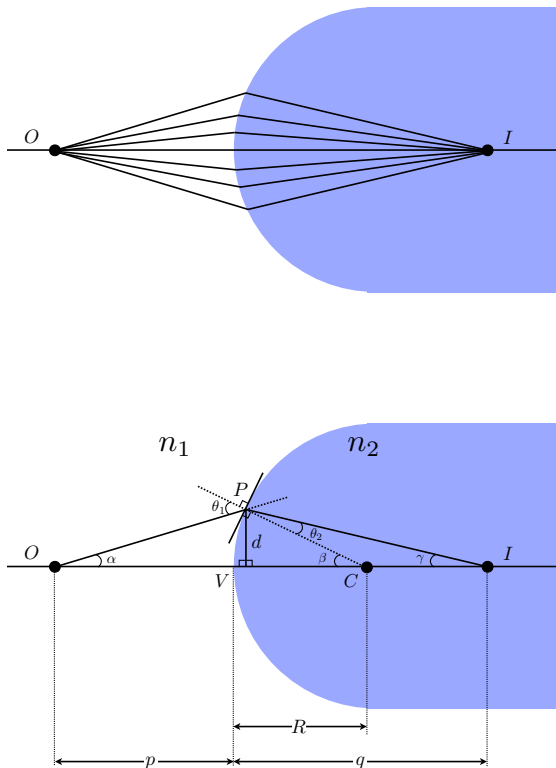


Fig. 11.1 A spherical refracting surface. **upper:** Rays incident from a distant object O are refracted toward the principle axis, and focused at a point I . **lower:** Construction for determining the relative image and object distances in terms of the radius of curvature and refractive indices.

With a bit of geometry, we can figure out exactly where the image must form, given the object distance and the radius of the spherical surface. Referring to the second portion of Fig. 11.1, let the object (O) and image (I) distances be p and q , respectively, measured from the intersection of the principle axis with the spherical surface (V). The center of the sphere of radius R making up the surface is at C . Trivially, a ray drawn from O through the principle axis pass through V , C , and I .

Now, draw a ray leaving the object and intersecting the surface at point P , \overrightarrow{OP} . At the point P , we draw surface normal and tangent lines to define the angle of incidence θ_1 and the angle of refraction θ_2 . The refracted ray will be bent toward the principle axis, intersecting it at point I . This ray \overrightarrow{PI} makes an angle α with the principle axis. Recall that any line perpendicular to the surface of a circle must pass through the center of the circle. Thus, if we extend the normal drawn at point P , it must intersect point C , forming ray \overrightarrow{PC} , which makes an angle β with the principle axis. Now we have everything labeled that we need, “all” that is left is to find a relationship between p , q , and R .

First, we can use right triangle $\triangle OPC$. The angles $\angle OPC$, α , and β making up this triangle must add up to 180° . We also know that the angles θ_1 and $\angle OPC$ by themselves define a straight line, and must therefore add up to 180° as well. Thus:

$$\alpha + \beta + \angle OPC = 180^\circ \quad (11.1)$$

$$\theta_1 + \angle OPC = 180^\circ \quad (11.2)$$

$$\implies \theta_1 = \alpha + \beta \quad (11.3)$$

Slowly, we are reducing the number of unknown quantities. Now examine the triangle $\triangle PCI$. We know that the angles θ_2 , γ , and $\angle PCI$ must add up to 180° . Further, we know that β and $\angle PCI$ must together make 180° , since they define the line \overrightarrow{OI} . Putting these facts together:

$$\theta_2 + \gamma + \angle PCI = 180^\circ \quad (11.4)$$

$$\beta + \angle PCI = 180^\circ \quad (11.5)$$

$$\implies \theta_2 = \beta - \gamma \quad (11.6)$$

Equations 11.3 and 11.6 give us the angles of incidence (θ_1) and refraction (θ_2) in terms of the interior angles α , β , and γ which can be more easily related to the distances of interest, viz., p , q , and R . Before we can do that, we have one trick up our sleeve: we haven’t yet used Snell’s law:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (11.7)$$

If we substitute equations 11.3 and 11.6 into this expression, we have:

$$n_1 \sin(\alpha + \beta) = n_2 \sin(\beta - \gamma) \quad (11.8)$$

We can apply the sum and difference identities for $\sin(a \pm b)$ to this, which yields the following:

$$n_1 \sin \alpha \cos \beta + n_1 \cos \alpha \sin \beta = n_2 \sin \beta \cos \gamma - n_2 \cos \beta \sin \gamma \quad (11.9)$$

$$n_1 \cos \alpha (\tan \alpha \cos \beta + \sin \beta) = n_2 \cos \gamma (\sin \beta - \cos \beta \tan \gamma)$$

$$n_1 \cos \alpha \sin \beta \left(\frac{\tan \alpha}{\tan \beta} + 1 \right) = n_2 \cos \gamma \sin \beta \left(1 - \frac{\tan \gamma}{\tan \beta} \right)$$

$$n_1 \cos \alpha \cancel{\sin \beta} \left(\frac{\tan \alpha}{\tan \beta} + 1 \right) = n_2 \cos \gamma \cancel{\sin \beta} \left(1 - \frac{\tan \gamma}{\tan \beta} \right) \quad (\beta \neq 0)$$

$$n_1 \cos \alpha \left(\frac{\tan \alpha}{\tan \beta} + 1 \right) = n_2 \cos \gamma \left(1 - \frac{\tan \gamma}{\tan \beta} \right) \quad (\beta \neq 0) \quad (11.10)$$

For the last line, we must take care that $\beta \neq 0$, otherwise canceling the $\sin \beta$ terms would be division by zero - strictly not allowed. This is not a problem - β is only zero for the trivial case of the ray traveling on the principle axis, which we already know how to deal with. In order to proceed

further, we need to make a crucial approximation. Namely, we assume that the object is very distant relative to the radius of the spherical surface, $p \gg R$, and we only consider rays incident near the principle axis, $d \ll R$. If this is true, then the tangents of α , β , and γ can be nicely approximated:

$$\begin{aligned}\tan \alpha &\approx \frac{d}{OV} = \frac{d}{p} \\ \tan \beta &\approx \frac{d}{VC} = \frac{d}{R} \\ \tan \gamma &\approx \frac{d}{q}\end{aligned}$$

Basically, we have just decided to ignore the tiny distance between point V and the intersection of PV with the principle axis. Qualitatively, these approximations seem reasonable. It would be equivalent to say that we only consider large p and small α - the same approximations result - if α is small, so too are β and γ . Using these approximations, Eq. 11.10 reduces to:

$$\begin{aligned}n_1 \cos \alpha \left(1 + \frac{d/p}{d/R}\right) &= n_2 \cos \gamma \left(1 - \frac{d/q}{d/R}\right) \\ n_1 \cos \alpha \left(1 + \frac{R}{p}\right) &= n_2 \cos \gamma \left(1 - \frac{R}{q}\right)\end{aligned}\quad (11.11)$$

Now, given that the angles α and β are supposed to be tiny and the object distance large, we know that $p \gg d$ and $q \gg d$. Thus, the ratios d/p and d/q will be very small compared to 1. We can use this fact to simplify things even further. Using the same logic behind the tangent approximations, we find $\cos \alpha \approx 1$, and $\cos \gamma \approx 1$

$$\begin{aligned}\cos \alpha &\approx \frac{p}{\sqrt{d^2 + p^2}} = \frac{p}{p\sqrt{1 + d^2/p^2}} = \frac{1}{\sqrt{1 + d^2/p^2}} \approx 1 \\ \cos \gamma &\approx \frac{q}{\sqrt{d^2 + q^2}} = \frac{q}{q\sqrt{1 + d^2/q^2}} = \frac{1}{\sqrt{1 + d^2/q^2}} \approx 1\end{aligned}$$

Thus, so long as d/p and d/q are very small (and their squares are even smaller), we can simply ignore the cosine terms, which leaves us:

$$n_1 \left(1 + \frac{R}{p}\right) = n_2 \left(1 - \frac{R}{q}\right) \quad (11.12)$$

$$n_1 \frac{R}{p} + n_2 \frac{R}{q} = n_2 - n_1 \quad (11.13)$$

$$\Rightarrow \frac{n_1}{p} + \frac{n_2}{q} = \frac{n_2 - n_1}{R} \quad (11.14)$$

This is the result we desire: the image and object distances are simply related by the radius of curvature of the spherical surface, and the indices of refraction of the lens material and its surrounding.

Spherical refracting surfaces:

$$\frac{n_1}{p} + \frac{n_2}{q} = \frac{n_2 - n_1}{R} \quad (11.15)$$

Here q is the image distance inside the dense material n_2 , and p is the object distance in the less dense material n_1 ($n_1 < n_2$). The results holds for rays not far from the principle axis.

11.1.1 Flat Refracting Surfaces

If we let R tend toward infinity, $R \rightarrow \infty$, our spherical surface becomes a flat one.¹ If R tends toward infinity, then $1/R$ tends toward zero, and our spherical lens equation reduces to:

Flat refracting surfaces:

$$q = -\frac{n_2}{n_1}p \quad (11.16)$$

Here q is the image distance inside the dense material n_2 , and p is the object distance in the less dense material n_1 ($n_1 < n_2$).

This derives a result with important everyday consequences: since $n_2 \neq n_1$, then $p \neq q$. This is why, when looking into a pool of water, objects are actually much farther below the surface than we think they are.

11.2 Spherical Lenses

Armed with a knowledge of spherical refracting surfaces, we can move on to *spherical lenses*. All of the lenses we will consider can be defined only by the surfaces of spheres, hence the name. Figure 11.2 shows how one can construct either biconvex (upper) or biconcave (lower) spherical lenses, defined by the intersection and region between two spheres, respectively.

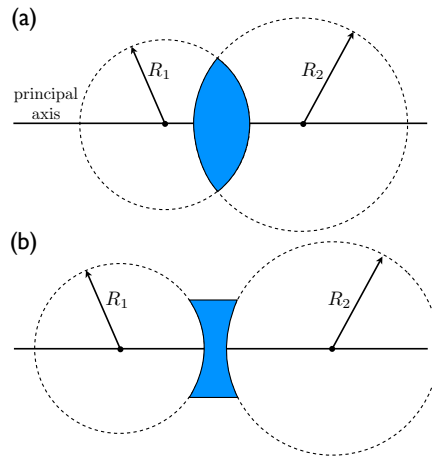


Fig. 11.2 Spherical lenses can also be either concave or convex, and their surfaces are defined by the surfaces of two spheres. **(a)** Biconvex lenses are formed by the intersection of two spheres, and **(b)** biconcave lenses are formed by the region between two spheres. When $R_1 = R_2$, the lens is spherically symmetric.

How can we analyze a lens like this? A lens can be considered the combination of two spherical interfaces, so all we need to do is use our solution to the case of the spherical refracting surface and

¹ One can say that the radius of curvature of a flat plane is infinite, or equivalently, that a plane is just the surface of a sphere with infinite radius.

apply it twice. First, we find the image due to (for instance) the left-hand spherical surface, and the image formed by that surface serves as the object for the right-hand spherical surface. This is shown in Fig. 11.3, where we consider a lens of thickness d formed by overlapping spheres of radii R_1 and R_2 , both of which are made of a material of refractive index n_2 . Surrounding the model lens is a material of refractive index n_1 .

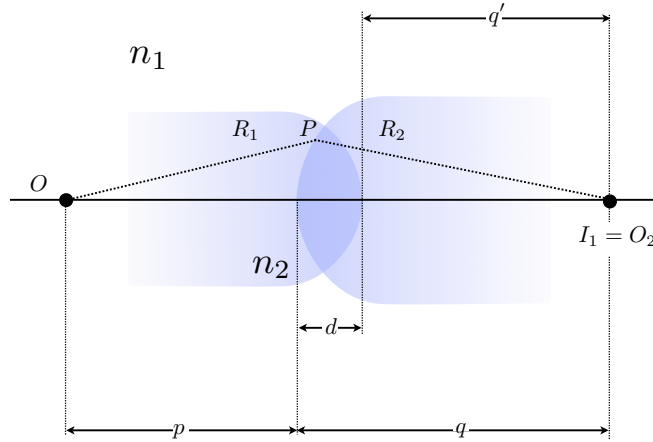


Fig. 11.3 Our model spherical lens is built out of two separate spherical refracting surfaces.

First, consider only the object on the right-hand side by itself, and pretend the right-hand refracting surface is not there at all. That is, we'll consider what happens due to one semi-infinite refracting object at a time. This means, implicitly, that we have defined distances to the *left* of point P to be positive, and distances to the *right* of point P to be negative. Light from point O , a distance p from the spherical surface, reaches the spherical interface at point P . Since we are only worrying in the end about the region where the two spherical surfaces overlap, we presume that the light is not refracted on the way from O to P . After refraction, the ray is refracted toward point I_1 on the principle axis. Since this is just refraction from a spherical surface as we solved above, we know

$$\frac{n_1}{p} + \frac{n_2}{q} = \frac{n_2 - n_1}{R_1} \quad (11.17)$$

This forms an image at point I_1 . This image now serves as an *object* for the second spherical surface - $I_1 = O_2$. Now ignore the right-hand side and consider only the left-hand side. Light from the image formed at O_2 will be incident on the spherical surface defined by R_2 in this case. Now, since point O_2 is on the *right* side of the lens, the object distance is *negative*, $p' < 0$. This distance is related to the object distance of the first lens, q , by the thickness of the lens:

$$p' = -(q - d) = d - q \quad (11.18)$$

where we made sure to carefully follow our sign convention, since p' is to the right of point P . Refraction from the spherical surface R_2 can be calculated in the same way:

$$\frac{n_2}{p'} + \frac{n_1}{q} = \frac{n_1 - n_2}{R_2} \quad (11.19)$$

$$\frac{n_2}{d - q} + \frac{n_1}{q} = \frac{n_1 - n_2}{R_2} \quad (11.20)$$

Now, add Eqns. 11.17 and 11.20:

$$\frac{n_1}{p} + \frac{n_2}{q} + \frac{n_2}{d-q} + \frac{n_1}{q} = \frac{n_1 - n_2}{R_2} + \frac{n_2 - n_1}{R_1} \quad (11.21)$$

$$\frac{n_1}{p} + \frac{n_2 + n_1}{q} + \frac{n_2}{d-q} = (n_2 - n_1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \quad (11.22)$$

This is the general equation for a spherical lens.

General equation for a spherical lens:

$$\frac{n_1}{p} + \frac{n_2 + n_1}{q} + \frac{n_2}{d-q} = (n_2 - n_1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \quad (11.23)$$

Here R_1 and R_2 are the radii of the spherical sections making up the lens, d is the thickness of the lens, n_2 the refractive index of the lens material, and n_1 of the surrounding material. The result holds for rays not far from the principle axis.

Most of the time, we are interested in the so-called *thin lens approximation*, in which we neglect the thickness of the lens. That is, we presume that the image and object distances are so large compared to the thickness of the lens, $p, q \gg d$, that we can safely neglect d . If we let $d \rightarrow 0$, we have what is known as the *lensmaker's formula*:

$$\frac{n_1}{p} + \frac{n_1}{q} = (n_2 - n_1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \quad (11.24)$$

We can find the focal length of the lens by considering the case of an extremely distant object, where we let p tend toward infinity just like we did with a spherical mirror. In that case, parallel rays will be converged on to a single focal point, just as with a spherical mirror, which we define to be the focal length f . Thus, we let p tend toward infinity (which makes $1/p$ tend toward zero), and find the corresponding value of q , which is the focal length f . This yields the more common form of the lensmaker's equation:

Lensmaker's equation:

$$\frac{1}{f} = \frac{1}{p} + \frac{1}{q} = \left(\frac{n_2 - n_1}{n_1} \right) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \quad (11.25)$$

here n_1 is the index of refraction of the surrounding material, n_2 of the lens. The lens is defined by the surfaces of spheres of radius R_1 and R_2 .

Comparing this to the preceding equation, we can also immediately relate the focal length to the image and object distance, which yields the 'lens equation':

Lens equation:

$$\frac{1}{f} = \frac{1}{p} + \frac{1}{q} \quad (11.26)$$

Surprise, surprise, the mirror equation is the same as the lens equation! A convex lens like the one we just considered will have a positive focal length f . Even though we derived these lens equations for the case of a convex lens, they are valid for thin concave lenses as well, so long as they are spherical. We will consider some other types of lenses shortly, but we have one bit of pressing business: we still don't know the magnification factor of the lens!

In order to determine the image magnification, it is easier at this point to construct a ray diagram, just as we did with mirrors. The rules are only slightly different:

How to construct ray diagrams:

Ray 1 is drawn parallel to the principle axis, and refracts through one focal point.

Ray 2 is drawn through the (other) focal point, and refracts parallel to the axis.

Ray 3 is drawn through the center of the lens, and continues in a straight line.

Figure 11.4 shows a ray diagram for a simple convex lens. Using the geometry of this figure, we can readily figure out the magnification factor, and verify our lens equation above to boot.

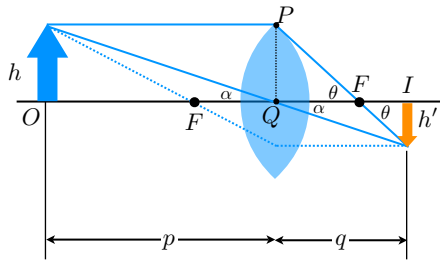


Fig. 11.4 Image construction with a biconvex lens.

Consider the triangle formed by points O , Q , and the tip of the object arrow. The tangent of the angle α is the object height over the object distance:

$$\tan \alpha = \frac{h}{p} \quad (11.27)$$

The triangle formed by points I , Q , and the tip of the *image* arrow give us another expression for $\tan \alpha$:

$$\tan \alpha = \frac{-h'}{q} \quad (11.28)$$

Comparing these two expressions, and using the definition of the magnification factor, we have our answer:

Magnification for a spherical lens:

$$M \equiv \frac{h'}{h} = -\frac{q}{p} = \frac{f-q}{f} = \frac{f}{f-p} \quad (11.29)$$

The last two forms are derived below. They follow by using the lens equation (11.26) in the first relationship.

Once again, the lens and mirror equations are the same - same spherical geometry, same equations. This formula is also much more general than our derivation suggests - it is valid for any spherical lens, not just the symmetric concave one we considered here.

We can also verify the lens equation by using the geometry of the uppermost ray. The triangle $\triangle PQF$ gives us another relationship, noting that the distance from the center of the lens (Q) to the focal point (F) is by definition the focal length ($\overline{QF} = f$) and $\overline{PQ} = h$:

$$\tan \theta = \frac{\overline{PQ}}{\overline{QF}} = \frac{h}{f} \quad (11.30)$$

The triangle defined by F , I , and the tip of the object arrow gives us one more equation:

$$\tan \theta = \frac{-h'}{q-f} \quad (11.31)$$

Comparing the last two equations, we have

$$\frac{h}{f} = \frac{-h'}{q-f} \quad (11.32)$$

$$\Rightarrow \frac{h'}{h} = -\frac{q-f}{f} \equiv M \quad (11.33)$$

Now we have two different expressions for M , which we can combine:

$$M = -\frac{q}{p} = -\frac{q-f}{f} \quad (11.34)$$

$$\frac{q}{p} = \frac{q}{f} - 1 \quad (11.35)$$

$$\frac{q}{p} + 1 = \frac{q}{f} \quad (11.36)$$

$$\frac{q}{p} + \frac{q}{q} = \frac{q}{f} \quad (11.37)$$

$$\Rightarrow \frac{1}{p} + \frac{1}{q} = \frac{1}{f} \quad (11.38)$$

A result that should be reassuring: we have now independently derived the lens equation. We can derive a third relationship between the magnification and focal length using the lens equation and our result above:

$$\frac{1}{q} = \frac{1}{f} - \frac{1}{p} \quad (11.39)$$

$$\Rightarrow q = \frac{fp}{p-f} \quad (11.40)$$

$$M = \frac{f-q}{f} \quad (11.41)$$

$$= \frac{f - \frac{fp}{p-f}}{f} \quad (11.42)$$

$$= \frac{fp - f^2 - fp}{f(p-f)} \quad (11.43)$$

$$= \frac{-f^2}{f(p-f)} \quad (11.44)$$

$$= \frac{f}{f-p} \quad (11.45)$$

This gives us three different relationships for the magnification factor, each one involving only two of the three quantities f , p , and q .

We now have all the mathematical and geometric ammunition we need for spherical lenses of any kind. Though we derived our results for the special case of convex lenses, they are more generally valid (it would take much more mathematics and geometry to demonstrate this, however), and hold for any spherical lenses we wish to consider. What we need to do next is figure out how different sorts of spherical lenses behave and what sorts of images they form on a qualitative level.

11.3 Types of spherical lenses

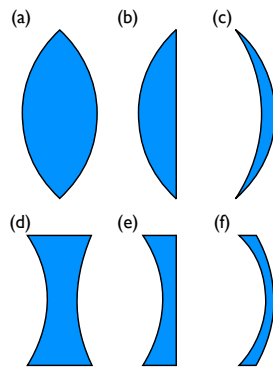


Fig. 11.5 There are a variety of common lens shapes, all essentially based on the intersection of two spheres or the space between two spheres. **(a)** Double convex, **(b)** plano-convex, **(c)** convex meniscus, **(d)** double concave, **(e)** plano-concave, **(f)** and concave meniscus lenses.

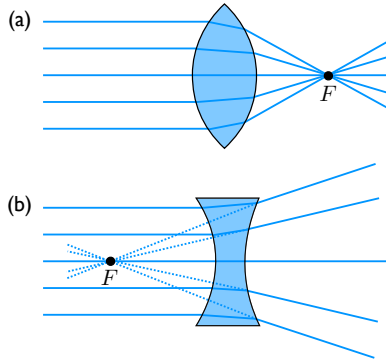


Fig. 11.6 **(a)** A biconvex lens converges distant light rays and focuses them onto a point – hence the name ‘focusing lens.’ **(b)** A biconcave lens causes distant light rays to *diverge*. They appear to diverge outward from a focal point on the incident side of the lens.

Solutions to Problems

Chapter 1 Problems

1.1 $2.3 \times 10^{13} \text{ J}$, $2.56 \times 10^{-4} \text{ kg}$. First part: relativistic kinetic energy is given by:

$$\text{KE} = (\gamma - 1)mc^2 \quad (11.46)$$

First, we'll calculate γ based on the given velocity:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{(0.75c)^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.75^2}} = 1.51 \quad (11.47)$$

Next, we'll calculate the mc^2 bit:

$$mc^2 = (5 \times 10^{-4} \text{ kg}) (3 \times 10^8 \text{ m/s})^2 = 4.5 \times 10^{13} \text{ kg} \cdot \text{m}^2/\text{s}^2 = 4.5 \times 10^{13} \text{ J} \quad (11.48)$$

Putting it all together:

$$\text{KE} = (\gamma - 1)mc^2 = (1.51 - 1)(4.5 \times 10^{13} \text{ J}) = 2.30 \times 10^{13} \text{ J} = 23.0 \text{ TJ} \quad (11.49)$$

Second part: what rest mass is equivalent to this amount of kinetic energy? We just need to use the mass-energy equivalence formula:

$$E_R = mc^2 = \text{KE} \quad (11.50)$$

$$\Rightarrow m = \frac{\text{KE}}{c^2} = \frac{(\gamma - 1)mc^2}{c^2} \quad (11.51)$$

$$= (\gamma - 1)m = 0.51m \quad (11.52)$$

$$= 2.56 \times 10^{-4} \text{ kg} \quad (11.53)$$

$$(11.54)$$

In other words, it takes fully half the mass of the bullet itself, completely converted to pure energy, to fire one round. Using more conventional propellants, that would mean 5760 kg (~ 6 tons) of TNT per round.

1.2 **1.8 taps/sec.** The 'proper time' Δt_p is that measured by the astronaut herself, which is 1/3 of a second between taps (so that there are 3 taps per second). The time interval *between taps* measured on earth is dilated (longer), so there are *less* taps per second. For the astronaut:

$$\Delta t_p = \frac{1 \text{ s}}{3 \text{ taps}} \quad (11.55)$$

On earth, we measure the dilated time:

$$\Delta t' = \gamma \Delta t_p = \frac{1}{\sqrt{1 - \frac{0.8^2 c^2}{c^2}}} \cdot \left(\frac{1 \text{ s}}{3 \text{ taps}} \right) = \frac{1}{\sqrt{1 - 0.8^2}} \cdot \left(\frac{1 \text{ s}}{3 \text{ taps}} \right) \approx \frac{0.56 \text{ s}}{\text{tap}} = \frac{1 \text{ s}}{1.8 \text{ taps}} \quad (11.56)$$

1.3 slow; slow. The time-dilation effect is symmetric, so observers in each frame measure a clock in the other to be running slow. Put another way, the *relative* velocity of the earth and the ship is the same no matter who you ask – each says the other is moving with some speed v , and they are sitting still. Therefore, the dilation effect is the same in both cases.

1.4 no; yes. There is no relative speed between you and your own pulse, since you are in the same reference frame, so there is no difference in your pulse rate (possible space-travel-related anxieties aside). There is a relative velocity between you and the people back on earth, however, so you would find their pulse rate *slower* than normal. Similarly, they would find *your* pulse rate slower than normal, since you are moving relative to them. Relativistic effects are always attributed to the other party – you are always at rest in your own reference frame.

1.5 9.61 sec. The proper time is that measured by in the reference frame of the pendulum itself, $\Delta t_p = 3.00 \text{ sec}$. The moving observer has to observe a *longer* period for the pendulum, since from the observer's point of view, the pendulum is moving relative to it. Observers always perceive clocks moving relative to them as running slow. The factor between the two times is just γ :

$$\Delta t' = \gamma \Delta t_p = \frac{3.0 \text{ sec}}{\sqrt{1 - \frac{0.95^2 c^2}{c^2}}} = \frac{3.0 \text{ sec}}{\sqrt{1 - 0.95^2}} \approx 9.61 \text{ sec} \quad (11.57)$$

1.6 134m. The electron in its own reference frame sees the *accelerator* moving toward it at $0.999c$, and sees a contracted length:

$$L = \frac{L_p}{\gamma} = 3 \text{ km} \cdot \sqrt{1 - \frac{0.999^2 c^2}{c^2}} = 3 \text{ km} \cdot \sqrt{1 - 0.999^2} = 0.134 \text{ km} = 134 \text{ m} \quad (11.58)$$

1.7 ellipsoid. The sphere is length contracted only along its direction of motion, *i.e.*, only along one axis. Squishing a sphere along one axis makes an ellipsoid.

1.8 The earth's clock. Less time will have passed in your reference frame, since you are moving relative to the earth. The earth's clock will have registered more time elapsed than yours.

1.9 We'll run it both forwards and backwards:

$$\text{KE} = \frac{1}{2}mv^2 = \frac{mv \cdot v}{2} = \frac{mv \cdot v}{2} \frac{m}{m} = \frac{mv \cdot mv}{2m} = \frac{p \cdot p}{2m} = \frac{p^2}{2m} \quad (11.59)$$

Or, since you know the answer you want ...

$$\frac{p^2}{2m} = \frac{(mv)^2}{2m} = \frac{m^2 v^2}{2m} = \frac{mv^2}{2} = \frac{1}{2}mv^2 \quad (11.60)$$

1.10 4.08 MeV for the muon, 29.6 MeV for the antineutrino. This one is a bit lengthier than most of the others! Before the collision, we have only the pion, and since it is at rest, it has zero momentum and zero kinetic energy. After it decays, we have a muon and an antineutrino created and speed off in opposite directions (to conserve momentum). Both total energy - including rest energy - and momentum must be conserved before and after the collision.

First, conservation of momentum. Before the decay, since the pion is at rest, we have zero momentum. Therefore, afterward, the muon and antineutrino must have equal and opposite momenta. This means we can essentially treat this as a one-dimensional problem, and not bother with vectors. A consolation prize of sorts.

$$\text{initial momentum} = \text{final momentum} \quad (11.61)$$

$$p_\pi = p_\mu + p_\nu \quad (11.62)$$

$$0 = p_\mu + p_\nu \quad (11.63)$$

$$\Rightarrow p_\nu = -p_\mu = -\gamma_\mu m_\mu v_\mu \quad (11.64)$$

For the last step, we made use of the fact that relativistic momentum is $p = \gamma mv$. Now we can also write down conservation of energy. Before the decay, we have only the rest energy of the pion. Afterward, we have the energy of both the muon and antineutrino. The muon has both kinetic energy and rest energy, and we can write its total kinetic energy in terms of γ and its rest mass, $E = \gamma mc^2$. The antineutrino has negligible mass, and therefore no kinetic energy, but we can still assign it a total energy based on its momentum, $E = pc$.

$$\text{initial energy} = \text{final energy} \quad (11.65)$$

$$E_\pi = E_\mu + E_\nu \quad (11.66)$$

$$m_\pi c^2 = \gamma_\mu m_\mu c^2 + p_\nu c \quad (11.67)$$

$$m_\pi = \gamma_\mu m_\mu + \frac{p_\nu}{c} \quad (11.68)$$

Now we can combine these two conservation results and try to solve for the velocity of the muon:

$$m_\pi = \gamma_\mu m_\mu + \frac{p_\nu}{c} \quad (11.69)$$

$$m_\pi = \gamma_\mu m_\mu - \gamma_\mu m_\mu \frac{v_\mu}{c} \quad (11.70)$$

$$\frac{m_\pi}{m_\mu} = \gamma_\mu - \gamma_\mu \frac{v_\mu}{c} \quad (11.71)$$

$$\frac{m_\pi}{m_\mu} = \gamma \left[1 - \frac{v_\mu}{c} \right] \quad (11.72)$$

We will need to massage this quite a bit more to solve for v_μ ...

$$\frac{m_\pi}{m_\mu} = \gamma \left[1 - \frac{v_\mu}{c} \right] = \frac{1 - \frac{v_\mu}{c}}{\sqrt{1 - \frac{v_\mu^2}{c^2}}} \quad (11.73)$$

$$\left(\frac{m_\pi}{m_\mu} \right)^2 = \frac{\left(1 - \frac{v_\mu}{c} \right)^2}{1 - \frac{v_\mu^2}{c^2}} \quad (11.74)$$

$$= \frac{\left(1 - \frac{v_\mu}{c} \right)^2}{\left(1 - \frac{v_\mu}{c} \right) \left(1 + \frac{v_\mu}{c} \right)} \quad (11.75)$$

$$= \frac{\left(1 - \frac{v_\mu}{c} \right)^2}{\cancel{\left(1 - \frac{v_\mu}{c} \right)} \left(1 + \frac{v_\mu}{c} \right)} \quad (11.76)$$

$$= \frac{1 - \frac{v_\mu}{c}}{1 + \frac{v_\mu}{c}} \quad (11.77)$$

Now we're getting somewhere. Take what we have left, and solve it for v_μ ... we will leave that as an exercise to the reader, and quote only the result, using the given masses of the pion and muon:

$$\frac{v_\mu}{c} = \frac{1 - \left(\frac{m_\pi}{m_\mu} \right)^2}{1 + \left(\frac{m_\pi}{m_\mu} \right)^2} \approx -0.270 \quad (11.78)$$

From here, we are home free. We can calculate γ_μ and the muon's kinetic energy first. It is convenient to remember that the electron mass is $511 \text{ keV}/c^2$.

$$\gamma_\mu = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{(0.27c)^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.27^2}} \approx 1.0386 \quad (11.79)$$

$$\text{KE}_\mu = (\gamma_\mu - 1) m_\mu c^2 = (1.0386 - 1) (207 m_e) c^2 \quad (11.80)$$

$$= 0.0386 (207 \cdot 511 \text{ keV}/c^2) c^2 \approx 4.08 \times 10^6 \text{ eV} = 4.08 \text{ MeV} \quad (11.81)$$

Finally, we can calculate the energy of the antineutrino as well:

$$E_\nu = p_\nu c = -p_\mu c \quad (11.82)$$

$$= -\gamma_\mu m_\mu v_\mu \quad (11.83)$$

$$= -1.0386 \cdot (207 \cdot 5.11 \text{ keV}/c^2) \cdot (-0.270c) \quad (11.84)$$

$$\approx 2.96 \times 10^7 \text{ eV} = 29.6 \text{ MeV} \quad (11.85)$$

1.11 $1.96 \times 10^{13} \text{ m}$. The 15 h set on the alarm clock in the spaceship is the proper time interval, Δt_p . Since the space ship is moving away from the earth at $v = 0.77c$, an earthbound observer observes a longer dilated time interval, $\Delta t'$. Based on this longer time interval, the earthbound observer will measure that the space ship has covered a distance of $v\Delta t'$. So, first: we need to calculate γ , then the dilated time interval, then finally the distance measured by the earthbound observer.

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{(0.77c)^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.77^2}} = 1.57 \quad (11.86)$$

$$\Delta t' = \gamma \Delta t_p \quad (11.87)$$

$$= 1.57 \cdot 15 \text{ h} = 1.57 \cdot 5.4 \times 10^4 \text{ s} \approx 8.48 \times 10^4 \text{ s} \quad (11.88)$$

$$d' = v \Delta t' \quad (11.89)$$

$$= 0.77c \cdot 8.48 \times 10^4 \text{ s} = 0.77 \cdot 3 \times 10^8 \text{ m/s} \cdot 8.48 \times 10^4 \text{ s} \quad (11.90)$$

$$d' \approx 1.96 \times 10^{13} \text{ m} \quad (11.91)$$

1.12 $1.31 \times 10^{-7} \text{ s}$, **38.4 m**, **7.64 m** The π meson's lifetime in its own frame is the proper time interval, $\Delta t_p = 2.6 \times 10^{-8} \text{ s}$. An earthbound observer measures a longer dilated time interval $\Delta t'$. To calculate it, we need only calculate γ for the velocity given, $v_\pi = 0.98c$.

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{(0.98c)^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.98^2}} = 5.03 \quad (11.92)$$

$$\Delta t' = \gamma \Delta t_p \quad (11.93)$$

$$= 5.03 (2.6 \times 10^{-8} \text{ s}) \quad (11.94)$$

$$\approx 1.31 \times 10^{-7} \text{ s} \quad (11.95)$$

The distance the π meson travels in the earthbound observer's reference frame, d' is the π meson's velocity multiplied by the time interval measured by the earthbound observer. We don't need to worry about whether the velocity is measured in the π meson's or the observer's frame - since it is a relative velocity, it is the same either way.

$$d' = \gamma v_\pi \Delta t_p = v_\pi \Delta t' = (0.98c) \cdot (1.31 \times 10^{-7} \text{ s}) = (0.98 \cdot 3 \times 10^8 \text{ m/s}) \cdot (1.31 \times 10^{-7} \text{ s}) \approx 38.4 \text{ m} \quad (11.96)$$

Without time dilation, the distance traveled would just be the proper lifetime multiplied by the meson's velocity:

$$d = v_{\pi} \Delta t_p = (0.98c) \cdot (2.6 \times 10^{-8} \text{ s}) = (0.98 \cdot 3 \times 10^8 \text{ m/s}) \cdot (2.6 \times 10^{-8} \text{ s}) \approx 7.64 \text{ m} \quad (11.97)$$

1.13 No. There is no relative speed between you and your cabin, since you are in the same reference frame. You and your bed will remain at the same lengths relative to each other.

1.14 8.42s. The time interval in the probe's reference frame is the proper one Δt_p ... which makes sense, since the antenna is part of the probe itself! The probe and antenna are moving relative to the earth, and therefore the earthbound observer measures a longer, dilated time interval $\Delta t'$:

$$\text{probe} = \Delta t_p \quad (11.98)$$

$$\text{earth} = \Delta t' \quad (11.99)$$

$$\Delta t' = \gamma \Delta t_p \quad (11.100)$$

As usual, we first need to calculate γ . No problem, given the probe's velocity of $0.88c$ relative to earth:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{(0.88c)^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.88^2}} = 2.11 \quad (11.101)$$

The proper time interval for one revolution Δt_p in the probe's reference frame is 4.0s, so we can readily calculate the time interval observed by the earthbound observer:

$$\Delta t' = \gamma \Delta t_p = 2.11 \cdot (4.0 \text{ s}) = 8.42 \text{ s} \quad (11.102)$$

1.15 24m; 18m; 0.661c. Once again: if you are observing something in your own reference frame, there is no length contraction or time dilation. You always observe your own ship to be the same length. If your friend's ship is 24m long, and yours is identical, you will measure it to be 24m.

On the other hand, you are moving relative to his ship, so you would observe his ship to be length contracted, and measure a shorter length. Your friend, on the other hand, will observe *exactly the same thing* - he will see *your* ship contracted, by precisely the same amount. Your observation of his ship has to be the same as his observation of his ship - since you are only the two observers, and you both have the same *relative* velocity, you must observe the same length contraction. If he sees your ship as 18m long, then you would also see his (identical) ship as 18m long.

Given the relationship between the contracted and proper length, we can find the relative velocity easily. Your measurement of your own ship is the proper length L_p , while your measurement of your friend's ship is the contracted length L' :

$$L_p = \gamma L' \quad (11.103)$$

$$\Rightarrow \gamma = \frac{L_p}{L'} = \frac{24}{18} = \frac{4}{3} \quad (11.104)$$

$$\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{4}{3} \quad (11.105)$$

$$1 - \frac{v^2}{c^2} = \frac{3^2}{4^2} = \frac{9}{16} \quad (11.106)$$

$$\frac{v^2}{c^2} = 1 - \frac{9}{16} = \frac{7}{16} \quad (11.107)$$

$$v = \sqrt{\frac{7}{16}}c = \frac{\sqrt{7}}{4}c \approx 0.661c \quad (11.108)$$

1.16 0.541c. This is just a problem of relativistically adding velocities, if we can keep them all straight. Let the unprimed system denote velocities measured relative to the earth, and the primed system those measured relative to the enterprise. We have, then:

$$v_e = 0.900c = \text{Enterprise relative to earth} \quad (11.109)$$

$$v_k = 0.700c = \text{Klingon ship relative to earth} \quad (11.110)$$

$$v'_k = ? = \text{Klingon ship, relative to Enterprise} \quad (11.111)$$

Since the Enterprise is moving faster relative to the earth than the Klingon ship, that means that from the Enterprise's point of view, the Klingons are actually moving backwards toward them. If we plug what we know into the velocity addition formula ...

$$v_k = \frac{v_e + v'_k}{1 + \frac{v_e v'_k}{c^2}} \quad (11.112)$$

It takes a bit of algebra, but we can readily solve this for v'_k :

$$v'_k = \frac{v_e - v_k}{1 - \frac{v_e v_k}{c^2}} \quad (11.113)$$

Not so surprisingly, what we have just done is to re-write the 'velocity addition formula' as a 'velocity subtraction formula.' It is just rearranging same formula (you can verify that both equations above are equivalent ...), but the second form is far more convenient for our present purposes.

We can find the velocity of the Klingon ship relative to the enterprise in terms of both ships' velocities relative to the earth. In the limit that both velocities are much smaller than c , we see that $v'_k \approx v_e - v_k = 0.200c$, just as we would expect from normal Newtonian physics. Since in this case, neither velocity is negligible compared to c , the actual v'_k will be significantly larger. At this point, we can just plug in the numbers we have and see:

$$v'_k = \frac{v_e - v_k}{1 - \frac{v_e v_k}{c^2}} \quad (11.114)$$

$$= \frac{0.900c - 0.700c}{1 - \frac{(0.900c)(0.700c)}{c^2}} \quad (11.115)$$

$$= \frac{0.200c}{1 - (0.900)(0.700)} = \frac{0.200c}{0.37} \quad (11.116)$$

$$v'_k \approx 0.541c \quad (11.117)$$

So, as far as the crew of the Enterprise is concerned, they are overtaking the Klingon ship at a rate of 0.541c.

1.17 $0.99995c$. Let the observer be in frame O' . In the reference frame of one of the particles, labeled O , the observer is traveling at $v = 0.99c$, and the second particle is traveling at $v'_2 = 0.99c$ relative to the observer. We can then find the velocity of the second particle relative to the first, v_2 , through velocity addition:

$$v_2 = \frac{v + v'_1}{1 + \frac{vv'_1}{c^2}} \quad (11.118)$$

$$= \frac{0.99c + 0.99c}{1 + \frac{(0.99c)(0.99c)}{c^2}} \quad (11.119)$$

$$= \frac{1.98c}{1 + 0.9801} \approx 0.99995c \quad (11.120)$$

This is an example of a problem where you need to make sure to use enough significant digits!

1.18 $0.87c$. The proper length of a meter stick, measured in its own reference frame, is obviously 1 m. For a moving observer to see the meter stick as only $L' = 0.5$ m long, we need a length contraction of a factor 2:

$$L' = \frac{L_p}{\gamma} \implies \frac{L_p}{L'} = \gamma = 2 \quad (11.121)$$

Thus, for the meter stick to be contracted by a factor 2, we need $\gamma = 2$. Using the equation for γ in terms of γ above, you should find $v = 0.87c$:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (11.122)$$

$$\sqrt{1 - \frac{v^2}{c^2}} = \frac{1}{\gamma} \quad (11.123)$$

$$1 - \frac{v^2}{c^2} = \frac{1}{\gamma^2} \quad (11.124)$$

$$v^2 = c^2 (1 - \gamma^{-2}) \quad (11.125)$$

$$v = c \sqrt{1 - \frac{1}{\gamma^2}} = c \sqrt{1 - \frac{1}{4}} = \frac{c\sqrt{3}}{2} \approx 0.87c \quad (11.126)$$

1.19 $0.305c$. We can use the result of the last problem here - once again, we know γ , and want to find the corresponding v . Really, just an exercise to make sure you have your algebra down cold ...

$$v = c \sqrt{1 - \frac{1}{\gamma^2}} = c \sqrt{1 - \frac{1}{1.05^2}} \approx 0.305c \quad (11.127)$$

1.20 $0.87c$ Once again, the factor between the two times is just γ . The clock at rest measures the proper time Δt_p . If the moving clock runs only half as fast, its time intervals $\Delta t'$ are twice as long $\Delta t' = 2\Delta t_p$. Thus:

$$\frac{\Delta t'}{\Delta t_p} = \gamma = 2 \quad (11.128)$$

From the first problem, we know that $\gamma = 2$ occurs when $v \approx 0.87c$.

1.21 23.4 m We presume that the motion is purely along the direction of the spaceship's length. Since length contraction occurs only along the direction of motion, the width is unaffected, it still appears to be 25.0 m for the external observer. Along the direction of motion, the length should appear contracted by a factor γ . remember that the proper length is that measured at rest. As usual, the primed quantities are for the external observer.

$$L' = \frac{L_p}{\gamma} \quad (11.129)$$

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.95^2}} \approx 3.2 \quad (11.130)$$

$$\Rightarrow L' = \frac{75 \text{ m}}{3.2} \approx 23.4 \text{ m} \quad (11.131)$$

1.22 $0.99c$ The first thing we need to do in order to avoid confusion is label everything properly. We will say the observer on earth is in the unprimed reference frame, and those in the first ship are in the primed frame. Since the spacecraft are moving in opposite directions, *one of them has to be negative*. Let us say that spaceship 1 is moving in the positive direction, so that spaceship 2 has a negative velocity.

$$v_1 = \text{velocity of first ship observed from earth} = 0.8c \quad (11.132)$$

$$v_2 = \text{velocity of second ship observed from earth} = -0.9c \quad (11.133)$$

$$v'_2 = \text{velocity of second ship observed from first} = ? \quad (11.134)$$

What we want to find is v'_2 , the relative velocity of the two ships. If we completely ignore relativity just for a minute, what would the answer be? The relative velocity of the two ships would just be the velocity of one minus the other - we subtract the two velocities as measured from earth to get their relative velocity. Now, to do it correctly, we just need to use our relativistic velocity subtraction formula, taking care that one of them is negative:

$$v'_2 = \frac{v_2 - v_1}{1 - \frac{v_2 v_1}{c^2}} \quad (11.135)$$

$$= \frac{-0.9c - 0.8c}{1 - (0.8)(-0.9)} = \frac{-1.7c}{1.72} \approx -0.99c \quad (11.136)$$

The overall answer comes out negative, which makes sense: the velocity of ship 2 is still in the negative direction when viewed from ship 1.

1.23 $0.98c$ Just like the last problem, let us first label what we know. Let the observer on the ground be in the unprimed frame, and the passenger in the car the primed frame:

$$v_b = \text{velocity of the ball relative to the ground} = ? \quad (11.137)$$

$$v_c = \text{velocity of the car relative to the ground} = 0.9c \quad (11.138)$$

$$v'_b = \text{velocity of the ball relative to the car} = 0.7c \quad (11.139)$$

Again, ask yourself how you would figure this out without relativity first, and that will help you pick the proper relativistic formula. Without relativity, you would just add the velocity of the car relative to the ground and the velocity of the ball relative to the car. Thus, all we need to do use our correct relativistic velocity addition formula:

$$v_b = \frac{v_c + v'_b}{1 + \frac{v_c v'_b}{c^2}} \quad (11.140)$$

$$= \frac{1.6c}{1 + (0.9)(0.7)} \approx 0.98c \quad (11.141)$$

1.24 $15.4 \mu\text{s}$; 649 m Let the earth be in reference frame O' (primed frame), and the muon itself in O (unprimed frame). First, since we know we will need it, for $v = 0.990c$, $\gamma = 7.09$. Next, the numbers we are given are measured in the earth's reference frame, so it will be easiest to calculate the time in the earth's frame first. The muon, according to earthbound observers, travels 4600 m at a speed of $0.990c$, so the apparent decay time is just distance divided by velocity.

$$\Delta t'_{\text{earth}} = \frac{4600 \text{ m}}{0.990(3 \times 10^8 \text{ m/s})} \approx 1.54 \times 10^{-5} \text{ s} = 15.4 \mu\text{s} \quad (11.142)$$

This is *not* the proper time - proper time is measured in the muon's own frame. According to the muon, the earth is moving toward them! Given γ and time measured on earth, we can find the proper time in the muon's frame easily:

$$\Delta t_p = \frac{\Delta t'_{\text{earth}}}{\gamma} \approx \frac{1.54 \mu\text{s}}{7.09} = 2.18 \mu\text{s} \quad (11.143)$$

This makes sense - since the people on earth are the moving observers in this case, they should see a longer time interval. About seven times longer, in this case, since $\gamma \approx 7$. The muon is at rest in its own frame, and measures the shorter proper time interval. Now we have the proper time, measured in the muon's reference frame, and the relative velocity, so we can calculate the distance from the muon's point of view using quantities valid in its reference frame.

$$d_\mu = v\Delta t_p = v \frac{\Delta t'_{\text{earth}}}{\gamma} \approx 649 \text{ m} \quad (11.144)$$

1.25 $v \leq 0.14c$, $v \leq 0.31c$. First of all, what do we mean by error? You want to find percent error between momentum calculated with the relativistic formula (*viz.*, $|\vec{p}_{\text{rel}}| = \gamma m |\vec{v}|$) and the classical formula (*viz.*, $|\vec{p}_{\text{class}}| = m |\vec{v}|$). First, we will drop the vector notation now, since error in momentum will only be in magnitude, not direction. Let $p_{\text{rel}} \equiv |\vec{p}_{\text{rel}}|$ and $p_{\text{class}} \equiv |\vec{p}_{\text{class}}|$. The definition of error you want is the difference between the two, divided by the correct one - the relativistic formula.

$$100\% \cdot \left| \frac{p_{\text{rel}} - p_{\text{class}}}{p_{\text{rel}}} \right| \leq \text{error desired} \quad (11.145)$$

For the last line, we drop the percent. Now we can just plug in what we know:

$$\left| \frac{p_{\text{rel}} - p_{\text{class}}}{p_{\text{rel}}} \right| = \left| \frac{\gamma mv - mv}{\gamma mv} \right| = \left| \frac{\gamma \cancel{mv} - \cancel{mv}}{\gamma \cancel{mv}} \right| = \left| \frac{\gamma - 1}{\gamma} \right| \leq \text{error} \quad (11.146)$$

We can further simplify this:

$$\left| \frac{\gamma - 1}{\gamma} \right| = \left| 1 - \frac{1}{\gamma} \right| \leq \text{error} \quad (11.147)$$

$$1 - \text{error} \leq \frac{1}{\gamma} \quad (11.148)$$

$$(11.149)$$

Here the last step is valid since γ is always positive and greater than one (unless the body is at rest, which is a bit silly in this context). What we really want is v . Remember the equation for v in terms of γ from problem 1.18? Take that, and plug in the expression above:

$$v = c \sqrt{1 - \frac{1}{\gamma^2}} \leq c \sqrt{1 - \left| (1 - \text{error}) \right|^2} \quad (11.150)$$

Now all we need to do is plug in the desired minimum errors - 1% or 0.01 for **(a)**, and 5% or 0.05 for **(b)**:

$$(a) \quad v \leq c \sqrt{1 - \left| (1 - \text{error}) \right|^2} = c \sqrt{1 - \left| (1 - 0.01) \right|^2} \approx c \sqrt{0.02} \approx 0.14c \quad (11.151)$$

$$(b) \quad v \leq c \sqrt{1 - \left| (1 - \text{error}) \right|^2} = c \sqrt{1 - \left| (1 - 0.05) \right|^2} \approx c \sqrt{0.098} \approx 0.31c \quad (11.152)$$

$$(11.153)$$

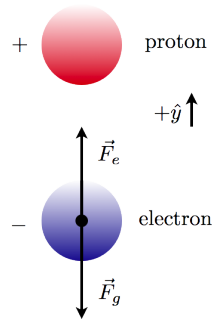
Chapter 2 Problems

2.1 Conducting shoes are worn to avoid building up static charge while walking. Rubber-soled shoes as they rub against the floor build up a static charge (conduction charging), which could discharge with a spark. The spark could cause a fire or explosion in an oxygen-rich environment. Conducting shoes provide a constant connection to the earth, “grounding” the wearer and allowing any excess charge to leak away.

2.2 dimensional analysis ...

2.3 Given: A proton and electron separated by some distance. The force between the oppositely-charged proton and electron will be attractive. If the force on the electron is to be upward, it must be true that the proton is sitting directly above the electron. This results in the electric force pulling the electron up, while the gravitational force (its weight) pulls it down. From Table 2.1, we also know the charge and mass of the proton and electron.

Find: We want to find the distance at which the attractive electric force balances the electron’s weight, which means we must balance the gravitational and electric forces.



Sketch: Let both charges be aligned along the y axis, with the gravitational force \vec{F}_g acting in the $-\hat{y}$ direction. The attractive electric force \vec{F}_e then acts in the $+\hat{y}$ direction. Our origin will be at the proton’s position. The sum of the gravitational force on the electron and the attractive electric force must be zero.

Relevant equations: We have two charges interacting, and thus the force between them must be given by Eq. 2.1. The electron’s weight is the gravitational force acting on it, $F_g = mg$. Additionally, we need Newton’s second law, $\sum F = ma$.

Symbolic solution: Using Eq. 2.1, we can easily write down the electric force. A force balance will give us a relationship involving d and known quantities. Let the electron have mass m_e and charge $-e$, the proton charge e , and the distance between them we will call d .

$$\sum F_y = 0 = F_e + F_g = -m_e g + \frac{k_e e (-e)}{d^2} \quad (11.154)$$

$$\implies d^2 = \frac{k_e e^2}{m_e g} \implies d = \sqrt{\frac{k_e e^2}{m_e g}} \quad (11.155)$$

Numeric solution: All the numbers we need can be found in Table 2.1. Since this is a rather hypothetical problem, we suppose that one significant digit in the answer is enough – the point is

merely to get a feeling for the relative magnitudes of electric and gravitational forces, so an order-of-magnitude estimate is fine. Plugging in our numbers, and keeping careful track of units,

$$d = \sqrt{\frac{(9 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2) (1.6 \times 10^{-19} \text{ C})^2}{(9.1 \times 10^{-31} \text{ kg}) (9.8 \text{ m/s}^2)}} \quad (11.156)$$

$$\approx \sqrt{26 \frac{\text{N} \cdot \text{m} \cdot \text{s}^2}{\text{kg}}} \approx 5 \sqrt{\frac{(\text{kg} \cdot \text{m/s}^2) \cdot \text{m} \cdot \text{s}^2}{\text{kg}}} \approx 5 \text{ m} \quad (11.157)$$

Is it reasonable? Dimensionally, our answer is correct – we carried units throughout the numerical calculation, and ended up with the proper dimensions of meters.

2.7 90N, or roughly 20lbs!

2.8 Given: The initial and final speeds v_i and v_f of a proton in a uniform electric field E . Additionally, from Table 2.1 we know the proton's mass and charge.

Find: The acceleration on the proton.



Sketch: There is not much too this one: we have a single proton present in an electric field. The proton's speed and direction of motion is not relevant to finding the acceleration.

Relevant equations: Since we know the electric field present and the charge of a proton, we know the electric force from Eq. 2.3. Additionally, we need Newton's second law, $\sum F = ma$.

Symbolic solution: The acceleration the proton experiences. If the proton, with charge e , is in a uniform electric field E , it experiences a constant electric force $F_e = eE$. If this is the only force acting on it, then by Newton's second law $F = qE = m_p a$, where m_p is the proton mass and a its acceleration. That's it – the speed is superfluous.

$$\sum F = qE = m_p a \implies a = \frac{qE}{m_p} \quad (11.158)$$

Numeric solution: Using numbers from Table 2.1,

$$F = qE = m_p a \quad (11.159)$$

$$\implies a = \frac{qE}{m_p} = \frac{(1.60 \times 10^{-19} \text{ C}) (800 \text{ N/C})}{1.67 \times 10^{-27} \text{ kg}} \approx 7.7 \times 10^{10} \text{ N/kg} = 7.7 \times 10^{10} \text{ m/s}^2 \quad (11.160)$$

For the last line, we had to use the fact that $1 \text{ N} = 1 \text{ kg} \cdot \text{m/s}^2$.

Is it reasonable?

2.9 Starting this one is not complicated: write down the electric field at a point along the x axis for each charge. The superposition principle says that the total electric field at that point is the sum of the fields from each charge alone. If we are at a point $(x, 0)$, then the $-q$ charge is a distance $x+a$ away, and the $+q$ charge is $x-a$ away. Thus:

$$E_{\text{tot}} = E_q + E_{-q} = \frac{k_e q}{(x-a)^2} + \frac{k_e (-q)}{(x+a)^2} = \frac{k_e q (x+a)^2}{(x-a)^2 (x+a)^2} - \frac{k_e q (x-a)^2}{(x-a)^2 (x+a)^2} \quad (11.161)$$

$$= \frac{k_e q (x^2 + 2ax + a^2) - k_e q (x^2 - 2ax + a^2)}{(x^2 - a^2)^2} = \frac{4k_e q ax}{(x^2 - a^2)^2} \quad (11.162)$$

Now what? The key is that when we specify that we want the field at a “distant” point, we mean the distance x is much, much larger than the spacing a , *i.e.*, $x \gg a$. Large enough that we can use mathematical approximations, basically. First, some rearranging:

$$E_{\text{tot}} = \frac{4k_e q ax}{(x^2 - a^2)^2} = \frac{4k_e q ax}{x^4 (1 - a^2/x^2)^2} \quad (11.163)$$

If we specify that $x \gg a$, then the larger x gets, the smaller a^2/x^2 gets, and for large distances $1 - a^2/x^2 \approx 1$. More directly, before rearranging anything we might have just claimed that since when $x \gg a$, $x^2 - a^2 \approx x^2$ - ignore the a^2 since it is much smaller anyway. Formally, this is considered Not OK, even though it works here. Typically, to make an approximation like this you want to get an expression such that in the limit x tends toward infinity, some term goes to zero and can be ignored - in this case, a^2/x^2 goes to zero, so we drop it. In some sense this is just being pedantic, but this more general trick is very useful for more complicated equations.

In any case, the effect here is the same: the denominator can be approximated as x^4 . Using this approximation,

$$E_{\text{tot}} \approx \frac{4k_e q ax}{x^4} = \frac{4k_e qa}{x^3} \quad (11.164)$$

A positive and a negative charge like this is a *dipole*, something that comes up a lot - for instance, it is a reasonable approximation of a diatomic molecule (*e.g.*, HCl).

2.4 1 - inside the hollow sphere. 2 - inside the hollow sphere. 3 - true everywhere, check for yourself. 4 - ends of the plates.

2.5 Choosing a cube would not give us any nice surfaces with a constant electric field on them.

2.6 From Gauss' law, we know that the *total* flux through any closed surface is just the total charge Q contained by the surface divided by ϵ_0 . If the charge is exactly at the center of the cube, the flux through each face of the cube should be the same. Six faces, each with one sixth the flux:

$$\Phi_{E,\text{face}} = \frac{1}{6} \Phi_{E,\text{total}} = \frac{1}{6} \frac{100 \mu\text{C}}{\epsilon_0} = 1.88 \times 10^7 \text{ N} \cdot \text{m}^2/\text{C} \quad (11.165)$$

The total flux is just six times this number, clearly. What about if the charge isn't at the center? The total flux remains the same no matter where the charge is, so long as it is inside the cube. However, to find the flux through a given face we assumed that the flux through all faces was equal based on the symmetry of the original problem. If the charge were closer to one face, that face would have higher flux, and the opposite side lower flux. Moving the charge from the center changes the distribution of flux over each face, and they will no longer all be equal, but the total remains the same. In other words, the first answer would change, the second would not.

2.10 **0.** Halfway between, the magnitude of the field from each individual charge is the same, but *they act in opposite directions*. Therefore, exactly in the middle, they cancel, and the field is zero. This is the same as the field exactly at the midpoint of an electric dipole. It might be easier to convince yourself the field is zero if you draw a picture including the electric field lines.

2.11 **0.** The field at the center from a point on the ring is always canceled by the field from another point 180° away.

2.12 $-q, -q$. The thing to remember is that any charge on a conductor spreads out evenly over its surface. When we have the conducting spheres isolated, they have q and $-3q$ respectively, and this charge is spread evenly over each sphere. When we connect them with a conducting wire, suddenly

charges are free to move from one conductor, across the wire, into the other conductor. It's just the same as if we had one big conductor, and all the *total net charge of the two conductors combined* will spread out evenly over *both* spheres and the wire.

If the charge from each sphere is allowed to spread out evenly over both spheres, then the $-3q$ and $+q$ will both be spread out evenly everywhere. The $+q$ will cancel part of the $-3q$, leaving a total net charge of $-2q$ spread over evenly over both spheres, or $-q$ on each sphere. Once we disconnect the two spheres again, the charge remains equally distributed between the two.

2.13 $|\vec{E}| = \frac{k_e q}{r^2}$. The easiest way out of this one is Gauss' law. First, Gauss' law told us that any spherically symmetric charge distribution behaves as a point charge. Second, Gauss' law tells us that the electric flux out of some surface depends only on the enclosed charge. If we draw a spherical surface of radius r and area A around the shell and point charge, centered on the center of the conducting sphere, Gauss' law gives:

$$\Phi_E = \frac{q_{encl}}{\epsilon_0} = 4\pi k_e q_{encl} \quad (11.166)$$

$$EA = 4\pi k_e q_{encl} \quad (11.167)$$

$$E = \frac{4\pi k_e q_{encl}}{A} \quad (11.168)$$

The surface area of a sphere is $A = 4\pi r^2$. In this case, the enclosed charge is just q , since the hollow conducting sphere itself has no charge of its own. Gauss' law only cares about the *total net charge* inside the surface of interest. This gives us:

$$E = \frac{4\pi k_e q}{4\pi r^2} = \frac{\cancel{4}\pi k_e q}{\cancel{4}\pi r^2} = \frac{k_e q}{r^2} \quad (11.169)$$

There we have it, it is just the field of a point charge q at a distance r .

If we want to get formal, we should point out that the point charge q induces a negative charge $-q$ on the inner surface of the hollow conducting sphere. Since the sphere is overall neutral, the outer surface must therefore have a net positive charge $+q$ on it. This makes no difference in the result – the total *enclosed* charge, for radii larger than that of the hollow conducting sphere ($r > R$), is still just q . If we start with an uncharged conducting sphere, and keep it physically isolated, any induced charges have to cancel each other over all.

If this is still a bit confusing, go back and think about induction charging again. A charged rod was used to induce a positive charge on one side of a conductor, and a negative charge on the other. *Overall*, the 'induced charge' was just a rearrangement of existing charges, so if the conductor started out neutral, no amount of 'inducing' will change that. We only ended up with a *net charge* on the conductor when we used a ground connection to 'drain away' some of the induced charges. Or, if you like, when we used a charged rod to repel some of the conductor's charges through the ground connection, leaving it with a net imbalance.

2.14 (b). If the charges are of the *opposite* sign, then the field lines would have to run from one charge directly to the other. Field lines start on a positive charge and end on a negative one, and there should be many lines which run from one charge to the other. Since opposite charges attract, the field between them is extremely strong, the lines should be densest right between the charges. This is the case in (a) and (c), so they are not the right ones.

By the same token, for charges of the *same* sign, the force is repulsive, and the electric field midway between them cancels. The field lines should "push away" from each other, and no field line from a given charge should reach the other charge – field lines cannot start and end on the same sign charge. This means that only (b) and (d) could possibly correspond to two charges of the same sign.

Next, the field lines leaving or entering a charge has to be proportional to the magnitude of the charge. In (d) there are the same number of lines entering and leaving each charge, so the charges are of the same magnitude. One can also see this from the fact that the lines are symmetric about a vertical line drawn midway between the charges. In (b) there are clearly many more lines near the left-most charge.

Or, right off the bat, you could notice that only (a) and (b) are asymmetric, and only (b) and (d) look like two like charges. No sense in over-thinking this one.

2.15 (a). By similar reasoning as above, only figure **a** could represent two opposite charges of different magnitude.

2.16 The third and fourth statements are true. The electric force on the proton is equal in magnitude to the force on the electron, but in the opposite direction. The magnitude of the acceleration of the electron is greater than that of the proton.

2.17 $E=0$. The simplest way to solve this one is with Gauss' law. First, Gauss law told us that any spherically symmetric charge distribution behaves as a point charge. Second, Gauss law tells us that the electric flux out of some surface depends only on the enclosed charge. If we draw a spherical surface of radius r and area A enclosing the shell *and* the point charge, centered on the center of the conducting sphere, the total enclosed charge is that of the shell plus that of the point charge: $q_{\text{encl}} = q + (-q) = 0$. If the enclosed charge is zero for any sphere drawn outside of and enclosing the spherical shell, then the electric field for all points outside the spherical shell.

2.18 $E = k_e q / r^2$. Just like the last question, we need Gauss' law. This time, we have to draw a sphere surrounding the point charge, but *inside* of the spherical shell. Gauss' law tells us that the electric field depends only on the *enclosed* charge within our sphere. The only charge enclosed is the point charge at the center of the shell, q – the charge on the spherical shell is outside of our spherical surface, so it is not enclosed and does not contribute to the electric field inside. Now we just apply Gauss' law, knowing that the enclosed charge is q , and the surface area of the sphere is $4\pi r^2$:

$$\Phi_E = \frac{q_{\text{encl}}}{\epsilon_0} = 4\pi k_e q \quad (11.170)$$

$$EA = 4\pi k_e q \quad (11.171)$$

$$E = \frac{4\pi k_e q}{4\pi r^2} = \frac{k_e q}{r^2} \quad (11.172)$$

2.19 $+6C/\epsilon_0$. Again, this question requires Gauss' law. We know that the electric flux through this surface only depends on the total amount of enclosed charge. All we need to do is add up the *net* charge inside the surface, since any charges outside the surface do not contribute to the flux. There are only three charges enclosed by the surface ... so:

$$\text{net charge} = 3C + 5C - 2C = 6C \quad (11.173)$$

The electric flux Φ_E is then just the enclosed charge divided by ϵ_0 , or $+6C/\epsilon_0$.

2.20 $-Q$. The charge $+Q$ on object *A* induces a negative charge $-Q$ on the inner surface of the conducting container *B*.

2.21 1.8 m to the left of the negative charge. By symmetry, we can figure out on which side the field should be zero. In between the two charges, the field from the positive and negative charges *add together*. The force on a fictitious positive test charge placed in between the two would experience a force to the left due to the positive charge, and another force to the left due to the negative charge. There is no way the fields can cancel here.

If we place a positive charge to the *right of the positive charge*, it will feel a force to the right from the positive charge, and a force to the left from the negative charge. The directions are opposite, but

the fields still cannot cancel because the test charge is closest to the larger charge.

This leaves us with points to the left of the negative charge. The forces on a positive test charge will be in opposite directions here, and we are closer to the smaller charge. What position gives zero field? First, we will call the position of the negative charge $x = 0$, which means the positive charge is at $x = 1$ m. We will call the position where electric field is zero x . The distance from this point to the negative charge is just x , and the distance to the positive charge is $1 + x$. Now write down the electric field due to each charge:

$$E_{\text{neg}} = \frac{k_e(-2.5 \mu\text{C})}{x^2} \quad (11.174)$$

$$E_{\text{pos}} = \frac{k_e(6 \mu\text{C})}{(1+x)^2} \quad (11.175)$$

The field will be zero when $E_{\text{neg}} + E_{\text{pos}} = 0$:

$$E_{\text{neg}} + E_{\text{pos}} = 0 \quad (11.176)$$

$$\frac{k_e(-2.5 \mu\text{C})}{x^2} + \frac{k_e(6 \mu\text{C})}{(1+x)^2} = 0 \quad (11.177)$$

$$\frac{\cancel{k_e}(-2.5 \mu\cancel{\text{C}})}{x^2} + \frac{\cancel{k_e}(6 \mu\cancel{\text{C}})}{(1+x)^2} = 0 \quad (11.178)$$

$$\frac{-2.5}{x^2} + \frac{6}{(1+x)^2} = 0 \quad (11.179)$$

$$\implies \frac{2.5}{x^2} = \frac{6}{(1+x)^2} \quad (11.180)$$

$$(11.181)$$

Cross multiply, apply the quadratic formula:

$$2.5(1+x)^2 = 6x^2 \quad (11.182)$$

$$2.5 + 5x + 2.5x^2 = 6x^2 \quad (11.183)$$

$$3.5x^2 - 5x - 2.5 = 0 \quad (11.184)$$

$$\implies x = \frac{-(-5) \pm \sqrt{5^2 - 4(-2.5)(3.5)}}{2(3.5)} \quad (11.185)$$

$$x = \frac{5 \pm \sqrt{25 + 35}}{7} \quad (11.186)$$

$$x = \frac{5 \pm 7.75}{7} = 1.82, -0.39 \quad (11.187)$$

Which root do we want? We wrote down the distance x the distance to the *left* of the negative charge. A negative value of x is then in the wrong direction, in between the two charges, which we already ruled out. The positive root, $x = 1.82$, means a distance 1.82 m to the *left* of the negative charge. This is what we want.

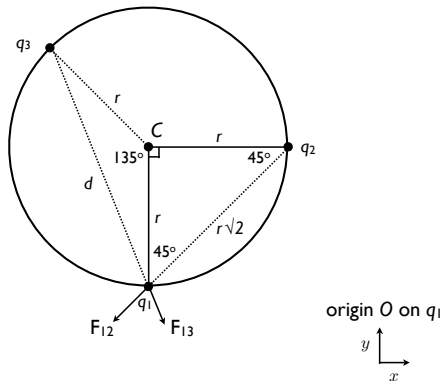
2.22 $0 \text{ N} \cdot \text{m}^2/\text{C}$. Remember that electric flux is $\Phi_E = EA \cos \theta$, where θ is the angle between a line perpendicular to the surface and the electric field. If E is *parallel* to the surface, then $\theta = 90$ and $\Phi_E = 0$.

Put more simply, there is only an electric flux if field lines penetrate the surface. If the field is parallel to the surface, no field lines penetrate, and there is no flux.

2.23 16 nN, down (-90°). The easiest way to solve this one is by symmetry and elimination. The negative charge q_2 feels an attractive force from both q_1 and q_3 . Since both charges are the same vertical distance away and below q_2 , both will give a force in the vertical downward direction of equal magnitude and direction. Since both charges are horizontally the same distance away but *on opposite sides*, the horizontal forces will be equal in magnitude but *opposite* in direction – the horizontal forces will cancel. Therefore, the net force has to be purely in the vertical direction and downward, so the second choice is the only option! Of course, you can calculate all of the forces by components and add them up ... you will arrive at the same answer.

2.24 The first thing we need to do is figure out the geometry and draw a picture. First, all three charges are confined to a circular track, which we will say has radius r . Two of the charges are the same, which we will call q_1 and q_2 , and they sit 90° apart on the circle. Where will the third, unequal charge (q_3) sit? In order for the forces on it due to charges 1 and 2 to be balanced, it must be equidistant from both on the circle. If charges 1 and 2 are 90° apart, then there are 270° left in the circle, and the third charge must sit halfway around that – the third charge must be 135° from both of the other charges.

Next, we should pick a coordinate system and origin. For reasons I hope will be clear soon, we will choose the origin to be on charge q_1 , with the $+y$ direction pointing toward the center of the circle and the x axis tangential to the circle, as shown below. We could have equally chosen q_2 as the origin, since it is identical to q_1 , it makes no difference.² For convenience, we label the center of the circle as point C so we can easily refer to it later.



Since charges q_1 and q_2 are 90° apart on the circle, we can form a 45-45-90 triangle with point C . Based on this, we can find the distance between q_1 and q_2 in terms of the radius r : $r_{12} = r\sqrt{2}$. Charges q_1 and q_2 are identical, and therefore experience a repulsive force of magnitude F_{12} directed along the line connecting them. This force must be at a 45° angle to the x and y axes, based on the geometry above. Charge q_3 has a different magnitude, but the same sign as q_1 , and thus the force between them F_{13} is also repulsive.

In order for the charges to stay in the positions above, what must be true? For charge q_1 , the forces in the y direction are irrelevant, since q_1 is constrained to stay on the circle anyway. Only net forces along the x direction will force it to move around the circle one way or the other. Thus, in order for this situation to be the equilibrium configuration, the forces in the x direction on q_1 must cancel. Since q_1 and q_2 are identical, the forces along the direction of the circle will also vanish for q_2 automatically. Finally, since the system is symmetric, q_3 must also have no net force along the direction of the circle if neither of the other charges do. Thus, it is sufficient to find the forces in the x direction for q_1 and equate them. This means we need to find the x components of F_{12} and F_{13} , set

² One could choose any point as the origin and get the same result, but in my opinion the geometry is more transparent in the present case.

them equal to one another, and solve for q_3 .

First, we focus on F_{12} , whose x component we will label $F_{12,x}$. We now know the distance between q_1 and q_2 , so the magnitude of the *total* force is easily written down with Coulomb's law:

$$F_{12} = \frac{k_e q_1 q_2}{r_{12}^2} = \frac{k_e q_1 q_2}{(r\sqrt{2})^2} = \frac{k_e q_1 q_2}{2r^2} \quad (11.188)$$

In order to find the x component, we just need to know the angle that \vec{F}_{12} makes with the x axis - 45° . You should be able to convince yourself this is true based on the geometry above (the inset to the second figure below may help). The x component is then just $F_{12,x} = F_{12} \sin 45^\circ$. Noting that $\sin 45^\circ = \sqrt{2}/2$:

$$F_{12,x} = F_{12} \sin 45^\circ = F_{12} \frac{\sqrt{2}}{2} = \frac{\sqrt{2} k_e q_1 q_2}{4r^2} \quad (11.189)$$

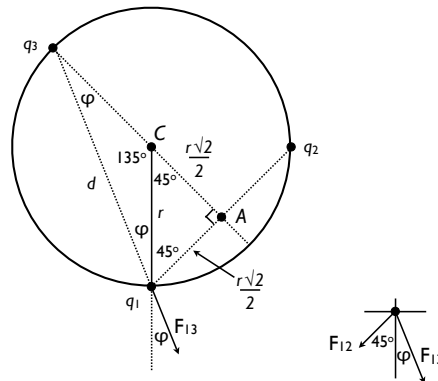
Now, what about the force between charges 1 and 3, F_{31} ? We can write down the force between them easily:

$$F_{13} = \frac{k_e q_1 q_3}{r_{13}^2} = \frac{k_e q_1 q_3}{d^2} \quad (11.190)$$

What is the distance d between q_1 and q_3 ? For this, we will need the law of cosines (and the fact that $\cos 135^\circ = -\sqrt{2}/2$):

$$d^2 = r^2 + r^2 - 2 \cdot r \cdot r \cdot \cos 135^\circ = 2r^2 - 2r^2 \left(-\frac{\sqrt{2}}{2} \right) = 2r^2 \left(1 + \frac{\sqrt{2}}{2} \right) \quad (11.191)$$

Before we combine that with our expression for F_{13} , let us find the x component, for which we need the angle that \vec{F}_{13} makes with our axes. The figure below will help us:



The triangle defined by q_1 , q_3 , and C gives us two equal angles φ . Since the angles of a triangle must add up to 180° , we must have $\varphi = (180^\circ - 135^\circ)/2 = 22.5^\circ$. This is the angle that \vec{F}_{13} makes with the y axis, and thus $F_{13,x} = F_{13} \sin \varphi$. The inset in the lower right of the figure should help you see this. If we look at the triangle formed by q_1 , q_3 , and point A , we can find $\sin \varphi$ analytically. Look at the φ nearest q_3 : $\sin \varphi = \frac{r\sqrt{2}/2}{d} = \frac{\sqrt{2}r}{2d}$. now we have everything to find $F_{13,x}$:

$$F_{13,x} = F_{13} \sin \varphi = \frac{k_e q_1 q_3}{d^2} \frac{r\sqrt{2}}{2d} = \frac{\sqrt{2} r k_e q_1 q_3}{2d^3} \quad (11.192)$$

Finally, we have the x components of both forces acting on q_1 . All we need to do now is equate them, and solve for q_3 :

$$F_{13,x} = F_{12,x} \quad (11.193)$$

$$\frac{\sqrt{2}k_e q_1 q_3}{2d^3} = \frac{\sqrt{2}k_e q_1 q_2}{4r^2} \quad (11.194)$$

$$\frac{\cancel{\sqrt{2}k_e q_1} q_3}{2d^3} = \frac{\cancel{\sqrt{2}k_e q_1} q_2}{4r^2} \quad (11.195)$$

$$\frac{r q_3}{d^3} = \frac{q_2}{2r^2} \quad (11.196)$$

$$\Rightarrow q_3 = \frac{q_2 d^3}{2r^3} \quad (11.197)$$

Plugging in our expression for d^2 we can find q_3 in terms of only q_2 and numerical factors:

$$q_3 = \frac{1}{2} \left(\frac{d}{r} \right)^3 q_2 \quad (11.198)$$

$$= \frac{1}{2} \left(\frac{d^2}{r^2} \right)^{\frac{3}{2}} q_2 \quad (11.199)$$

$$= \frac{1}{2} \left(\frac{2r^2 \left(1 + \frac{\sqrt{2}}{2} \right)}{r^2} \right)^{\frac{3}{2}} q_2 \quad (11.200)$$

$$= \frac{1}{2} \left(2 + \sqrt{2} \right)^{\frac{3}{2}} q_2 \approx 3.15 q_2 \quad (11.201)$$

Thus, the charge q_3 must be approximately 3.15 times as big as q_1 and q_2 in order for the latter two charges to be 90° apart. Physically, it makes sense that q_3 is bigger - q_1 and q_2 are closer together than they would be if all three charges are equal, so they must be feeling more repulsion from q_3 than from each other, which means q_3 must be bigger.

2.25 -0.0244 N . We are only interested in the horizontal component of the force, which makes things easier. First, we are trying to find the force on a negative charge due to two positive charges. Both positive charges are to the left of the negative charge, and both forces will be attractive. We will adopt the usual convention that the x axis lies horizontally, with the $+x$ direction to the right and the $-x$ direction to the left.

First, we will find the force on the negative charge due to the positive charge in the lower left, which we will call “1” to keep things straight. We will call the negative charge “2.” This is easy, since the force is purely in the $-x$ direction:

$$F_{x,1} = k_e \frac{q_1 q_2}{r_{12}^2} \quad (11.202)$$

$$= (9 \times 10^9 \text{ N} \cdot \text{m}^2 / \text{C}^2) \frac{(10^{-6} \text{ C}) \cdot (-2 \times 10^{-6} \text{ C})}{(1 \text{ m})^2} \quad (11.203)$$

$$= (9 \times 10^9 \text{ N} \cdot \cancel{\text{m}^2} / \cancel{\text{C}^2}) (-2 \times 10^{-12} \cancel{\text{C}^2} / \cancel{\text{m}^2}) \quad (11.204)$$

$$= -18 \times 10^{-3} \quad (11.205)$$

So far so good, but now we have to include the force from the upper left-hand positive charge, which we'll call “3.” We calculate the force in exactly the same way, with two little difference: the separation distance is slightly larger, and now the force has both a horizontal and vertical component. First, let's calculate the magnitude of the net force, we'll find the horizontal component after that.

Plane geometry tells us that the separation between charges 3 and 2 has to be $\sqrt{2} \cdot 1 \text{ m}$, or $\sqrt{2} \text{ m}$ – connecting the charges with straight lines forms a 1-1- $\sqrt{2}$ right triangle, with 45° angles.

$$F_{\text{net},3} = k_e \frac{q_2 q_3}{r_{23}^2} \quad (11.206)$$

$$= (9 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2) \frac{(10^{-6} \text{ C}) \cdot (-2 \times 10^{-6} \text{ C})}{(\sqrt{2} \text{ m})^2} \quad (11.207)$$

$$= (9 \times 10^9 \text{ N} \cdot \cancel{\text{m}^2}/\cancel{\text{C}^2}) \frac{-2 \times 10^{-12} \cancel{\text{C}^2}}{2 \cancel{\text{m}^2}} \quad (11.208)$$

$$= -9 \times 10^{-3} \text{ N} \quad (11.209)$$

So the *net* force from the upper left charge is just half as much, since it is a factor $\sqrt{2}$ farther away. We only want the horizontal component though! Since we are dealing with a 45-45-90 triangle here, the horizontal component is just the net force times $\cos 45^\circ$:

$$F_{x,3} = F_{\text{net},3} \cos 45^\circ \quad (11.210)$$

$$= -9 \times 10^{-3} \cdot \frac{\sqrt{2}}{2} \text{ N} = -9 \times 10^{-3} \cdot 0.707 \text{ N} \quad (11.211)$$

$$\approx -6.4 \times 10^{-3} \text{ N} \quad (11.212)$$

The total horizontal force is just the sum of the horizontal forces from the two positive charges:

$$F_{x,\text{total}} = F_{x,1} + F_{x,3} \quad (11.213)$$

$$= (-18 \times 10^{-3}) + (-6.4 \times 10^{-3}) \text{ N} \quad (11.214)$$

$$= -24.4 \times 10^{-3} \text{ N} = -0.0244 \text{ N} \quad (11.215)$$

2.26 $x = 0.77 \text{ m}$. We have two positive charges and one negative charge along a straight line. If we want there to be no net force on the negative charge, the electric forces from both of the positive charges on it must cancel. For that to happen, there is only one possibility: the negative charge has to be between the two positive charges. Outside that middle region, both positive charges will exert an attractive force on the negative charge in the same direction, and there is no way they can cancel each other. Only in the middle region do the forces from both positive charges act in opposite directions on a negative charge, and only there can they cancel each other. We want to find the position r_{23} such that both forces are equal in magnitude. All charges are on the x axis, so the problem is one-dimensional and does not require vectors.

Intuitively, we know that the negative charge q_3 must be closer to the smaller of the positive charges. Since electric forces get larger as separation decreases, the only way the force due to the larger charge can be the same as that due to the smaller charge is if *the negative charge is farther away from the larger charge*.

Let F_{32} be the force on q_3 due to q_2 , and F_{31} be the force on q_3 due to q_1 , and we will take the positive x direction to be to the right. Since both forces are repulsive, F_{32} acts in the $-x$ direction *and must therefore be negative*, while F_{31} acts in the $+x$ direction and is positive. This is only true for the region between the two positive charges! Elsewhere, both positive charges would give an attractive force, and there is no way they could cancel each other. We are not told about any other forces acting, so our force balance is this:

$$-F_{32} + F_{31} = 0 \quad \implies \quad F_{32} = F_{31} \quad (11.216)$$

It didn't really matter which one we called negative and which one we called positive, just that they have different signs. The separation between q_2 and q_3 is r_{23} , and the separation between q_1 and q_3 is then $2 - r_{23}$. Now we just need to down the electric forces. We will keep everything perfectly general, and plug in actual numbers at the end ... this is always safer.

$$F_{32} = F_{31} \quad (11.217)$$

$$\frac{k_e q_3 q_2}{r_{23}^2} = \frac{k_3 q_3 q_2}{(2 - r_{23})^2} \quad (11.218)$$

$$\frac{\cancel{k_e} q_3 q_2}{r_{23}^2} = \frac{\cancel{k_3} q_3 q_1}{(2 - r_{23})^2} \quad (11.219)$$

$$\frac{q_2}{r_{23}^2} = \frac{q_1}{(2 - r_{23})^2} \quad (11.220)$$

Note how this doesn't depend at all on the actual magnitude *or sign* of the charge in the middle! From here, there are two ways to proceed. We could cross-multiply, use the quadratic formula, and that would be that. On the other hand, since we know that q_3 is supposed to be between the other two charges, then r_{23} must be positive, and less than 2. That means that we can just take the square root of both sides of the equation above without problem, since neither side would be negative afterward.³ Using this approach first:

$$\frac{q_2}{r_{23}^2} = \frac{q_1}{(2 - r_{23})^2} \quad (11.221)$$

$$\Rightarrow \frac{\sqrt{q_2}}{r_{23}} = \frac{\sqrt{q_1}}{2 - r_{23}} \quad (11.222)$$

Now we can cross-multiply, and solve the resulting linear equation:

$$\sqrt{q_2} (2 - r_{23}) = \sqrt{q_1} r_{23} \quad (11.223)$$

$$2\sqrt{q_2} - \sqrt{q_2} r_{23} = \sqrt{q_1} r_{23} \quad (11.224)$$

$$2\sqrt{q_2} = (\sqrt{q_2} + \sqrt{q_1}) r_{23} \quad (11.225)$$

$$r_{23} = \frac{2\sqrt{q_2}}{\sqrt{q_2} + \sqrt{q_1}} \quad (11.226)$$

Plugging in the numbers we were given (and noting that all the units cancel):

$$r_{23} = \frac{2\sqrt{q_2}}{\sqrt{q_2} + \sqrt{q_1}} = \frac{2\sqrt{6\mu C}}{\sqrt{6\mu C} + \sqrt{15\mu C}} = \frac{2\sqrt{6}}{\sqrt{6} + \sqrt{15}} = \frac{2\sqrt{2}}{\sqrt{2} + \sqrt{5}} \approx 0.77 \text{ m} \quad (11.227)$$

For that very last step, we factored out $\sqrt{3}$ from the top and the bottom. An unnecessary step if you are using a calculator anyway, but we prefer to stay in practice.

The more general solution is to go back before we took the square root of both sides of the equation and solve it completely:

³ This would not work if we wanted the point to the left of q_2 .

$$\frac{q_2}{r_{23}^2} = \frac{q_1}{(2 - r_{23})^2} \quad (11.228)$$

$$q_2 (2 - r_{23})^2 = q_1 r_{23}^2 \quad (11.229)$$

$$q_2 (4 - 4r_{23} + r_{23}^2) = q_1 r_{23}^2 \quad (11.230)$$

$$(q_2 - q_1) r_{23}^2 - 4q_2 r_{23} + 4q_2 = 0 \quad (11.231)$$

$$(11.232)$$

Now we just have to solve the quadratic ...

$$r_{23} = \frac{4q_2 \pm \sqrt{(-4q_2)^2 - 4(q_2 - q_1) \cdot 4q_2}}{2(q_1 - q_2)} \text{ m} \quad (11.233)$$

$$= \frac{4 \cdot 6 \mu\text{C} \pm \sqrt{(-4 \cdot 6 \mu\text{C})^2 - 4(6 \mu\text{C} - 15 \mu\text{C}) \cdot 4 \cdot 6 \mu\text{C}}}{2(6 \mu\text{C} - 15 \mu\text{C})} \text{ m} \quad (11.234)$$

$$(11.235)$$

We can cancel all of the μC ...

$$r_{23} = \frac{24 \pm \sqrt{24^2 - 4(-9)(4)(6)}}{2(-9)} \text{ m} \quad (11.236)$$

$$= \frac{24 \pm \sqrt{24^2 + 36(24)}}{-18} \text{ m} \quad (11.237)$$

$$= \frac{-24 \mp \sqrt{1440}}{18} \text{ m} \quad (11.238)$$

$$= (0.775, -3.44) \text{ m} \quad (11.239)$$

Just as we expected: one solution ($r_{23} = 0.775 \text{ m}$) is right between the two charges, a little bit closer to the smaller charge. What about the positive solution? This corresponds to a position far away from both charges 3.44 m to the left of q_2 . As stated above, the forces act in the same direction outside of the middle region, and cannot cancel! This solution is physically impossible, just an artifact of the mathematics. We specified originally that the equations were only good for the middle region, so if we get an answer that falls outside we must discard it as outside the scope of our equations.

Our equations as we have written them do not take into account the fact that the fields change direction on one side of a charge versus the other. Properly speaking, outside the middle region between the positive charges, we should write $F_{32} = -F_{31}$ since the forces act in the same direction. Try repeating the problem starting there, and you will find that there are no real (non-imaginary) solutions outside the middle region - two positive forces cannot add up to zero.

Remember: in the end, we always need to make sure that the solutions are physically sensible in addition to being mathematically correct.

2.27 Two solid spheres, both of radius R , carry identical total charges, Q . One sphere is a good conductor while the other is an insulator. If the charge on the insulating sphere is uniformly distributed throughout its interior volume, how do the electric fields outside these two spheres compare? Are the fields identical inside the two spheres?

First off: if this problem didn't make any sense at all, you may want to re-read Sect. 2.8 covering Gauss' law, that is the key to the whole problem.

Outside of the two spheres, we have only to remember our key result from Gauss' law: the field from spherically symmetric charge distributions is equivalent to that of a point charge. The insulating sphere has a uniform charge distribution, and is therefore spherically symmetric. We know that for any isolated conductor, all excess charge must reside on the surface, and must be uniformly distributed, so the conducting sphere also has a spherically symmetric charge distribution. Since both have a spherically symmetric charge distribution and contain a total charge Q , outside the spheres at a distance r the electric field is the same for both, and the same as for a point charge Q :

$$E = \frac{k_e Q}{r^2} \quad r \geq R \quad \text{both spheres} \quad (11.240)$$

Inside the spheres, the two situations are qualitatively different. We know that inside any conductor in electrostatic equilibrium, the electric field is zero, since all excess charge resides on the surface.

$$E = 0 \quad r < R \quad \text{conducting sphere} \quad (11.241)$$

For the insulating sphere, this is not true. The charge is distributed over the whole volume. Based on Gauss' law, for any radius $r < R$, we know that only the charge that resides *inside* the radius r contributes to the electric field. The fraction of the total charge Q that resides inside a radius r is just the volume fraction of a sphere radius r to the total volume:

$$\text{fraction of charge within } r = Q \cdot \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi R^3} = Q \cdot \frac{r^3}{R^3} \quad r \leq R \quad (11.242)$$

This makes sense - when we are right at the radius of the sphere, $R=r$, we have the full charge Q , and when we are at the dead center ($r=0$) there is no charge effectively. The electric field inside the insulating sphere at radius r is just Coulomb's law for the fraction of charge within the radius r :

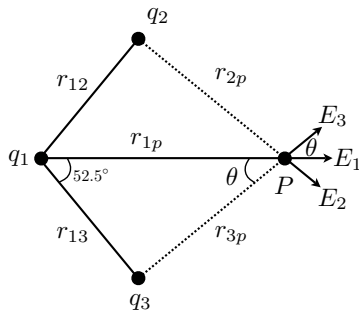
$$E = \frac{k_e Q}{r^2} \frac{r^3}{R^3} = \frac{k_e Q r}{R^3} \quad r \leq R \quad \text{insulating sphere} \quad (11.243)$$

Again, this is sensible - the electric field increases linearly from zero at the center of the sphere as more and more charge is outside the radius r .

Chapter 3 Problems

3.1 Decreases; stays the same. The capacitance of a parallel plate capacitor is $C = \frac{\epsilon_0 A}{d}$. If we pull the plates apart and increase the spacing d , the capacitance decreases. Nothing happens to the charges already on the plates if the capacitor is disconnected, though - they have no where to go!

3.2 First, we need to define the geometry of the situation a bit more clearly, and label things properly. Have a look:



Rather than worry about which nucleus is which, we will simply label the charges q_1 , q_2 , and q_3 and be as general as possible. We will also label the distances in a generic but self-explanatory way: the distance from charge 1 to charge 2 is r_{12} , the distance from charge 3 to the point P is r_{3p} , and so

on.

First, connect q_1 and P with a straight line. This is our x axis, and it nicely splits the problem into two symmetric halves. Since the bond angle was given as 105° , we know that the angle $\angle Pq_1q_3$ must be 52.5° , as must the angle $\angle Pq_1q_2$. The electric field due to charge 1 will clearly point directly along the x axis toward point P . The electric field due to charge 3 will make an angle θ with the x axis. Clearly, by symmetry, since $q_3 = q_2$ the electric fields from charges 2 and 3 will have the same x components, but equal and opposite y components - $E_{2x} = E_{3x}$, $E_{2y} = -E_{3y}$. Thus, the fields from charges 2 and 3 will in total have only an x component - so it is enough to compute only the x component of the field. And, since the x components are the same, we really only need to find one of them. In total, the field at P is then only composed of x components, and requires only two calculations:

$$\vec{E}_P = [E_{2x} + E_{3x} + E_1] \hat{x} = [2E_{3x} + E_1] \hat{x} \quad (11.244)$$

First, we can easily find E_1 , since we are told $r_{1p} = 1.2 \times 10^{-10}$ m:

$$E_1 = \frac{k_e q_1}{r_{1p}^2} \quad (11.245)$$

In order to find E_{3x} , we need two things: the angle θ , and the distance of charge 3 to point P , viz., r_{3p} . We can find the latter in terms of known quantities using the law of cosines⁴ on the triangle $\triangle q_1Pq_3$ with the 52.5° angle

$$r_{3p}^2 = r_{1p}^2 + r_{13}^2 - 2r_{13}r_{1p} \cos 52.5^\circ \approx 9.79 \times 10^{-11} \text{ m} \quad (11.246)$$

Once we have r_{3p} , we can find the angle θ by using the law of cosines on the same triangle, this time about the angle θ :

$$r_{13}^2 = r_{1p}^2 + r_{3p}^2 - 2r_{1p}r_{3p} \cos \theta \quad (11.247)$$

$$\Rightarrow \cos \theta = \frac{r_{1p}^2 + r_{3p}^2 - r_{13}^2}{2r_{1p}r_{3p}} \approx 0.631 \quad (11.248)$$

$$\Rightarrow \theta \approx 50.9^\circ \quad (11.249)$$

Once we have the angle and distance, we can easily find E_3 , and then its x component:

$$E_3 = \frac{k_3 q_3}{r_{3p}^2} \quad (11.250)$$

$$E_{3x} = E_3 \cos \theta = \frac{k_3 q_3}{r_{3p}^2} \cos \theta \quad (11.251)$$

Since the x component of the field from charge 2 is the same (and the y components of E_2 and E_3 cancel), we are ready to find the total field at point P :

$$\vec{E}_P = [2E_{3x} + E_1] \hat{x} = \left[2 \left(\frac{k_3 q_3}{r_{3p}^2} \cos \theta \right) + \frac{k_e q_1}{r_{1p}^2} \right] \hat{x} \approx [9.9 \times 10^{11} \text{ V/m}] \hat{x} \quad (11.252)$$

Now, when you get to the point of actually plugging in numbers, remember: the charge on a hydrogen nucleus, with a single proton, is $+e$, while that on an oxygen nucleus is $+8e$.

What about the potential at point P ? Far easier, no vectors! We have two charges a distance r_{3p} away, and one a distance r_{1p} away (again, we know that the contributions from charges 2 and 3 will be the same):

⁴ This is a very useful trick, worth reviewing if you have forgotten.

$$V_P = \frac{k_e q_1}{r_{1p}} + \frac{k_e q_2}{r_{2p}} + \frac{k_e q_3}{r_{3p}} = \frac{k_e q_1}{r_{1p}} + 2 \frac{k_e q_3}{r_{3p}} \approx 125 \text{ V} \quad (11.253)$$

3.3 This is probably another question most easily answered by elimination. In (a), the charges are clearly of the same magnitude, since the graph is perfectly symmetric, while in (b) the charges must be of different magnitude to explain the asymmetric graph. Therefore, the third answer cannot be correct.

In (a), the potential is constant along a vertical line separating the two charges (since there is a perfectly vertical line running halfway between the charges). This would only be true if they are of *opposite* signs. If the charges were of the same sign, there would be equipotential lines running horizontally from charge to charge. Similarly, the charges must also be of opposite sign in (b). This also rules out the first answer.

Based on similarity of (a) and (b), it must be that if (a) has charges of opposite magnitude, then so does (b). This also means that the fourth answer is out, which leaves only the second answer as a possibility. If you are still not clear on why the correct answer must be the second one, you may want to look carefully at the examples of equipotential lines in different situations presented in this chapter.

3.4 0. The charge is moved along the surface of the conductor, which is always at the same electric potential. Since the charge has moved through no net potential difference, no work has been done.

3.5 $KE_e = KE_p$. All of the potential energy gained by the proton and electron has to be converted into kinetic energy, and both particles lose the same potential energy by moving through the potential difference. Both particles have equal but opposite charges and move through equal and opposite potential differences – since the negatively charged electron moves through a positive potential difference, and the positively charged proton moves through a negative potential difference, the net loss of potential energy $q\Delta V$ is the same. Therefore, the amount of kinetic energy gained by each particle is the same. Since both particles started at rest, their resulting kinetic energies have to be the same. The *velocity* of the electron will be much greater, however, owing to its smaller mass – recall that kinetic energy is $\frac{1}{2}mv^2$.

3.6 $\frac{1}{2}C_0$. The capacitance of a parallel plate capacitor whose plates have an area A and a separation d is $C = \frac{\epsilon_0 A}{d}$. If we imagine the plates to be rectangular of length l and width w , the area A is $A = lw$. Let the capacitance of the capacitor be $C_0 = \frac{\epsilon_0 lw}{d}$ *before* dimensions are shrunk. Once we reduce the length, width, and separation by two times, we have:

$$C = \frac{\epsilon_0 \left(\frac{1}{2}l\right) \left(\frac{1}{2}w\right)}{\left(\frac{1}{2}d\right)} = \frac{\epsilon_0 \frac{1}{2}lw}{d} = \frac{1}{2}C_0 \quad (11.254)$$

It is easy to prove that if we chose, *e.g.*, circular plates, the answer would be the same – for any reasonable shape, the area goes down as the *square* of the dimensional decrease, while the separation just goes down as the factor itself.

3.7 4. Without the piece of glass, our capacitor has a value we'll call C . The charge stored on the capacitor is $Q = CV = 120C$ when the initial voltage is $V_{\text{initial}} = 120 \text{ V}$. The piece of glass acts as a dielectric, which increases the capacitance to κC (κ is always greater than 1).

Since the battery was disconnected, after inserting the piece of glass the total amount of charge Q stays the same – there is no source for additional charge to enter the capacitor. Now, however, the voltage V_{final} is less and the capacitance is more. We can set the initial amount of charge before inserting the glass equal to the final charge after inserting the glass, and solve for κ :

$$Q = CV_{\text{initial}} = \kappa CV_{\text{final}} \quad (11.255)$$

$$\mathcal{C}V_{\text{initial}} = \kappa \mathcal{C}V_{\text{final}} \quad (11.256)$$

$$V_{\text{initial}} = \kappa V_{\text{final}} \quad (11.257)$$

$$\kappa = \frac{V_{\text{initial}}}{V_{\text{final}}} = \frac{120}{30} = 4 \quad (11.258)$$

3.8 $8.00 \times 10^4 \text{ V/m}$. In a constant electric field, the electric field, potential difference and displacement are related by:

$$\Delta V = -|\vec{E}||\Delta \vec{x}| \cos \theta \quad (11.259)$$

Since the displacement and electric field are parallel everywhere, $\theta = 0$, and we have just $\Delta V = E\Delta x$. We have a potential difference $\Delta V = 2 \times 10^4 \text{ V}$ developed over a displacement of $\Delta x = 25 \text{ cm}$ (0.25 m). Plugging in the numbers:

$$\Delta V = -E\Delta x \quad (11.260)$$

$$2 \times 10^4 \text{ V} = -E(0.25 \text{ m}) \quad (11.261)$$

$$\Rightarrow E = -\frac{2 \times 10^4 \text{ V}}{0.25 \text{ m}} = -8.00 \times 10^4 \text{ V/m} \quad (11.262)$$

Since we want only the magnitude of the electric field, it is sufficient to write $8.00 \times 10^4 \text{ V/m}$.

3.9 $5.8 \times 10^{-19} \text{ J}$ The work done in moving a single charge through a constant electric field is given by:

$$W = qE_x\Delta x \quad (11.263)$$

where E_x is the component of the electric field parallel to the displacement. In this case, the displacement is always parallel to the electric field, so E_x is just the total field and Δx the displacement. Now we just plug in the numbers, remembering to put the displacement in meters:

$$W = qE\Delta x \quad (11.264)$$

$$= (1.6 \times 10^{-19} \text{ C})(240 \text{ N/C})(0.015 \text{ m}) \quad (11.265)$$

$$\approx 5.8 \times 10^{-19} \text{ N} \cdot \text{m} = 5.8 \times 10^{-19} \text{ J} \quad (11.266)$$

In the last line we used the fact that one Joule is defined to be one Newton times one meter.

3.10 2×10^{24} electrons. The energy required to charge the battery is just the amount that the potential energy of all the charges changes by. Each electron is moved through 9 V, which means each electron changes its potential energy by $-e \cdot 9 \text{ V}$, where e is the charge on one electron. The total potential energy is the potential energy per electron times the number of electrons, n . Basically, this is conservation of energy: the total energy into the battery has to equal the amount of energy to move one electron across 9 V times the number of electrons.

$$\Delta E_{in} + \Delta PE = 0 \quad (11.267)$$

$$3.6 \times 10^6 \text{ J} + n(-e \cdot 9 \text{ V}) = 0 \quad (11.268)$$

$$ne \cdot 9 \text{ V} = 3.6 \times 10^6 \text{ J} \quad (11.269)$$

$$n = \frac{3.6 \times 10^6 \text{ J}}{e \cdot 9 \text{ V}} \quad (11.270)$$

$$= \frac{3.6 \times 10^6 \text{ J}}{(1.6 \times 10^{-19} \text{ C})(9 \text{ V})} \quad (11.271)$$

$$= \frac{3.6 \times 10^6}{(1.6 \times 10^{-19})(9)} \quad (11.272)$$

$$\approx 2 \times 10^{24} \quad (11.273)$$

Here we make use of the fact that a coulomb times a volt is a joule. As usual, if you just use proper SI units throughout, the units will work out on their own.

3.11 $6.02 \mu\text{F}$. See page 91, this is the same capacitor layout!

3.12 $1.41 \times 10^5 \text{ m/s}$. When the proton is accelerated through a potential difference ΔV , it loses a potential energy of $e\Delta V$, which is converted into kinetic energy. We only need to apply conservation of energy, noting that the proton started at rest, and choosing our zero of potential energy such that the final potential energy is zero:

$$\begin{aligned} E_{\text{initial}} &= E_{\text{final}} \\ KE_{\text{initial}} + PE_{\text{initial}} &= KE_{\text{final}} + PE_{\text{final}} \\ 0 + q\Delta V &= \frac{1}{2}m_p v_f^2 + 0 \\ \Rightarrow v_f^2 &= \frac{2q\Delta V}{m_p} \\ v_f &= \sqrt{\frac{2q\Delta V}{m_p}} = \sqrt{\frac{2 \cdot 1.6 \times 10^{-19} \text{ C} \cdot 104 \text{ V}}{1.67 \times 10^{-27} \text{ kg}}} \\ &\approx 1.41 \times 10^5 [\text{C} \cdot \text{V}/\text{kg}]^{\frac{1}{2}} \\ &= 1.41 \times 10^5 [\text{J}/\text{kg}]^{\frac{1}{2}} = 1.41 \times 10^5 [\text{kg} \cdot \text{m}^2/\text{s}^2 \cdot \text{kg}]^{\frac{1}{2}} \\ &= 1.41 \times 10^5 \text{ m/s} \end{aligned}$$

The units are a bit tricky here, but remember that if you keep everything in proper SI units from the start, they will always work out ok. From the definition of electrical potential you know that one Volt is equal to one Joule per Coulomb, $1 \text{ V} = 1 \text{ J/C}$, and it then follows that $1 \text{ C} \cdot \text{V} = 1 \text{ J}$.

3.13 42.0 km/s . The proton starts from rest, and hence has no kinetic energy. It is accelerated by an electric field, and thus gains kinetic energy. The kinetic energy gained must come from the electric field. A charge q moving parallel to a constant electric field E over a distance Δx changes its potential energy by:

$$\Delta PE = qE\Delta x \quad (11.274)$$

The charge on a proton is just $+e$, and E and Δx are given. The change in kinetic energy is just the final kinetic energy of the proton, since it started from rest. The gain in kinetic energy must equal the change in potential energy:

$$\Delta PE = PE_{\text{initial}} - PE_{\text{final}} = -\Delta KE = -(KE_{\text{initial}} - KE_{\text{final}}) \quad (11.275)$$

$$eE\Delta x - 0 = -\left(0 - \frac{1}{2}m_p v_{\text{final}}^2\right) \quad (11.276)$$

$$eE\Delta x = \frac{1}{2}m_p v_{\text{final}}^2 \quad (11.277)$$

$$\Rightarrow v_{\text{final}}^2 = \frac{2eE\Delta x}{m_p} \quad (11.278)$$

$$v_{\text{final}} = \sqrt{\frac{2eE\Delta x}{m_p}} \quad (11.279)$$

Plugging in what we are given ...

$$v_{\text{final}} = \sqrt{\frac{2(1.6 \times 10^{-19} \text{ C})(8.36 \text{ V/m})(1.10 \text{ m})}{1.67 \times 10^{-27} \text{ kg}}} \quad (11.280)$$

$$\approx 42000 \sqrt{\text{C} \cdot \text{V}/\text{kg}} \quad (11.281)$$

$$= 42000 \sqrt{\text{J}/\text{kg}} \quad (11.282)$$

$$= 42000 \sqrt{\frac{\text{kg} \cdot \text{m}^2}{\text{s}^2 \cdot \text{kg}}} \quad (11.283)$$

$$= 42 \text{ km/s} \quad (11.284)$$

Making absolutely sure that the units work out, one should note that Coulombs times Volts is Joules, or $\text{kg} \cdot \text{m}^2/\text{s}^2$. If you always use proper SI units, it will work out though, and you won't have to remember lots of unit conversions.

3.14 The potential energy of a system of charges can be found by superposition, by adding together the potential energy of all *unique* pairs of charges. In this case, we have three distinct pairs of charges – (1,2), (1,3), and (2,3). The potential energy of the pair (1,2) is the electric potential that charge 2 feels due to charge 1, times charge 2:

$$PE_{(1,2)} = k_e q_2 \frac{q_1}{r_{12}} = k_e \frac{q_1 q_2}{r_{12}^2} \quad (11.285)$$

Here r_{12} is the separation between charges 1 and 2, or just 1.0 m in this case. We do the same for the other two pairs of charges, and add all three energies together (being very careful with signs):

$$PE_{\text{total}} = PE_{(1,2)} + PE_{(1,3)} + PE_{(2,3)} \quad (11.286)$$

$$= k_e \frac{q_1 q_2}{r_{12}} + k_e \frac{q_1 q_3}{r_{13}} + k_e \frac{q_2 q_3}{r_{23}} \quad (11.287)$$

$$= k_e \left(\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \quad (11.288)$$

$$= \left(9 \times 10^9 \frac{\text{N} \cdot \text{m}^2}{\text{C}^2} \right) \left[\frac{(-10^{-9} \text{ C})(10^{-9} \text{ C})}{1 \text{ m}} + \frac{(10^{-9} \text{ C})(10^{-9} \text{ C})}{3 \text{ m}} + \frac{(-10^{-9} \text{ C})(10^{-9} \text{ C})}{2 \text{ m}} \right] \quad (11.289)$$

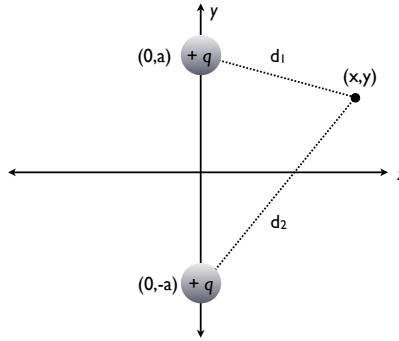
$$= \left(9 \times 10^9 \frac{\text{N} \cdot \text{m}^2}{\text{C}^2} \right) \left(\frac{10^{-18} \text{ C}^2}{\text{m}} \right) \left[-1 + \frac{1}{3} - \frac{1}{2} \right] \quad (11.290)$$

$$= (9 \times 10^9 \text{ N} \cdot \text{m}) \left[\frac{-7}{6} \right] \quad (11.291)$$

$$\approx -1.1 \times 10^{-8} \text{ J} \quad (11.292)$$

Here we used the fact that a $1\text{ J} \equiv 1\text{ N} \cdot \text{m}$.

3.15 $\frac{k_e q}{\sqrt{x^2 + (a-y)^2}} + \frac{k_e q}{\sqrt{x^2 + (a+y)^2}}$. For this one, it is perhaps easier to draw ourselves a picture:



We will label the upper charge 1, and the lower charge 2. The principle of superposition tells us that we only need to find the potential at point (x, y) due to each separately, and then add the results together. First, we focus on charge 1, located at $(0, a)$. First, we need the distance d_1 from charge 1 to the point (x, y) . The horizontal distance is just x , and the vertical distance has to be $a - y$. Therefore,

$$d_1 = \sqrt{x^2 + (a - y)^2} \quad (11.293)$$

The potential due the first charge, which we'll call V_1 is then found from Eq. 3.14:

$$V_1 = \frac{k_e q}{d_1} = \frac{k_e q}{\sqrt{x^2 + (a - y)^2}} \quad (11.294)$$

The potential due to the second charge at $(0, -a)$ is found in an identical manner, only noting that the vertical distance is now $a + y$:

$$d_2 = \sqrt{x^2 + (a + y)^2} \quad (11.295)$$

$$V_2 = \frac{k_e q}{d_2} = \frac{k_e q}{\sqrt{x^2 + (a + y)^2}} \quad (11.296)$$

Finally, since potential is a scalar quantity (it has only magnitude, not direction), the superposition principle tells us that the total electric potential at point (x, y) is just the sum of the individual potentials due to charges 1 and 2:

$$V_{\text{tot}} = V_1 + V_2 = \frac{k_e q}{\sqrt{x^2 + (a - y)^2}} + \frac{k_e q}{\sqrt{x^2 + (a + y)^2}} \quad (11.297)$$

Without resorting to approximations, there isn't really a much more aesthetically pleasing form for this one.

3.16 First of all, we should notice that the $7\mu\text{F}$ capacitor has nothing connected to its right wire, so it can't possibly be doing anything in this circuit. We can safely ignore it. Next, the $3\mu\text{F}$ and $14\mu\text{F}$ capacitors are simply in series, so we can readily find their equivalent capacitor:

$$C_{\text{eff}, 3\&14} = \frac{(3\mu\text{F})(14\mu\text{F})}{(3\mu\text{F}) + (14\mu\text{F})} \approx (2.65\mu\text{F}) \quad (11.298)$$

This $2.65\mu\text{F}$ effective capacitor is purely in parallel with the $6\mu\text{F}$ capacitor. We can therefore just add the two capacitances together and come up with an equivalent capacitance for the 3, 14, and $6\mu\text{F}$ capacitors:

$$C_{\text{eff},3,14,\&6} = C_{\text{eff},3\&14} + 6\mu\text{F} = 8.65\mu\text{F} \quad (11.299)$$

Finally, *that* equivalent capacitance is just in series with the $20\mu\text{F}$ capacitor, so the overall equivalent capacitance is readily found:

$$C_{\text{eff, total}} = \frac{C_{\text{eff},3,14,\&6} 20\mu\text{F}}{C_{\text{eff},3,14,\&6} + 20\mu\text{F}} \approx 6\mu\text{F} \quad (11.300)$$

3.17 Once again, we can simply use the principle of superposition. The total electric potential at any point is just the sum of the electric potentials due to each point charge. We'll label the charges 1-3 from left to right, and calculate the potential due to each first.

If we take an arbitrary point on the y axis $(0, y)$, what is the distance to charge 1? The vertical distance will always be just y , and the horizontal distance is just d . Therefore, the distance d_1 to the first charge is:

$$d_1 = \sqrt{d^2 + y^2} \quad (11.301)$$

The electric potential V_1 due to charge 1, $+Q$, is then found from Eq. 3.14:

$$V_1 = \frac{k_e Q}{d_1} = \frac{k_e Q}{\sqrt{d^2 + y^2}} \quad (11.302)$$

The distance to charge 2 is simply y , since it is also located on the y axis. The electric potential V_2 due to charge 2 is then:

$$V_2 = \frac{-2k_e Q}{y} \quad (11.303)$$

Finally, the distance to charge 3 is just the same as the distance to charge 1. Since both charges also have the same magnitude, $V_1 = V_3$. The total potential at a point $(0, y)$ is then just the sum of the potentials from all three individual charges:

$$V_{\text{tot}} = V_1 + V_2 + V_3 \quad (11.304)$$

$$= \frac{k_e Q}{d_1} = \frac{k_e Q}{\sqrt{d^2 + y^2}} + \frac{-2k_e Q}{y} + \frac{k_e Q}{d_1} = \frac{k_e Q}{\sqrt{d^2 + y^2}} \quad (11.305)$$

$$= \frac{2k_e Q}{\sqrt{d^2 + y^2}} + \frac{-2k_e Q}{y} \quad (11.306)$$

$$= 2k_e Q \left[\frac{1}{\sqrt{d^2 + y^2}} - \frac{1}{y} \right] \quad (11.307)$$

3.18 $-9\mathbf{nJ}$. The potential energy of a system of charges can be found by calculating the potential energy for every unique pair of charges and adding the results together. In this case, we have three unique pairings: charges 1 and 2, charges 2 and 3, and charges 1 and 3:

$$PE = PE_{1\&2} + PE_{2\&3} + PE_{1\&3} \quad (11.308)$$

$$= k_e \left(\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \quad (11.309)$$

Here r_{12} is the distance between charge 1 and 2, and so on. Since we have an equilateral triangle, all distances are 1 m. Since all charges are equal in magnitude, we can simplify this quite a bit once we plug in what we know - we just need to keep track of the signs of the charges:

$$\begin{aligned}
PE_{\text{total}} &= k_e \left(\frac{q_1 q_2}{r_{12}} + \frac{q_1 q_3}{r_{13}} + \frac{q_2 q_3}{r_{23}} \right) \\
&= \left(9 \times 10^9 \frac{\text{N} \cdot \text{m}^2}{\text{C}^2} \right) \left[\frac{(10^{-9} \text{C})(-10^{-9} \text{C})}{1 \text{m}} + \frac{(10^{-9} \text{C})(10^{-9} \text{C})}{1 \text{m}} + \frac{(-10^{-9} \text{C})(10^{-9} \text{C})}{1 \text{m}} \right] \\
&= \left(9 \times 10^9 \frac{\text{N} \cdot \text{m}^2}{\text{C}^2} \right) \left(\frac{10^{-18} \text{C}^2}{\text{m}} \right) [-1 + 1 - 1] \\
&= (9 \times 10^{-9} \text{N} \cdot \text{m}) [-1] \\
&\approx -9 \times 10^{-9} \text{J}
\end{aligned}$$

We used the conversion $1 \text{J} \equiv 1 \text{N} \cdot \text{m}$.

3.19 (a) Dielectric parallel to the plates: $C_{\text{eff}} = \frac{2K}{1+K} C$.

It is easiest to think of this as two capacitors in series, both with half the plate spacing - one filled with dielectric, one with nothing. First, without *any* dielectric, we will say that the original capacitor has plate spacing d and plate area A . The capacitance is then:

$$C_0 = \frac{\epsilon_0 A}{d} \quad (11.310)$$

The upper half capacitor with dielectric then has a capacitance:

$$C_d = \frac{K\epsilon_0 A}{d/2} = \frac{2K\epsilon_0 A}{d} = 2KC_0 \quad (11.311)$$

The half capacitor without then has

$$C_{\text{none}} = \frac{\epsilon_0 A}{d/2} = \frac{2\epsilon_0 A}{d} = 2C_0 \quad (11.312)$$

Now we just add the two like capacitors in series:

$$\frac{1}{C_{\text{eff}}} = \frac{1}{2KC_0} + \frac{1}{2C_0} \quad (11.313)$$

$$C_{\text{eff}} = \frac{4KC_0^2}{2KC_0 + 2C_0} \quad (11.314)$$

$$= \frac{2K}{1+K} C_0 \quad (11.315)$$

(b) Dielectric “perpendicular” to the plates: $C_{\text{eff}} = \frac{K+1}{2} C$.

In this case, we think of the half-filled capacitor as two capacitors in *parallel*, one filled with dielectric, one with nothing. Now each half capacitor has half the plate *area*, but the same spacing. The upper half capacitor with dielectric then has a capacitance:

$$C_d = \frac{K\epsilon_0 \frac{1}{2}A}{d} = \frac{K\epsilon_0 A}{2d} = \frac{1}{2}KC_0 \quad (11.316)$$

The half capacitor without then has

$$C_{\text{none}} = \frac{\epsilon_0 \frac{1}{2}A}{d} = \frac{\epsilon_0 A}{2d} = \frac{1}{2}C_0 \quad (11.317)$$

Now we just add our parallel capacitors:

$$C_{\text{eff}} = \frac{1}{2}KC_0 + \frac{1}{2}C_0 \quad (11.318)$$

$$= \frac{1}{2}(K+1)C_0 \quad (11.319)$$

$$= \frac{K+1}{2}C_0 \quad (11.320)$$

Chapter 4 Problems

4.1 The resistance of the conductor is $1500\,\Omega$. Using Ohm's Law:

$$R = \frac{\Delta V}{I} = \frac{1.5\text{ mV}}{1\,\mu\text{A}} = \frac{1.5 \times 10^{-3}\text{ V}}{1.0 \times 10^{-6}\text{ A}} = \frac{1.5 \times 10^{-3}\cancel{\text{V}}}{1.0 \times 10^{-6}\cancel{\text{A}}} \Omega = \frac{1.5}{10^{-3}} \Omega = 1.5 \times 10^3 \Omega$$

4.2 Diodes, insulators, and capacitors do not obey Ohm's law. A resistor by definition obeys Ohm's law. A normal conductor like copper also obeys Ohm's law. A diode has a non-linear $I-V$ relationship, and therefore does not obey Ohm's law. An insulator has no mobile charges, and cannot conduct current, so therefore does not obey Ohm's law. A capacitor also does not let a constant current pass through it, and does not obey Ohm's law.

4.3 A has the largest current. There are really only three rules to keep in mind: (1) a negative charge moving in one direction is the same thing as a positive charge moving in the opposite direction, (2) a positive and negative charge moving in the same direction cancel out, and (3) two charges of the same sign moving in the opposite direction cancel out. With that in mind ... In figure A, there are 3 positive charges moving to the right, and two negative charges moving to the left, the same as 5 positive charges moving to the right.

In B, four positive charges move to the left, which gives a *negative* current.

In C, two positive charges moving to the right and two negative charges moving to the left gives the same as four positive charges moving to the right.

In D, this is the same as two positive charges moving to the left, for a *negative* current.

Ranked from highest to lowest, we would have A, C, D, B.

4.4 Charges are already in the wire. When the circuit is completed, there is a rapid rearrangement of surface charges in the circuit. Turning on the light switch pushes charges in one end of the wire, and this displaces the charges already in the wire all along its length. The charges on the far end of the wire are pushed out as a result, and this is how current flows almost instantaneously – even though a single charge moves slowly, each charge pushes its neighbor further down the wire, and the *net* movement of charge occurs rapidly across the wire.

It is the same as turning on the hot water faucet in a way. Water comes out right away – the pipe is already filled with water. *Hot* water only comes out after some time, since it takes a while for water to go from the water heater to the faucet. Charges come out of the wire right away, but they are not the same charges entering the other end of the wire – the wire is already full of charge.

4.5 The drift speed increases as the cross section becomes smaller. We can relate current, area, and drift velocity using Eq. 4.5:

$$I = v_d n q A \quad \text{or} \quad v_d = \frac{I}{n q A}$$

This tells us that drift velocity scales inversely with the area, so if the area *decreases*, the drift velocity must *increase*. Again, it works the same way for water in pipes – the smaller the pipe, the higher the pressure and the larger the velocity.

4.6 First, we recall the relation between *current* and drift velocity:

$$I = nqAv_d$$

What we are really after is the resistance, however, which we can find with Ohm's law:

$$R = \frac{\Delta V}{I} = \frac{\Delta V}{nqAv_d} \propto \frac{1}{nv_d}$$

So the resistance is inversely proportional to the carrier density and drift velocity. Let's say the initial resistance is R_0 , and the resistance after changing n and v_d is just R . If we decrease the number of carriers by 100 times, the resistance goes *up* by 100 times. If we increase the drift velocity by 5 times, the resistance goes *down* by 5 times.

$$\begin{aligned} R_0 &\propto \frac{1}{nv_d} \\ R &\propto \frac{1}{\left(\frac{n}{100}\right)(5v_d)} = \frac{1}{\frac{nv_d}{20}} = \frac{20}{nv_d} \\ \implies R &= 20R_0 \end{aligned}$$

Even though we don't know what the actual resistance R_0 is, we can say that R is twenty times more. The one tricky step here is to write down the proper relationship between *resistance* and the given quantities, not just the relationship between *current* and the given quantities.

4.7 What we have to remember here is that grounding a point in circuit *defines its potential to be zero*, so $V_b = 0$. First, consider the resistor R . If there is a current I flowing through it from left to right, we know that the potential *difference* between points a and b must be $\Delta V_{ba} = V_b - V_a = -IR$. That is, the presence of a current I means that there is a *drop* of potential for charges going across the resistor. If we know that the potential at b is zero due to the ground point, $V_b = 0$, then in order to satisfy $\Delta V_{ba} = V_b - V_a = -IR$, we have to have $V_a = +IR$.

4.8 Half the resistance. Let's say the original resistance is R_0 , and the original wire has a length l_0 and radius r_0 . Since the material is the same, we can presume that the resistivity ρ is the same as well. The original resistance can be written in terms of the resistivity, length, and cross-sectional area ($A = \pi r_0^2$) of the wire:

$$R_0 = \frac{\rho l_0}{A} = \frac{\rho l_0}{\pi r_0^2}$$

The new wire, with every dimension doubled but the *same* resistivity ρ , has resistance:

$$R = \frac{\rho 2l_0}{\pi (2r_0)^2} = \frac{2\rho}{\pi} \frac{l_0}{4r_0^2} = \frac{1}{2} \frac{\rho l_0}{\pi r_0^2} = \frac{1}{2} R_0$$

4.9 $2.9 \times 10^{-4} \Omega \cdot \text{m}$. We first need to know the relation between resistivity and resistance, which includes the cross-sectional area of the wire A and its length l :

$$R = \frac{\rho l}{A} \quad \text{or} \quad \rho = \frac{RA}{l}$$

And then we add in the relation between current, voltage, and resistance, *viz.* $R = \Delta V / I$.

$$\rho = \frac{RA}{l} = \frac{\left(\frac{\Delta V}{I}\right)A}{l} = \frac{\Delta V \cdot A}{I \cdot l}$$

The wire is said to have a uniform radius, which can only be true if its cross section is circular. The area of the circular cross section is then just $A = \pi r^2$. Making sure we keep track of the units, we just plug everything in and run the numbers:

$$\rho = \frac{\Delta V \cdot A}{I \cdot l} = \frac{11 \text{ V} \cdot \pi (3.8 \times 10^{-3} \text{ m})^2}{0.45 \text{ A} \cdot 3.8 \text{ m}} = 2.9 \times 10^{-4} \frac{\text{V} \cdot \text{m}}{\text{A}} = 2.9 \times 10^{-4} \Omega \cdot \text{m}$$

4.10 3.95×10^{19} **electrons.** We know that each electron carries a charge of $-1.6 \times 10^{-19} \text{ C}$, so if we can figure out how much total charge has flowed through the bulb in 5 seconds, we can divide by the charge per electron to get the total number of electrons. First, we can calculate the amount of charge per second - the current - over the first 1.37 seconds from the given quantities:

$$I = \frac{\Delta Q}{\Delta t} = \frac{1.73 \text{ C}}{1.37 \text{ s}} = 1.26 \text{ C/s} = 1.26 \text{ A} \quad (11.321)$$

Next, we can find the total charge that passes in 5 seconds by rearranging the formula:

$$\Delta Q = I \Delta t = 1.26 \text{ A} \times 5.00 \text{ s} = 6.31 \text{ C} \quad (11.322)$$

Finally, we can divide the total charge by the charge per electron:

$$\# \text{ of electrons} = \frac{\text{total charge}}{\text{charge per electron}} = \frac{6.31 \text{ C}}{1.60 \times 10^{-19} \text{ C/electron}} = 3.95 \times 10^{19} \text{ electrons} \quad (11.323)$$

4.11 **4.23 A.** We know that Watts (W) are a unit of power, and that electrical power can be expressed as $\mathcal{P} = I \Delta V$. Since we know \mathcal{P} and ΔV , it is straightforward to find I , remembering that a Watt is the same as a Volt times an Ampere:

$$I = \frac{\mathcal{P}}{\Delta V} = \frac{550 \text{ W}}{130 \text{ V}} = 4.23 \text{ W/V} = 4.23 \text{ V} \cdot \text{A/V} = 4.23 \text{ A} \quad (11.324)$$

4.12 (a) about 16.7 years. First things first: to find out how long the electron takes to travel the length of the line, we need to know its velocity (since we already know the length). We can calculate drift velocity from the density of electrons, their individual charge, the current, and the cross-sectional area of the wire (noting that we are given the diameter, not the radius, and converting that to meters):

$$I = nq v_d A \quad \Rightarrow \quad v_d = \frac{I}{nqA} \quad (11.325)$$

$$v_d = \frac{1000 \text{ A}}{(8.20 \times 10^{28} \text{ e}^-/\text{m}^3)(1.60 \times 10^{-19} \text{ C/e}^-) \left[\pi \left(\frac{0.016 \text{ m}}{2} \right)^2 \right]} = 3.79 \times 10^{-4} \text{ m/s} \quad (11.326)$$

Here we used the fact that $1 \text{ A} = 1 \text{ C/s}$ to make the units come out properly. Next, given a velocity v_d and a distance d , we can calculate how long the journey takes:

$$\Delta t = \frac{d}{v_d} = \frac{200 \times 10^3 \text{ m}}{3.79 \times 10^{-4} \text{ m/s}} = 5.28 \times 10^8 \text{ s} \approx 16.7 \text{ yr} \quad (11.327)$$

(b) 7.5 MW. The power loss in the wire is most easily calculated from the current and resistance: $\mathcal{P} = I^2 R$. We can find the resistance of the whole wire from the length and the resistance per unit length:

$$R = (0.5 \Omega/\text{mi})(150 \text{ mi}) = 75 \Omega \quad (11.328)$$

Now we can readily calculate \mathcal{P} :

$$\mathcal{P} = I^2 R = (1000 \text{ A})^2 (75 \Omega) \quad (11.329)$$

$$= 7.5 \times 10^7 \text{ A}^2 \cdot \Omega = 7.5 \times 10^7 \text{ A}^2 \cdot \text{V/A} \quad (11.330)$$

$$= 7.5 \times 10^7 \text{ V} \cdot \text{A} = 7.5 \times 10^7 \text{ W} = 75 \text{ MW} \quad (11.331)$$

Here we used the conversions $1 \Omega = 1 \text{ V/A}$ and $1 \text{ W} = 1 \text{ V} \cdot \text{A}$.

Chapter 5 Problems

5.1 As low as possible. Power is delivered to the internal resistance of a battery, so decreasing the internal resistance will decrease this *lost* power and increase the percentage of the power delivered to the device.

5.2 Both are 500Ω . Call the two resistors R_1 and R_2 . Connected in series, their equivalent resistance is $R_1 + R_2 = 1000 \Omega$. Connected in parallel, their equivalent resistance is $1/R_1 + 1/R_2 = 250 \Omega$.

$$\begin{aligned} R_1 + R_2 &= 1000 \\ \frac{1}{R_1} + \frac{1}{R_2} &= \frac{1}{250} \\ \frac{1}{R_1} + \frac{1}{1000 - R_1} &= \frac{1}{250} \\ \frac{1000 - R_1 + R_1}{R_1(1000 - R_1)} &= \frac{1000}{R_1(1000 - R_1)} = \frac{1}{250} \\ 1000R_1 - R_1^2 &= (250)(1000) \\ R_1^2 - 1000R_1 + 250000 &= 0 \\ \Rightarrow R_1 &= 500 \Omega = R_2 \end{aligned}$$

5.3 1000Ω . See the text; this is the same example circuit.

5.4 The reading goes down. When the switch is closed, we have R_2 in parallel with a switch. Switches (ideally) have zero resistance, so all the current goes through the switch and none goes through R_2 – if we calculate the equivalent resistance between R_2 in parallel with zero, the equivalent resistance is still zero. Thus, the battery is connected effectively only to R_1 , and there is a current of:

$$I_{\text{closed}} = \frac{\Delta V}{R_1}$$

When the switch is opened, resistors R_1 and R_2 are now in series, so that the total circuit resistance is larger than when the switch was closed. As a result, the current decreases, since the applied voltage is the same in both cases. The total current is now:

$$I_{\text{open}} = \frac{\Delta V}{R_1 + R_2} < \frac{\Delta V}{R_1} = I_{\text{closed}}$$

No matter what R_1 and R_2 are, since resistances are always positive, the current has to be smaller when the switch is open.

5.5 522 mA. We know the resistance $R = 230 \Omega$, and the voltage $V = 230 \text{ Volts}$. We can get the current from Ohm's law:

$$\begin{aligned} R &= \frac{V}{I} \quad \Rightarrow \quad I = \frac{V}{R} \\ I &= \frac{230 \text{ V}}{230 \Omega} = 0.522 \frac{\text{V}}{\Omega} = 0.522 \text{ A} = 522 \text{ mA} \end{aligned}$$

5.6 $1.5\ \Omega$. We have 135 resistors in parallel R_1 through R_{135} , all of the same value. We know that the equivalent resistance must be:

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_{135}} = 135 \left(\frac{1}{R_1} \right) = \frac{135}{200} \quad (11.332)$$

So $R_{eq} = \frac{200}{135} \approx 1.5\ \Omega$.

5.7 $17.3\ \Omega$. First, note that you can combine the middle two resistors ($7\ \Omega$ and $11\ \Omega$) which are just in a simple parallel combination. The equivalent resistance for these two is:

$$\begin{aligned} \frac{1}{R_{eq, 7-11}} &= \frac{1}{7} + \frac{1}{11} = 0.234 \\ \Rightarrow R_{eq, 7-11} &= 4.28\ \Omega \end{aligned}$$

Now we have three resistors in *series* - $4\ \Omega$, $4.28\ \Omega$, and $9\ \Omega$. Resistors in series just add together, so the total equivalent resistance is:

$$R_{eq, total} = 4 + 4.28 + 9 = 17.28 \approx 17.3\ \Omega$$

5.8 $54\ \Omega$. Note that the $17\ \Omega$ resistor is only connected on one end, so it doesn't do anything! First, combine the $10\ \Omega$ and $15\ \Omega$ resistors in series to make $25\ \Omega$. This $25\ \Omega$ effective resistor is then in parallel with the $50\ \Omega$ resistor. Combining those two makes (approximately) $17\ \Omega$, which is now purely in series with the $37\ \Omega$ resistor. Adding those two together gives you, to two significant figures, $54\ \Omega$.

5.9 **2A**. After a long enough time, the capacitor will be completely charged. A current only flows in a capacitor while it is charging or discharging. Even during charging and discharging, the current steadily decreases with time until the capacitor is completely full or empty, respectively. Since the problem says "steady-state", we may assume that the capacitor is no longer charging – if it were, the current would not be steady, but decreasing, and after a long enough time, the capacitor should be fully charged anyway.

If the capacitor is fully charged and no current flows through it, then there is also no current through the $1\ \Omega$ resistor in series with it. If there is no current through the resistor either, then there is no voltage drop across it, and that whole branch of the circuit actually does nothing. Remember, if no current flows through a path in a circuit, it isn't doing anything except possibly storing energy. Portions of a circuit with no current can almost always be neglected when analyzing the rest of the circuit.

If the 1 mF - $1\ \Omega$ branch of the circuit can be neglected, then the only things left are a single 6 V battery, a $1\ \Omega$ resistor, and a $2\ \Omega$ resistor, all in series. Finding the current now is a simple matter, since the $1\ \Omega$ and $2\ \Omega$ resistors in series just make an equivalent resistance of $3\ \Omega$. Effectively, we have a single battery and resistor, for which we can easily calculate the current:

$$I = \frac{\Delta V}{R_{eq}} = \frac{6\text{ V}}{3\ \Omega} = 2\text{ A}$$

5.10 **$5=6>1=2=3=4$** . We only need to remember three things to figure this one out: (1) when a current encounters a junction, it splits up to take each path in amounts inversely proportional to the resistance of the path, (2) the current through a single loop of a circuit is the same everywhere, and (3) related to the last point, charge must be conserved, such that the same number of charges entering a wire have to leave it.

First, think about a current leaving the battery at point 5 and traveling clockwise around the circuit. The current reaches the junction leading to points 1 and 3, and must split up to take both paths. Since both paths have the same resistance (the resistors are equivalent, remember), the current will

split up equally between the two. Therefore, the current is the same at points 1 and 3.

The current in the path from 1-2 or 3-4 is in just a single wire, and the current can't change. Conservation of charge requires that every charge entering point 1 leaves through point 2 (and the same for points 3 and 4). Therefore, the currents at points 1 and 2 are equal, and so are those at points 3 and 4. Putting everything so far together, the current is the same at 1, 2, 3, and 4.

What about the currents at points 5 and 6? Conservation of charge again requires that the charges leaving the battery at 5 must eventually come back through point 6 – no charge can be gained or lost when going around the loop. Therefore, the currents at points 5 and 6 must be the same. Further, since the whole current leaving the battery at point 5 splits up into two separate (and equal) currents at points 1 and 3, the current at point 5 must be larger than the current at points 1 and 3. Therefore, overall the ranking from highest to lowest must be $5=6, 1=2=3=4$.

5.11 $4.91\ \Omega$. Each battery has an internal resistance that acts in *series*. Once the bulb is connected, it is also in series with both batteries. All we have two batteries and three resistors in series, nothing more. The sum of the voltage sources (the batteries) has to equal the sum of the voltage drops (current through the resistors) around the whole circuit - conservation of energy again. We know the current, the value of two of the resistors, and the voltages on the batteries. Let our unknown lamp resistance be r :

$$\text{total sources} = \text{total sinks} \quad (11.333)$$

$$1.6\text{ V} + 1.6\text{ V} = (0.6\text{ A})(0.151\ \Omega) + (0.6\text{ A})(0.270\ \Omega) + (0.6\text{ A})r \quad (11.334)$$

$$3.2\text{ V} = 0.0906\text{ V} + 0.162\text{ V} + (0.6\text{ A})r \quad (11.335)$$

$$(0.6\text{ A})r = 2.95\text{ V} \quad (11.336)$$

$$\Rightarrow r = 4.91\text{ V/A} = 4.91\ \Omega \quad (11.337)$$

5.12 0.67 A . Basically, all we need to remember is the relationship between power \mathcal{P} , current I , and voltage ΔV :

$$\mathcal{P} = I\Delta V$$

$$1\text{ W} = I(1.5\text{ V})$$

$$\Rightarrow I = \frac{1\text{ W}}{1.5\text{ V}} \approx 0.67\text{ A}$$

5.13 347 mA . In a preceding problem, we found the equivalent resistance of this circuit to be $17.3\ \Omega$. This single effective resistor is connected to a 6 V battery, so the current in the *effective* resistor has to be:

$$I_{\text{eq}} = \frac{\Delta V}{R_{\text{eq}}} = \frac{6\text{ V}}{17.3\ \Omega} \approx 0.347\text{ A} = 347\text{ mA} \quad (11.338)$$

Now, think about the circuit topology. The current through the equivalent resistor is the same as that through the $9\ \Omega$ resistor! If we work backwards from finding the overall equivalent resistor, the equivalent resistor decomposes into a composite of the 4 , 7 , and $11\ \Omega$ resistors and the $9\ \Omega$ resistor in series. Since two series resistors must both have the same current, they both have the same current as their equivalent resistance as well, and the current in the $9\ \Omega$ resistor must be 347 mA .

5.14 If we treat the battery as a perfect voltage source in series with its internal resistance, then the whole circuit under consideration is a perfect source of 9 V , a $1\ \Omega$ resistor, and a $10\ \Omega$ resistor all in *series*. The fact that they are all in series means they all have the same current. The internal resistance and the $10\ \Omega$ load resistance in series are equivalent to a single $11\ \Omega$ resistor, which means that effectively a perfect 9 V battery is connected to a single $11\ \Omega$ resistor. In that case, we can find the voltage across the $10\ \Omega$ resistor by first finding the current in the single loop of the circuit:

$$I = \frac{\Delta V}{R_{eq}} = \frac{9V}{11\Omega} \approx 0.818A$$

The voltage across the 10Ω resistor is then just given by Ohm's law:

$$\Delta V_{10\Omega} = I(10\Omega) \approx 8.18V$$

5.15 (a) is the only correct schematic. Remember: voltmeters have enormous internal resistances, and must be in *parallel* with what they are measuring. Ammeters have tiny internal resistances, and must be in *series* with what they are measuring. Based on this alone, (a) is the only correct schematic.

Circuit (b) is wrong because the ammeter is connected in parallel with the resistor. The ammeter's resistance is sufficiently low (zero, ideally) that it will 'steal' all of the current from the resistor instead of measuring it. The same effect could be had by just connecting a short-cut wire across the resistor – the ammeter effectively takes it out of the circuit by providing a far lower resistance path, such that little current will actually go through the resistor. The fact that a low equivalent resistance is connected to the battery means a large current will flow, quickly draining the battery. The voltmeter is connected correctly, but in this case it will basically only measure the voltage drop across the ammeter itself.

Circuit (c) is wrong because the ammeter is in series *and* the voltmeter is in parallel. The enormous resistance of the voltmeter (infinite, ideally) means that almost all of the battery's voltage will be dropped across the voltmeter itself, and almost none will be left for the ammeter and resistor. Since the ammeter effectively short-circuits the resistor anyway, this circuit will measure neither I nor ΔV correctly.

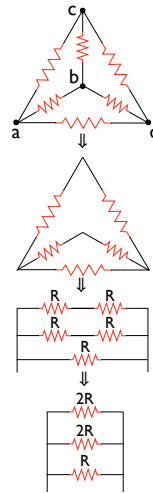
Circuit (d) is wrong because again the voltmeter is in series. The ammeter is correct, but the high resistance of the voltmeter will prevent all but the most miniscule currents from flowing anyway, so there will be nothing to measure!

5.16 Just by inspection, it is clear that we can't use our usual rules for parallel and series resistors to reduce this circuit - those simple rules cannot handle junctions with three branches. We can figure out how to reduce it by symmetry, however. Consider a current flowing out of point a below. The current must split up into three equal portions, since all three branches from point a are connected to the same resistance. Thus, the currents in branches ab, ac, and ad must be equal. If this is true, then the potential difference at points b, c, and d must be the same - at all three points, the same current has flowed from point a through the same resistance, so the voltage drops from a to c, a to b, and a to d are the same. This means that points b and c are at the same potential (as is point d), and there is no voltage across the resistor between points b and c. If there is no potential difference across the resistor bc, then there is no current by Ohm's law.

If there is no voltage difference and no current across the resistor between b and c, then it may be removed from the circuit - it isn't doing anything! If we take out that resistor, we have the second diagram above. If we rearrange this circuit - which we can always do, so long as no wires are broken in the process - then we see it is equivalent to the third diagram above. This circuit *does* allow a simple analysis based on series and parallel combinations.

First, combine the series resistors in the upper to branches. These add together to give two equivalent resistors of $2R$, which results in the simple parallel resistor diagram in the last panel above. These three parallel resistors give one equivalent resistance:

$$\begin{aligned} \frac{1}{R_{eq}} &= \frac{1}{2R} + \frac{1}{2R} + \frac{1}{R} = \frac{4}{2R} \\ \implies R_{eq} &= \frac{1}{2}R \end{aligned}$$



Thus, a regular tetrahedron of resistors can be replaced by a single resistor of half the value of the individual components. In this case, that means the equivalent resistance is 7Ω . If this 7Ω equivalent resistance is connected to a 9 V battery, the total current flowing out of the battery must be:

$$I = \frac{\Delta V}{R_{eq}} = \frac{9\text{ V}}{7\Omega} \approx 1.29\text{ A}$$

Incidentally, there is another way to analyze this circuit. Once we realize there is no potential difference between points b and c, rather than removing the resistor we could just connect those two points together. Since they are at the same voltage, this does not affect the operation of the circuit. You can verify for yourself that if you simply connect points b and c in the second diagram above, the same equivalent resistance results.

5.17 coming soon . . .

5.18 Put two 50Ω resistors in parallel, and connect that combination in series with a 20Ω resistor. There are many other ways, this is perhaps the simplest.

Chapter 6 Problems

6.1 Both the first and third are possible – they could be in the same direction or opposite directions, both would give zero force.

6.2 North.

6.3 coming soon . . .

6.4 First, the lack of a magnetic force when the proton moves north means that the magnetic field must be pointing either north or south – there is zero force only when velocity and magnetic field are parallel. The right-hand rule tells us that since the net force is upward, and the velocity is eastward, the magnetic field must be pointing north. (See Figure 6.3a.) The magnitude of the magnetic force is readily calculated from Equation 6.1:

$$|\vec{F}| = q|\vec{v}||\vec{B}|\sin\theta_{vB} = (1.6 \cdot 10^{-19}\text{ C})(1 \cdot 10^5\text{ m/s})(55 \times 10^{-6}\text{ T})\sin 90^\circ = 8.8 \cdot 10^{-19}\text{ N}$$

6.5 It shortens. The currents in neighboring spring coils are parallel, and therefore they attract each other.

6.6 No . . . coming soon

References

1. H. Muller, P. L. Stanwix, M. E. Tobar, E. Ivanov, P. Wolf, S. Herrmann, A. Senger, E. Kovalchuk, and A. Peters, "Tests of relativity by complementary rotating Michelson-Morley experiments," *Physical Review Letters*, vol. 99, no. 5, p. 050401, 2007.
2. http://en.wikipedia.org/wiki/Global_Positioning_System.
3. http://en.wikipedia.org/wiki/Hafele-Keating_experiment.
4. A. Einstein, *The meaning of relativity*. Princeton, New Jersey: Princeton University Press, 5 ed., 1988.
5. See <http://imdb.com/title/tt0116213/>.
6. http://en.wikipedia.org/wiki/Faraday_cage.
7. <http://bama.ua.edu/~lclavell/pages/>.
8. <http://bama.ua.edu/~jharrell/PH106-S06/vandegraaff.htm>.
9. R. J. van de Graaff, "Electrostatic Generator." Patent 1,991,236, 12 February, 1935. Filed 16 December, 1931. Patents are published as part of the terms of granting the patent to the inventor. Subject to limited exceptions reflected in 37 CFR 1.71(d) & (e) and 1.84(s), the text and drawings of a patent are typically not subject to copyright restrictions. In this case, no copyright reservations were stated.
10. K. T. Compton, L. C. V. Atta, and R. J. V. de Graaff, "Progress report on the M.I.T. high-voltage generator at Round Hill (typescript)," *MIT Office of the President Records*, vol. box 187, folder 5, 'Round Hill, 1932-1933', 1930-1959.
11. This photograph, from <http://flickr.com>, is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 2.0 license. It is the work of Tracy Lee Carroll (user StarrGazr on flickr.com). See <http://creativecommons.org/licenses/by-nc-nd/3.0/> for license details.
12. Crystal lattice images created with Jmol: an open-source Java viewer for chemical structures in 3D. <http://www.jmol.org/>.
13. This photograph, from <http://flickr.com>, is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 2.0 license. It is the work of user 'germanium' on flickr.com. See <http://creativecommons.org/licenses/by-nc-nd/3.0/> for license details.
14. http://en.wikipedia.org/wiki/Dielectric_constant.
15. C. Kittel, *Introduction to Solid State Physics*. New York: John Wiley and Sons, Inc., 7 ed., 1996.
16. L. Solymar and D. Walsh, *Lectures on the Electrical Properties of Materials*. Oxford: Oxford Science Publications, 4 ed., 1990.
17. <http://en.wikipedia.org/wiki/Resistivity>.
18. Image from <http://commons.wikimedia.org/wiki/Image:Mooninite2.jpg>. Permission is granted to copy, distribute and/or modify this image under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
19. Image in the public domain. From Practical Physics, publ. 1914 (Macmillan and Company).
20. <http://en.wikipedia.org/wiki/Pseudovector>.
21. F. J. Blatt, *Modern Physics*. McGraw-Hill, 1992.
22. Image from L. Keiner, <http://www.keiner.us/>. This image is licensed under the Creative Commons Attribution ShareAlike License v. 2.5 (<http://creativecommons.org/licenses/by-sa/2.5/>). You may use this image if attribution is given. Please notify the author of your use.
23. See http://en.wikipedia.org/List_of_indices_of_refraction. Image from <http://en.wikipedia.org/wiki/Image:Dispersion-curve.png>. The creator of this image has given permission to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.