Calculus-Based Physics II

by Jeffrey W. Schnick

This book is dedicated to Marie, Sara, and Natalie.

1 Charge & Coulomb's Law 2
2 The Electric Field: Description and Effect 9
3 The Electric Field Due to one or more Point Charges 14
4 Conductors and the Electric Field 24
5 Work Done by the Electric Field, and, the Electric Potential 32
6 The Electric Potential Due to One or More Point Charges 42
7 Equipotential Surfaces, Conductors, and Voltage 48
8 Capacitors, Dielectrics, and Energy in Capacitors 53
9 Electric Current, EMF, Ohm's Law 63
10 Resistors in Series and Parallel; Measuring *I* & *V* 70
11 Resistivity, Power 84
12 Kirchhoff's Rules, Terminal Voltage 90
13 RC Circuits 99
14 Capacitors in Series & Parallel 109
15 Magnetic Field Intro: Effects 115
16 Magnetic Field: More Effects 122
17 Magnetic Field: Causes 137
18 Faraday's Law, Lenz's Law 145
19 Induction, Transformers, and Generators 157
20 Faraday's Law and Maxwell's Extension to Ampere's Law 174
21 The Nature of Electromagnetic Waves 186
22 Huygens's Principle and 2-Slit Interference 192
23 Single-Slit Diffraction 211
24 Thin Film Interference 217
25 Polarization 223
26 Geometric Optics, Reflection 229
27 Refraction, Dispersion, Internal Reflection 237
28 Thin Lenses: Ray Tracing 243
29 Thin Lenses: Lens Equation, Optical Power 257
30 The Electric Field Due to a Continuous Distribution of Charge on a Line 267
31 The Electric Potential due to a Continuous Charge Distribution 279
32 Calculating the Electric Field from the Electric Potential 284
33 Gauss's Law 296
34 Gauss's Law Example 304
35 Gauss's Law for the Magnetic Field, and, Ampere's Law Revisited 309
36 The Biot-Savart Law 318
37 Maxwell's Equations 324

1 Charge & Coulomb's Law

Charge is a property of matter. There are two kinds of charge, positive "+" and negative "–".[1] An object can have positive charge, negative charge, or no charge at all. A particle which has charge causes a force-per-charge-of-would-be-victim vector to exist at each point in the region of space around itself. The infinite set of force-per-charge-of-would-be-victim vectors is called a vector field. Any charged particle that finds itself in the region of space where the force-per-charge-of-would-be-victim vector field exists will have a force exerted upon it by the force-per-charge-of-would-be-victim field. The force-per-charge-of-would-be-victim field is called the electric field. The charged particle causing the electric field to exist is called the source charge. (Regarding jargon: A charged particle is a particle that has charge. A charged particle is often referred to simply as "a charge.")

The source charge causes an electric field which exerts a force on the victim charge. The net effect is that the source charge causes a force to be exerted on the victim. While we have much to discuss about the electric field, for now, we focus on the net effect, which we state simply (neglecting the "middle man", the electric field) as, "A charged particle exerts a force on another charged particle." This statement is *Coulomb's Law* in its conceptual form. The force is called the *Coulomb force*, a.k.a. the *electrostatic force*.

Note that either charge can be viewed as the source charge and either can be viewed as the victim charge. Identifying one charge as the victim charge is equivalent to establishing a point of view, similar to identifying an object whose motion or equilibrium is under study for purposes of applying Newton's 2nd Law of motion, $\vec{\mathbf{a}} = \frac{\sum \vec{\mathbf{F}}}{m}$. In Coulomb's Law, the force exerted on one charged particle by another is directed along the line connecting the two particles, and, away from the other particle if both particles have the same kind of charge (both positive, or, both negative) but, toward the other particle if the kind of charge differs (one positive and the other negative). This fact is probably familiar to you as, "like charges repel and unlike attract."

The SI unit of charge is the coulomb, abbreviated C. One coulomb of charge is a lot of charge, so much that, two particles, each having a charge of +1 C and separated by a distance of 1 meter exert a force of 9×10^9 N, that is, 9 billion newtons on each other.

This brings us to the equation form of Coulomb's Law which can be written to give the magnitude of the force exerted by one charged particle on another as:

[1] It can be argued that, since the net charge on an object consisting of a bunch of particles, each of which has a positive amount of charge, and a bunch of particles, each of which has a negative amount of charge, is simply the algebraic sum of all the individual values of charge, there is really only one kind of charge and that it is the *value* of the charge of an object that can be either positive or negative. (See One Kind of Charge by John Denker, http://www.av8n.com/physics/one-kind-of-charge.htm .) However, it is conventional to refer to a negative amount of charge as an amount of negative charge and a positive amount of charge as an amount of positive charge. We use language consistent with that convention.

$$F = k\frac{|q_1||q_2|}{r^2} \tag{1-1}$$

where:

$k = 8.99\times10^9 \frac{\mathrm{N\cdot m^2}}{\mathrm{C^2}}$, a universal constant called the *Coulomb constant*,

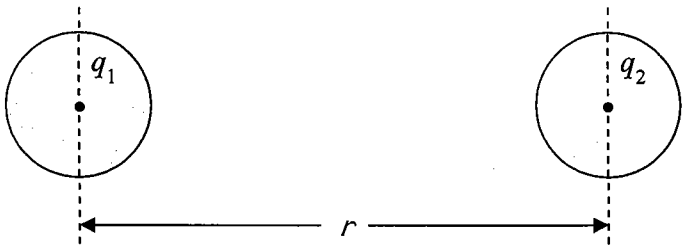
q_1 is the charge of particle 1,

q_2 is the charge of particle 2, and

r is the distance between the two particles.

The user of the equation (we are still talking about equation 1-1, $F = k\frac{|q_1||q_2|}{r^2}$) is expected to establish the direction of the force by means of "common sense" (the user's understanding of what it means for like charges to repel and unlike charges to attract each other).

While Coulomb's Law in equation form is designed to be exact for point particles, it is also exact for spherically symmetric charge distributions (such as uniform balls of charge) as long as one uses the center-to-center distance for r.



Coulomb's Law is also a good approximation in the case of objects on which the charge is not spherically symmetric as long as the objects' dimensions are small compared to the separation of the objects (the truer this is, the better the approximation). Again, one uses the separation of the centers of the charge distributions in the Coulomb's Law equation.

Coulomb's Law can be written in vector form as:

$$\vec{\mathbf{F}}_{12} = k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{12} \tag{1-2}$$

where:

$\vec{\mathbf{F}}_{12}$ is the force "of 1 on 2", that is, the force exerted by particle 1 on particle 2,

$\hat{\mathbf{r}}_{12}$ is a unit vector in the direction "from 1 to 2", and

k, q_1, and q_2 are defined as before (the Coulomb constant, the charge on particle 1, and the charge on particle 2 respectively).

Note the absence of the absolute value signs around q_1 and q_2. A particle which has a certain amount, say, 5 coulombs of the negative kind of charge is said to have a charge of –5 coulombs and one with 5 coulombs of the positive kind of charge is said to have a charge of +5 coulombs) and indeed the plus and minus signs designating the kind of charge have the usual arithmetic meaning when the charges enter into equations. For instance, if you create a composite object by combining an object that has a charge of $q_1 = +3$ C with an object that has a charge of $q_2 = -5$ C, then the composite object has a charge of

$$q = q_1 + q_2$$

$$q = +3 \text{ C} + (-5 \text{ C})$$

$$q = -2 \text{ C}$$

Note that the arithmetic interpretation of the kind of charge in the vector form of Coulomb's Law causes that equation to give the correct direction of the force for any combination of kinds of charge. For instance, if one of the particles has positive charge and the other negative, then the value of the product $q_1 q_2$ in equation 1-2

$$\vec{\mathbf{F}}_{12} = k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{12}$$

has a negative sign which we can associate with the unit vector. Now $-\hat{\mathbf{r}}_{12}$ is in the direction opposite "from 1 to 2" meaning it is in the direction "from 2 to 1." This means that $\vec{\mathbf{F}}_{12}$, the force of 1 on 2, is directed toward particle 1. This is consistent with our understanding that opposites attract. Similarly, if q_1 and q_2 are both positive, or both negative in $\vec{\mathbf{F}}_{12} = k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{12}$ then the value of the product $q_1 q_2$ is positive meaning that the direction of the force of 1 on 2 is $\hat{\mathbf{r}}_{12}$ (from 1 to 2), that is, away from 1, consistent with the fact that like charges repel.

We've been talking about the force of 1 on 2. Particle 2 exerts a force on particle 1 as well. It is given by $\vec{\mathbf{F}}_{21} = k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{21}$. The unit vector $\hat{\mathbf{r}}_{21}$, pointing from 2 to 1, is just the negative of the unit vector pointing from 1 to 2:

$$\hat{\mathbf{r}}_{21} = -\hat{\mathbf{r}}_{12}$$

If we make this substitution into our expression for the force exerted by particle 2 on particle 1, we obtain:

$$\vec{\mathbf{F}}_{21} = k\frac{q_1 q_2}{r^2}(-\hat{\mathbf{r}}_{12})$$

$$\vec{\mathbf{F}}_{21} = -k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{12}$$

Comparing the right side with our expression for the force of 1 on 2 (namely, $\vec{\mathbf{F}}_{12} = k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{12}$), we see that

$$\vec{\mathbf{F}}_{21} = -\vec{\mathbf{F}}_{12}.$$

So, according to Coulomb's Law, if particle 1 is exerting a force $\vec{\mathbf{F}}_{12}$ on particle 2, then particle 2 is, at the same time, exerting an equal but opposite force $-\vec{\mathbf{F}}_{12}$ back on particle 2, which, as we know, by Newton's 3rd Law, it must.

In our macroscopic[2] world we find that charge is not an inherent fixed property of an object but, rather, something that we can change. Rub a neutral rubber rod with animal fur, for instance, and you'll find that afterwards, the rod has some charge and the fur has the opposite kind of charge. Ben Franklin defined the kind of charge that appears on the rubber rod to be negative charge and the other kind to be positive charge. To provide some understanding of how the rod comes to have negative charge, we delve briefly into the atomic world and even the subatomic world.

The stable matter with which we are familiar consists of protons, neutrons, and electrons. Neutrons are neutral, protons have a fixed amount of positive charge, and electrons have the same fixed amount of negative charge. Unlike the rubber rod of our macroscopic world, you cannot give charge to the neutron and you can neither add charge to, nor remove charge from, either the proton or the electron. Every proton has the same fixed amount of charge, namely 1.60×10^{-19} C. Scientists have never been able to isolate any smaller amount of charge. That amount of charge is given a name. It is called the e, abbreviated e and pronounced "ee". The e is a non-SI unit of charge. As stated $1\,\mathrm{e} = 1.60\times10^{-19}$ C. In units of e, the charge of a proton is 1 e (exactly) and the charge of an electron is -1 e. For some reason, there is a tendency among humans to interpret the fact that the unit the e is equivalent to 1.60×10^{-19} C to mean that 1 e equals -1.60×10^{-19} C. This is wrong! Rather,

$$1\,\mathrm{e} = 1.60\times10^{-19}\ \mathrm{C}.$$

A typical neutral atom consists of a nucleus made up of neutrons and protons surrounded by orbiting electrons such that the number of electrons in orbit about the nucleus is equal to the number of protons in the nucleus. Let's see what this means in terms of an everyday object such as a polystyrene cup. A typical polystyrene cup has a mass of about 2 grams. It consists of roughly: 6×10^{23} neutrons, 6×10^{23} protons, and, when neutral, 6×10^{23} electrons. Thus, when neutral it has about 1×10^{5} C of positive charge and 1×10^{5} C of negative charge, for a total of 0 charge. Now if you rub a polystyrene cup with animal fur you can give it a noticeable charge. If you rub it all over with the fur on a dry day and then experimentally determine the charge on the cup, you will find it to be about -5×10^{-8} C. This represents an increase of about 0.00000000005 % in the number of electrons on the cup. They were

[2] Macroscopic means "of a size that we can see with the naked eye." It is to be contrasted with microscopic (you need a light microscope to see it), atomic (of or about the size of an atom), and subatomic (smaller than an atom, e.g. about the size of a nucleus of an atom).

transferred from the fur to the cup. We are talking about 3×10^{10} electrons, which sure would be a lot of marbles but represents a minuscule fraction of the total number of electrons in the material of the cup.

The main points of the preceding discussion are:

- A typical neutral macroscopic object consists of incredibly huge amounts of both kinds of charge (about 50 million coulombs of each for every kilogram of matter), the same amount of each kind.
- When we charge an object, we transfer a relatively minuscule amount of charge to or from that object.
- A typical everyday amount of charge (such as the amount of charge on a clingy sock just out of the dryer) is 10^{-7} coulombs.
- When we transfer charge from one object to another, we are actually moving charged particles, typically electrons, from one object to the other.

One point that we did not make in the discussion above is that *charge is conserved.* For instance, if, by rubbing a rubber rod with fur, we transfer a certain amount of negative charge to the rubber rod, then, the originally-neutral fur is left with the exact same amount of positive charge. Recalling the exact balance between the incredibly huge amount of negative charge and the incredibly huge amount of positive charge in any macroscopic object, we recognize that, in charging the rubber rod, the fur becomes positively charged not because it somehow gains positive charge, but, because it loses negative charge, meaning that the original incredibly huge amount of positive charge now (slightly) exceeds the (still incredibly huge) amount of negative charge remaining on and in the fur.

Charging by Rubbing

One might well wonder why rubbing a rubber rod with animal fur would cause electrons to be transferred from the fur to the rod. If one could imagine some way that even one electron might, by chance, find its way from the fur to the rod, it would seem that, then, the rod would be negatively charged and the fur positively charged so that any electron that got free from the fur would be attracted back to the fur by the positive charge on it and repelled by the negative charge on the rod. So why would any more charge ever be transferred from the fur to the rod? The answer comes under the heading of "distance matters." In rubbing the rod with the fur you bring lots of fur molecules very close to rubber molecules. In some cases, the outer electrons in the atoms of the fur come so close to nuclei of the atoms on the surface of the rubber that the force of attraction of these positive nuclei is greater than the force of attraction of the nucleus of the atom of which they are a part. The net force is then toward the rod, the electrons in question experience acceleration toward the rod that changes the velocity such that the electrons move to the rod. Charging by rubbing depends strongly on the molecular structure of the materials in question. One interesting aspect of the process is that the rubbing only causes lots of molecules in the fur to come very close to molecules in the rubber. It is not as if the energy associated with the rubbing motion is somehow given to the electrons causing them to jump from the fur to the rubber. It should be noted that fur is not the only material that has a tendency to give up electrons and rubber is not the only material with a tendency to acquire them. The phenomenon

of charging by rubbing is called triboelectrification. The following ordered list of the tendency of (a limited number of) materials to give up or accept electrons is called a *triboelectric sequence*:

Increasing tendency to take on electrons →

Air	Rabbit Fur	Glass	Wool	Silk	Steel	Rubber	Polyester	Styrofoam	Vinyl	Teflon

← Increasing tendency to give up electrons

The presence and position of air on the list suggests that it is easier to maintain a negative charge on objects in air than it is to maintain a positive charge on them.

Conductors and Insulators

Suppose you charge a rubber rod and then touch it to a neutral object. Some charge, repelled by the negative charge on the rod, will be transferred to the originally-neutral object. What happens to that charge then depends on the material of which the originally-neutral object consists. In the case of some materials, the charge will stay on the spot where the originally neutral object is touched by the charged rod. Such materials are referred to as insulators, materials through which charge cannot move, or, through which the movement of charge is very limited. Examples of good insulators are quartz, glass, and air. In the case of other materials, the charge, almost instantly spreads out all over the material in question, in response to the force of repulsion (recalling that force causes acceleration which leads to the movement) that each elementary particle of the charge exerts on every other elementary particle of charge. Materials in which the charge is free to move about are referred to as conductors. Examples of good conductors are metals and saltwater.

When you put some charge on a conductor, it immediately spreads out all over the conductor. The larger the conductor, the more it spreads out. In the case of a very large object, the charge can spread out so much that any chunk of the object has a negligible amount of charge and hence, behaves as if were neutral. Near the surface of the earth, the earth itself is large enough to play such a role. If we bury a good conductor such as a long copper rod or pipe, in the earth, and connect to it another good conductor such as a copper wire, which we might connect to another metal object, such as a cover plate for an electrical socket, above but near the surface of the earth, we can take advantage of the earth's nature as a huge object made largely of conducting material. If we touch a charged rubber rod to the metal cover plate just mentioned, and then withdraw the rod, the charge that is transferred to the metal plate spreads out over the earth to the extent that the cover plate is neutral. We use the expression "the charge that was transferred to the cover plate has flowed into the earth." A conductor that is connected to the earth in the manner that the cover plate just discussed is connected is called "ground." The act of touching a charged object to ground is referred to as grounding the object. If the object itself is a conductor, grounding it (in the absence of other charged objects) causes it to become neutral.

Charging by Induction

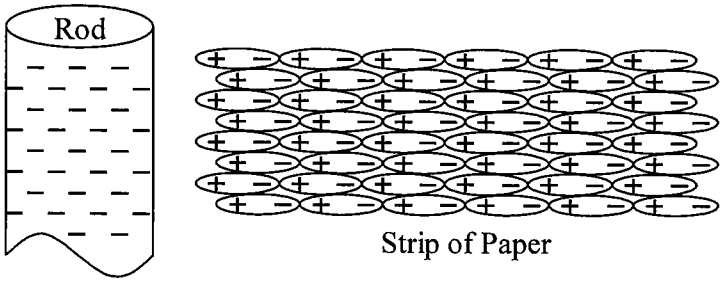
If you hold one side of a conductor in contact with ground and bring a charged object very near the other side of the conductor, and then, keeping the charged object close to the conductor without touching it, break the contact of the conductor with ground, you will find that the conductor is charged with the opposite kind of the charge that was originally on the charged object. Here's why. When you bring the charged object near the conductor, it repels charge in the conductor right out of the conductor and into the earth. Then, with those charges gone, if you break the path to ground, the conductor is stuck with the absence of those charged particles that were repelled into the ground. Since the original charged object repels the same kind of charge that it has, the conductor is left with the opposite kind of charge.

Polarization

Let's rub that rubber rod with fur again and bring the rubber rod near one end of a small strip of neutral aluminum foil. We find that the foil is attracted to the rubber rod, even though the foil remains neutral. Here's why:

The negatively charged rubber rod repels the free-to-move negative charge in the strip to the other end of the strip. As a result, the near end of the aluminum strip is positively charged and the far end is negatively charged. So, the rubber rod attracts the near end of the rod and repels the far end. But, because the near end is nearer, the force of attraction is greater than the force of repulsion and the net force is toward the rod. The separation of charge that occurs in the neutral strip of aluminum is called polarization, and, when the neutral aluminum strip is positive on one end and negative on the other, we say that it is polarized.

Polarization takes place in the case of insulators as well, despite the fact that charge is not free to move about within an insulator. Let's bring a negatively-charged rod near one end of a piece of paper. Every molecule in the paper has a positive part and a negative part. The positive part is attracted to the rod and the negative part is repelled. The effect is that each molecule in the paper is polarized and stretched. Now, if every bit of positive charge gets pulled just a little bit closer to the rod and every bit of negative charge gets pushed a little farther away, the net effect in the bulk of the paper is to leave it neutral, but, at the ends there is a net charge. On the near end, the repelled negative charge leaves the attracted positive charge all by itself, and, on the far end, the attracted positive charge leaves the repelled negative charge all by itself.



As in the case of the aluminum strip, the negative rubber rod attracts the near, positive, end and repels the far, negative, end, but, the near end is closer so the attractive force is greater, meaning that the net force on the strip of paper is attractive. Again, the separation of the charge in the paper is called polarization and the fact that one end of the neutral strip of paper is negative and the other is positive means that the strip of paper is polarized.

2 The Electric Field: Description and Effect

An electric field is an invisible entity[1] which exists in the region around a charged particle. It is caused to exist by the charged particle. The effect of an electric field is to exert a force on any charged particle (other than the charged particle causing the electric field to exist) that finds itself at a point in space at which the electric field exists. The electric field at an empty point in space is the force-per-charge-of-would-be-victim at that empty point in space. The charged particle that is causing the electric field to exist is called a source charge. The electric field exists in the region around the source charge whether or not there is a victim charged particle for the electric field to exert a force upon. At every point in space where the electric field exists, it has both magnitude and direction. Hence, the electric field is a vector at each point in space at which it exists. We call the force-per-charge-of-would-be-victim vector at a particular point in space the "electric field" at that point. We also call the infinite set of all such vectors, in the region around the source charge, the electric field of the source charge. We use the symbol $\vec{\mathbf{E}}$ to represent the electric field. I am using the word "victim" for any particle upon which an electric field is exerting a force. The electric field will only exert a force on a particle if that particle has charge. So all "victims" of an electric field have charge. If there does happen to be a charged particle in an electric field, then that charged particle (the victim) will experience a force

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}} \tag{2-1}$$

where q is the charge of the victim and $\vec{\mathbf{E}}$ is the electric field vector at the location of the victim. We can think of the electric field as a characteristic of space. The force experienced by the victim charged particle is the product of a characteristic of the victim (its charge) and a characteristic of the point in space (the electric field) at which the victim happens to be.

The electric field is not matter. It is not "stuff." It is not charge. It has no charge. It neither attracts nor repels charged particles. It cannot do that because its "victims", the charged particles upon which the electric field exerts force, are within it. To say that the electric field attracts or repels a charged particle would be analogous to saying that the water in the ocean attracts or repels a submarine that is submerged in the ocean. Yes, the ocean water exerts an upward buoyant force on the submarine. But, it neither attracts nor repels the submarine. In like manner, the electric field never attracts nor repels any charged particles. It is nonsense to say that it does.

If you have two source charge particles, e.g. one at point A and another at point B, each creating its own electric field vector at one and the same point P, the actual electric field vector at point P is the vector sum of the two electric field vectors. If you have a multitude of charged particles contributing to the electric field at point P, the electric field at point P is the vector sum of all the electric field vectors at P. Thus, by means of a variety of source charge distributions, one can create a wide variety of electric field vector sets in some chosen region of space. In the next chapter, we discuss the relation between the source charges that cause an electric field to exist,

[1] English rather than physics: An entity is something that exists. I use the word "entity" here rather than "thing" or "substance" because either of these words would imply that we are talking about matter. The electric field is not matter.

and the electric field itself. In this chapter, we focus our attention on the relation between an existing electric field (with no concern for how it came to exist) and the effect of that electric field on any charged particle in the electric field. To do so, it is important for you to be able to accept a given electric field as specified, without worrying about how the electric field is caused to exist in a region of space. (The latter is an important topic which we deal with at length in the next chapter.)

Suppose for instance that at a particular point in an empty region in space, let's call it point P, there is an eastward-directed electric field of magnitude 0.32 N/C. Remember, initially, we are talking about the electric field at an empty point in space. Now, let's imagine that we put a particle that has +2.0 coulombs of charge at point P. The electric field at point P will exert a force on our 2.0 C victim:

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}}$$

$$\vec{\mathbf{F}} = 2.0\,\mathrm{C}\,(0.32\frac{\mathrm{N}}{\mathrm{C}}\ \text{eastward})$$

Note that we are dealing with vectors so we did include both magnitude and direction when we substituted for $\vec{\mathbf{E}}$. Calculating the product on the right side of the equation, and including the direction in our final answer yields:

$$\vec{\mathbf{F}} = 0.64\,\mathrm{N}\ \text{eastward}$$

We see that the force is in the same direction as the electric field. Indeed, the point I want to make here is about the direction of the electric field: The electric field at any location is defined to be in *the direction of the force that the electric field would exert on a positively charged victim* if there was a positively charged victim at that location.

Told that there is an electric field in a given empty region in space and asked to determine its direction at the various points in space at which the electric exists, what you should do is to put a single positively-charged particle at each of the various points in the region in turn, and find out which way the force that the particle experiences at each location is directed. Such a positively-charged particle is called a positive *test charge*. At each location you place it, the direction of the force experienced by the positive test charge is the direction of the electric field at that location.

Having defined the electric field to be in the direction of the force that it would exert on a *positive* test charge, what does this mean for the case of a *negative* test charge? Suppose that, in the example of the empty point in space at which there was a 0.32 N/C eastward electric field, we place a particle with charge –2.0 coulombs (instead of +2.0 coulombs as we did before). This particle would experience a force:

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}}$$

$$\vec{\mathbf{F}} = -2.0\,\mathrm{C}\,(0.32\frac{\mathrm{N}}{\mathrm{C}}\ \text{eastward})$$

$$\vec{\mathbf{F}} = -0.64\,\mathrm{N}\ \text{eastward}$$

A negative eastward force is a positive westward force of the same magnitude:

$$\vec{\mathbf{F}} = 0.64\,\mathrm{N}\ \text{westward}$$

In fact, any time the victim particle has negative charge, the effect of the minus sign in the value of the charge q in the equation $\vec{\mathbf{F}} = q\vec{\mathbf{E}}$ is to make the force vector have the direction opposite that of the electric field vector. So the force exerted by an electric field on a negatively charged particle that is at any location in that field, is always in the exact opposite direction to the direction of the electric field itself at that location.

Let's investigate this direction business for cases in which the direction is specified in terms of unit vectors. Suppose that a Cartesian reference frame[2] has been established in an empty region of space in which there is an electric field. Further assume that the electric field at a particular point, call it point P, is:

$$\vec{\mathbf{E}} = 5.0\frac{\text{kN}}{\text{C}}\hat{\mathbf{k}}$$

Now suppose that a proton ($q = 1.60\times10^{-19}\,\text{C}$) is placed at point P. What force would the electric field exert on the proton?

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}}$$

$$\vec{\mathbf{F}} = (1.60\times10^{-19}\,\text{C})\,5.0\times10^{3}\frac{\text{N}}{\text{C}}\hat{\mathbf{k}}$$

$$\vec{\mathbf{F}} = 8.0\times10^{-16}\text{N}\,\hat{\mathbf{k}}$$

The force on the proton is in the same direction as that of the electric field at the location at which the proton was placed (the electric field is in the +z direction and so is the force on the proton), as it must be for the case of a positive victim.

If, in the preceding example, instead of a proton, an electron ($q = -1.60\times10^{-19}\,\text{C}$) is placed at point P, recalling that in the example $\vec{\mathbf{E}} = 5.0\frac{\text{kN}}{\text{C}}\hat{\mathbf{k}}$, we have

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}}$$

$$\vec{\mathbf{F}} = (-1.60\times10^{-19}\,\text{C})\,5.0\times10^{3}\frac{\text{N}}{\text{C}}\hat{\mathbf{k}}$$

$$\vec{\mathbf{F}} = -8.0\times10^{-16}\text{N}\,\hat{\mathbf{k}}$$

The negative sign is to be associated with the unit vector. This means that the force has a magnitude of $8.0\times10^{-16}\,\text{N}$ and a direction of $-\hat{\mathbf{k}}$. The latter means that the force is in the –z direction which is the opposite direction to that of the electric field. Again, this is as expected.

[2] A reference frame is a coordinate system.

The force exerted on a negatively charged particle by the electric field is always in the direction opposite that of the electric field itself.

In the context of the electric field as the set of all electric field vectors in a region of space, the simplest kind of an electric field is a *uniform electric field.* A uniform electric field is one in which every electric field vector has one and the same magnitude and one and the same direction. So, we have an infinite set of electric field vectors, one at every point in the region of space where the uniform electric field is said to exist, and every one of them has the same magnitude and direction as every other one. A charged particle victim that is either released from rest within such an electric field, or launched with some initial velocity within such a field, will have one and the same force exerted upon it, no matter where it is in the electric field. By Newton's 2nd Law, this means that the particle will experience a constant acceleration. If the particle is released from rest, or, if the initial velocity of the particle is in the same direction as, or the exact opposite direction to, the electric field, the particle will experience constant acceleration motion in one dimension. If the initial velocity of the particle is in a direction that is not collinear[3] with the electric field, then the particle will experience constant acceleration motion in two dimensions. The reader should review these topics from *Calculus-Based Physics I.*

Electric Field Diagrams

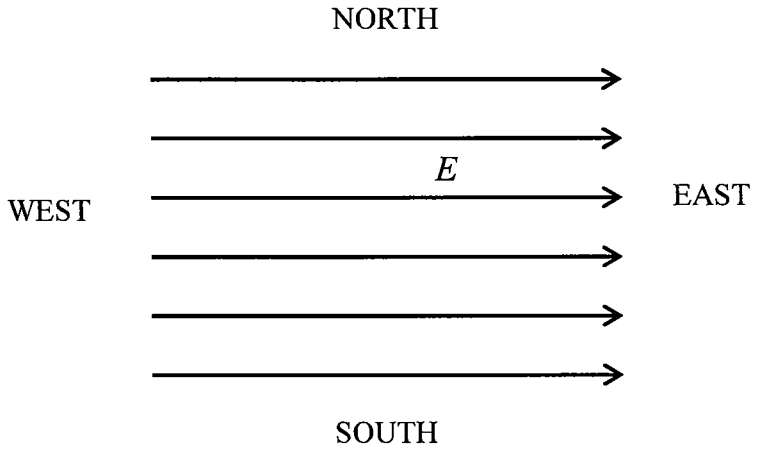
Consider a region in space in which there is a uniform, eastward-directed field. Suppose we want to depict this situation, as viewed from above, in a diagram. At every point in the region of space where the electric field exists, there is an electric field vector. Because the electric field is uniform, all the vectors are of the same magnitude and hence, we would draw all the arrows representing the electric field vectors, the same length. Since the field is uniform and eastward, we would draw all the arrows so that they would be pointing eastward. The problem is that it is not humanly possible to draw an arrow at every point on the region of a page used to depict a region of space in which there is an electric field. Another difficulty is that in using the convention that the length of a vector is representative of its magnitude, the arrows tend to run into each other and overlap.

Physicists have adopted a set of conventions for depicting electric fields. The result of the application of the conventions is known as an electric field diagram. According to the convention, the drawer creates a set of curves or lines, with arrowheads, such that, at every point on each curve, the electric field is, at every point on the curve, directed tangent to the curve, in the direction agreeing with that depicted by the arrowhead on that curve. Furthermore, the spacing of the lines[4] in one region of the diagram as compared to other regions in the diagram is representative of the magnitude of the electric field relative to the magnitude at other locations in

[3] "Collinear" means "along the same line as". Two vectors that are collinear are either in one and the same direction or in exact opposite directions to each other.

[4] In geometry, a line is a straight line. In physics, in the context of fields, lines can be curved or straight. The notion of a curved line also arises in nautical terminology—the waterline of a ship or boat is curved. Electric field lines can be curved or straight.

the same diagram. The closer the lines are, the stronger the electric field they represent. In the case of the uniform electric field in question, because the magnitude of the electric field is the same everywhere (which is what we mean by "uniform"), the line spacing must be the same everywhere. Furthermore, because the electric field in this example has a single direction, namely eastward, the electric field lines will be *straight* lines, with arrowheads:



3 The Electric Field Due to one or more Point Charges

A charged particle (a.k.a. a point charge, a.k.a. a source charge) causes an electric field to exist in the region of space around itself. This is Coulomb's Law for the Electric Field in conceptual form. The region of space around a charged particle is actually the rest of the universe. In practice, the electric field at points in space that are far from the source charge is negligible because the electric field due to a point charge "dies off like one over r-squared." In other words, the electric field due to a point charge obeys an inverse square law, which means, that the electric field due to a point charge is proportional to the reciprocal of the square of the distance that the point in space, at which we wish to know the electric field, is from the point charge that is causing the electric field to exist. In equation form, Coulomb's Law for the magnitude of the electric field due to a point charge reads

$$E = \frac{k|q|}{r^2} \tag{3-1}$$

where

E is the magnitude of the electric field at a point in space,

k is the universal Coulomb constant $k = 8.99 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}}$,

q is the charge of the particle that we have been calling the point charge, and

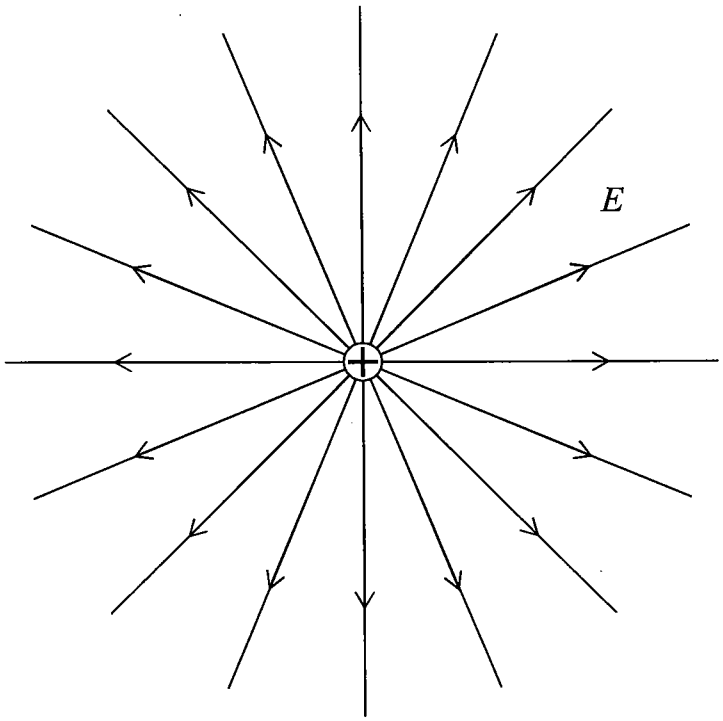
r is the distance that the point in space, at which we want to know E, is from the point charge that is causing E.

Again, Coulomb's Law is referred to as an inverse square law because of the way the magnitude of the electric field depends on the distance that the point of interest[1] is from the source charge.

Now let's talk about direction. Remember, the electric field at any point in space is a force-per-charge-of-would-be-victim *vector* and as a vector, it always has direction. We have already discussed the defining statement for the direction of the electric field: The electric field at a point in space is in the direction of the force that the electric field would exert on a positive victim if there were a positive victim at that point in space. This defining statement for the direction of the electric field is about the *effect* of the electric field. We need to relate this to the *cause* of the electric field. Let's use some grade-school knowledge and common sense to find the direction of the electric field due to a *positive source charge*. First, we just have to obtain an imaginary positive test charge. I recommend that you keep one in your pocket at all times (when not in use) for just this kind of situation. Place your positive test charge in the vicinity of the source charge, at the location at which you wish to know the direction of the electric field. We know that like charges repel, so, the positive source charge repels our test charge. This means that the source charge, the point charge that is causing the electric field under investigation to exist, exerts a force on the test charge that is directly away from the source charge. Again, the electric field at any point is in the direction of the force that would be exerted on a positive test charge if that charge was at that point, so, the direction of the electric field is "directly away from the positive source charge." You get the same result no matter where, in the region of space

[1] The point of interest is the point at which we wish to calculate the electric field due to the point charge.

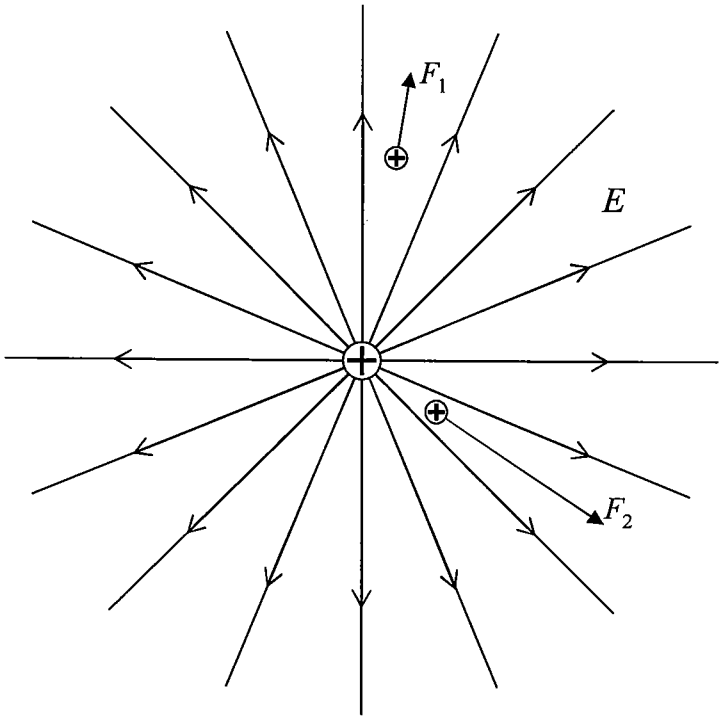
around the source charge, you put the positive test charge. So, put your imaginary positive test charge back in your pocket. It has done its job. We know what we needed to know. The electric field due to a positive source charge, at any point in the region of space around that positive source charge, is directed directly away from the positive source charge. At every point in space, around the positive source charge, we have an electric field vector (a force-per-charge-of-would-be-victim vector) pointing directly away from the positive source charge. So, how do we draw the electric field diagram for that? We are supposed to draw a set of lines or curves *with arrowheads* (NEVER OMIT THE ARROWHEADS!), such that, at every point on each line or curve, the electric field vector at that point is directed along the line or curve in the direction specified by the arrowhead or arrowheads on that line or curve. Let's give it a try.



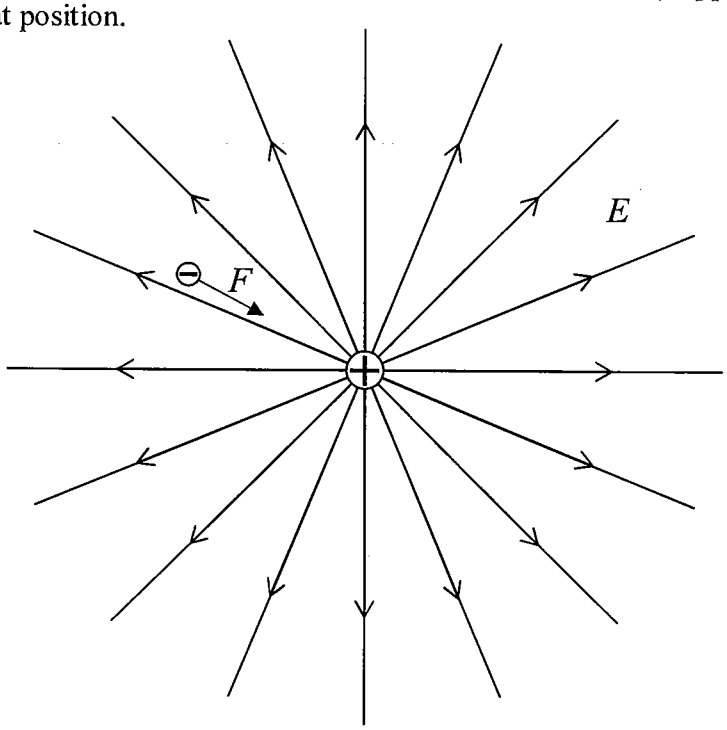
The number of lines drawn extending out of the positive source charge is chosen arbitrarily, but, if there was another positively charged particle, with twice the charge of the first one, in the same diagram, I would need to have twice as many lines extending out of it. That is to say that the line spacing has no absolute meaning overall, but it does have some relative meaning within a single electric field diagram. Recall the convention that the closer together the electric field lines are, the stronger the electric field. Note that in the case of a field diagram for a single source charge, the lines turn out to be closer together near the charged particle than they are farther away. It turned out this way when we created the diagram to be consistent with the fact that the electric field is always directed directly away from the source charge. The bunching of the lines close to the source charge (signifying that the electric field is strong there) is consistent with the inverse square dependence of the electric field magnitude on the distance of the point of interest from the source charge.

There are a few of important points to be made here. The first one is probably pretty obvious to you, but, just to make sure: The electric field exists between the electric field lines—its

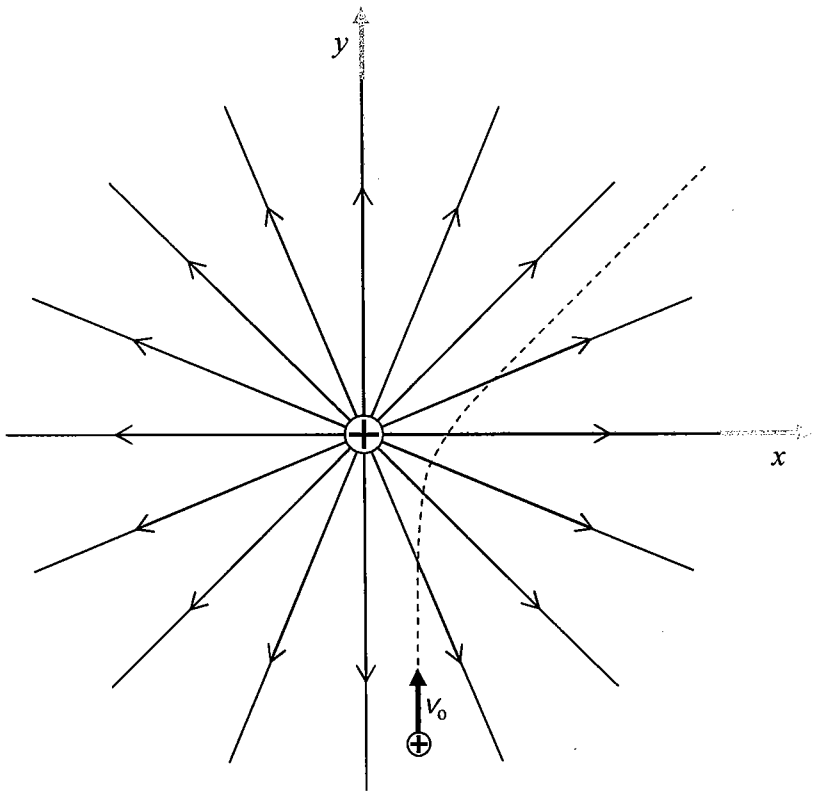
existence there is implied by the lines that are drawn—we simply can't draw lines everywhere that the electric field does exist without completely blackening every square inch of the diagram. Thus, a charged victim that finds itself at a position in between the lines will experience a force as depicted below for each of two different positively-charged victims.



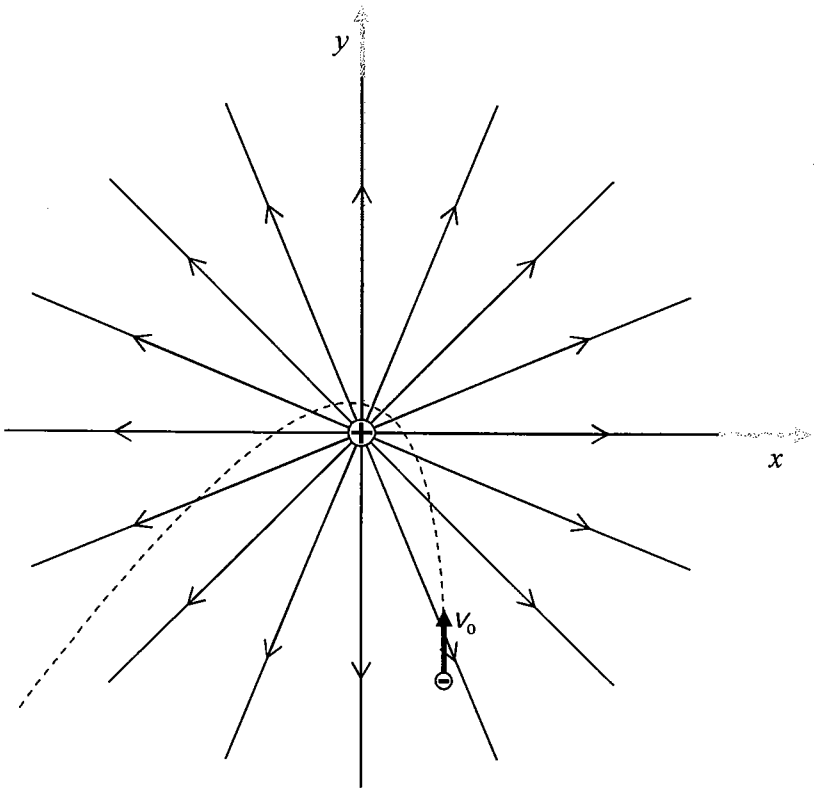
The next point is a reminder that a negatively-charged particle that finds itself at a position at which an electric field exists, experiences a force in the direction exactly opposite that of the electric field at that position.



The third and final point that should be made here is a reminder that the direction of the force experienced by a particle, is not, in general, the direction in which the particle moves. To be sure, the expression "in general" implies that there *are* special circumstances in which the particle would move in the same direction as that of the electric field but these are indeed special. For a particle on which the force of the electric field is the only force acting, there is no way it will stay on one and the same electric field line (drawn or implied) unless that electric field line is straight (as in the case of the electric field due to a single particle). Even in the case of straight field lines, the only way a particle will stay on one and the same electric field line is if the particle's initial velocity is zero, or if the particle's initial velocity is in the exact same direction as that of the straight electric field line. The following diagram depicts a positively-charged particle, with an initial velocity directed in the +y direction. The dashed line depicts the trajectory for the particle (for one set of initial velocity, charge, and mass values). The source charge at the origin is fixed in position by forces not specified.

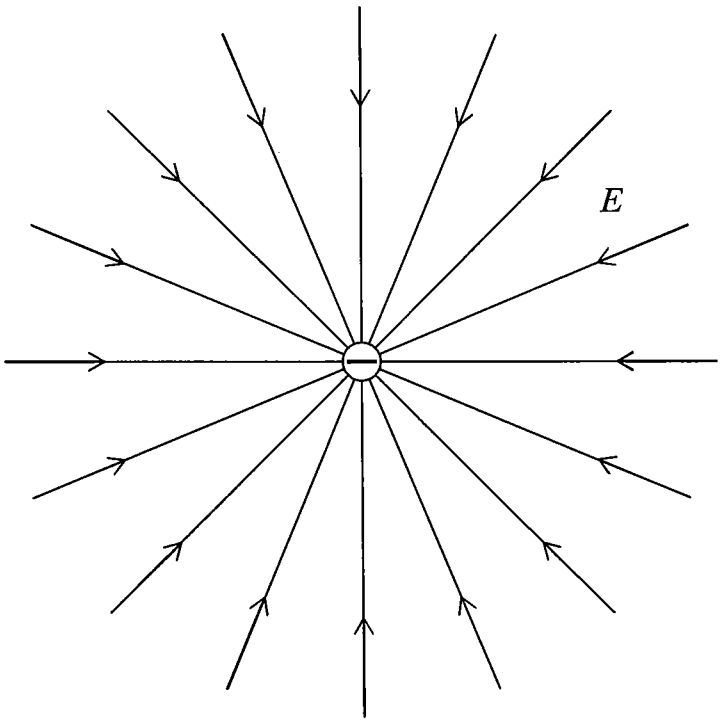


Here is an example of a trajectory of a negatively-charged particle, again for one set of values of source charge, victim charge, victim mass, and victim initial velocity:



Again, the point here is that, in general, charged particles do not move along the electric field lines, rather, they experience a force along (or, in the case of negative particles, in the exact opposite direction to) the electric field lines.

At this point, you should know enough about electric field diagrams to construct the electric field diagram due to a single *negatively*-charged particle. Please do so and then compare your work with the following diagram:



Some General Statements that can be made about Electric Field Lines

The following useful facts about electric field lines can be deduced from the definitions you have already been provided:

1) Every electric field line begins either at infinity or at a positive source charge.

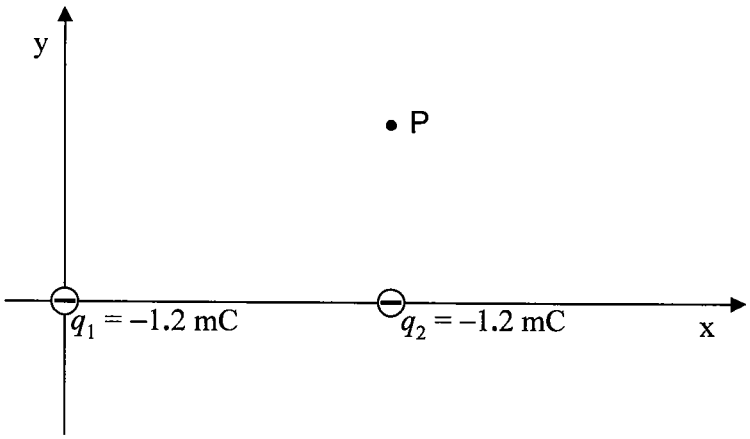
2) Every electric field line ends either at infinity or at a negative source charge.

3) Electric field lines never cross each other or themselves.

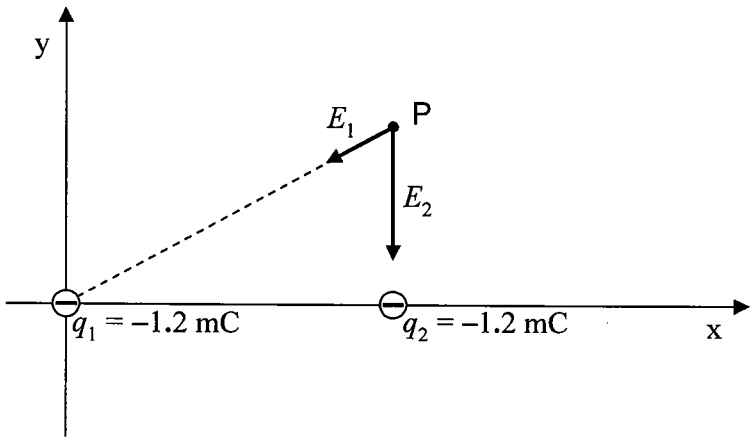
Superposition

If there is more than one source charge, each source charge contributes to the electric field at every point in the vicinity of the source charges. The electric field at a point in space in the vicinity of the source charges is the vector sum of the electric field at that point due to each source charge. For instance, suppose the set of source charges consists of two charged particles. The electric field at some point P will be the electric field vector at point P due to the first charged particle plus the electric field vector at point P due to the second particle. The determination of the total electric field at point P is a vector addition problem because the two electric field vectors contributing to it are, as the name implies, vectors.

Suppose, for instance, that you were asked to find the magnitude and direction of the electric field vector[2] at point P due to the two charges depicted in the diagram below:



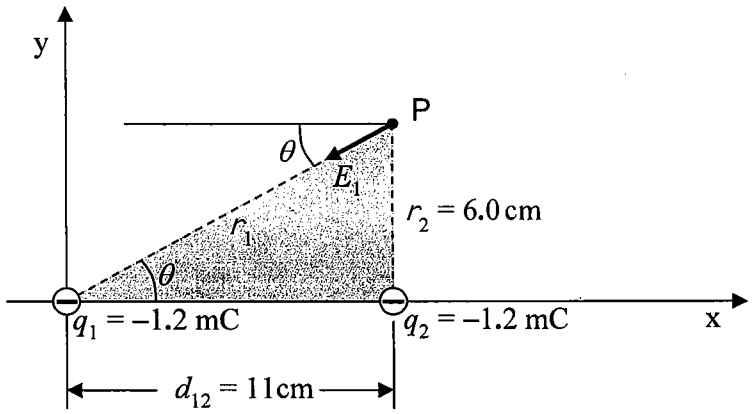
given that charge q_1 is at (0,0), q_2 is at (11 cm, 0) and point P is at (11 cm, 6.0 cm). The first thing that you would have to do is to find the direction and magnitude of $\vec{\mathbf{E}}_1$ (the electric field vector due to q_1) and the direction and magnitude of $\vec{\mathbf{E}}_2$ (the electric field vector due to q_2).



Referring to the diagram above, the direction of $\vec{\mathbf{E}}_2$ is "the –y direction" by inspection.

[2] We use the expression "the electric field vector at point P" for added clarity in distinguishing between the electric field as the infinite set of all electric field vectors and the electric field as the electric field vector at a particular point in space. The reader is warned that it is common practice to use the expression "the electric field at point P" and the reader is expected to tell from the context, that it means "the electric field *vector* at point P".

The angle θ specifying the direction of $\vec{\mathbf{E}}_1$ can be determined by analyzing the shaded triangle in the following diagram.



Analysis of the shaded triangle will also give the distance r_1 that point P is from charge q_1. The value of r_1 can then be substituted into

$$E_1 = \frac{k|q_1|}{r_1^2}$$

to get the magnitude of $\vec{\mathbf{E}}_1$. Based on the given coordinates, the value of r_2 is apparent by inspection and we can use it in

$$E_2 = \frac{k|q_2|}{r_2^2}$$

to get the magnitude of $\vec{\mathbf{E}}_2$. With the magnitude and direction for both $\vec{\mathbf{E}}_1$ and $\vec{\mathbf{E}}_2$, you follow the vector addition recipe to arrive at your answer:

The Vector Addition Recipe

1. For each vector:
   a. Draw a vector component diagram.
   b. Analyze the vector component diagram to get the components of the vector.
2. Add the x components to get the x component of the resultant.
3. Add the y components to get the y component of the resultant.
4. For the resultant:
   a. Draw a vector component diagram.
   b. Analyze the vector component diagram to get the magnitude and direction of the resultant.

Coulomb's Law for the Electric Field in Vector Equation Form

The magnitude and direction information on Coulomb's Law for the Electric Field can be combined in one equation. Namely,

$$\vec{\mathbf{E}} = \frac{kq}{r^2}\hat{\mathbf{r}} \tag{3-2}$$

where:

$\vec{\mathbf{E}}$ is the electric field at an empty point in space[3], call it point P, due to a point charge,

k is the Coulomb constant $8.99\times10^9 \frac{\mathrm{N\cdot m^2}}{\mathrm{C^2}}$,

q is the charge of the charged particle (the point charge) that is causing the electric field to exist,

r is the distance that point P is from the point charge that is causing the electric field, and

$\hat{\mathbf{r}}$ is a unit vector in the "from the point charge toward point P" direction.

Note the absence of the absolute value signs about the q in the expression $\vec{\mathbf{E}} = \frac{kq}{r^2}\hat{\mathbf{r}}$. (We did have them in the case of the expression $E = \frac{k|q|}{r^2}$ for the magnitude of the electric field.) In the vector equation $\vec{\mathbf{E}} = \frac{kq}{r^2}\hat{\mathbf{r}}$, the sign indicating what kind of charge the source charge is, treated algebraically, automatically yields the correct direction for the electric field. For example, if the charge is negative, after substituting the negative value of charge in $\vec{\mathbf{E}} = \frac{kq}{r^2}\hat{\mathbf{r}}$, the minus sign is associated with the unit vector, and, the direction $-\hat{\mathbf{r}}$ of the resulting electric field vector at point P is the direction "from point P, toward the source charge." This is consistent with our understanding that a positive test charge, placed at point P, would experience a force directly toward the negative source charge (since opposites attract), and, the direction of the force on a positive test charge at a specific location is the direction of the electric field vector at that location.

Self-Consistency

In chapter 1 we said that the force that one charged particle, call it particle 1, exerts on another charged particle, particle 2, is given by equation 1-2:

$$\vec{\mathbf{F}}_{12} = k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{12}$$

[3] The point in space doesn't really *have* to be empty. We use the expression "empty point in space" to emphasize the fact that we don't need a charged particle at the location at which we are calculating the electric field. The point is, it *can* be empty.

In chapter 2, we said that the force exerted on a charged particle by an electric field is given by equation 2-1:

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}}$$

In this chapter we said that the electric field at point P is given by equation 3-2:

$$\vec{\mathbf{E}} = \frac{kq}{r^2}\hat{\mathbf{r}}$$

If we call the source charge q_1, (rather than q) we can write this latter equation as

$$\vec{\mathbf{E}} = \frac{kq_1}{r^2}\hat{\mathbf{r}}$$

where the subscripts on the unit vector make it clear that it is in the direction "from particle 1 toward point P."

Substituting this expression into our force of the electric field equation written for the case of a victim q_2 at point P ($\vec{\mathbf{F}} = q_2\vec{\mathbf{E}}$) yields equation 1-2:

$$\vec{\mathbf{F}}_{12} = k\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}_{12}$$

which is the expression for the Coulomb force exerted on charged particle 2 by charged particle 1 introduced back in Chapter 1—the expression without the "middleman" (the electric field).

4 Conductors and the Electric Field

An ideal conductor is chock full of charged particles that are perfectly free to move around within the conductor. Like all macroscopic samples of material, an ideal conductor consists of a huge amount of positive charge, and, when neutral, the same amount of negative charge. When not neutral, there is a tiny fractional imbalance one way or the other. In an ideal conductor, some appreciable fraction of the charge is completely free to move around within the conducting material. The ideal (perfect) conductor is well-approximated by some materials familiar to you, in particular, metals. In some materials, it is positive charge that is free to move about, in some, it is negative, and in others, it is both. For our purposes, the observable effects of positive charge moving in one direction are so close to being indistinguishable from negative charge moving in the opposite direction that, we will typically treat the charge carriers as being positive without concern for what the actual charge carriers are[1].

Here, we make one point about conductors by means of an analogy. The analogy involves a lake full of fish. Let the lake represent the conductor and the fish the charge carriers. The fish are free to move around anywhere within the lake, but, and this is the point, they can't, under ordinary circumstances, escape the lake. They can go to every boundary of the body of water, you might even see some on the surface, but, they cannot leave the water. This is similar to the charge carriers in a conductor surrounded by vacuum or an insulating medium such as air. The charges can go everywhere in and on the conductor, but, they cannot leave the conductor.

The facts we have presented on the nature of charge, electric fields, and conductors allow one to draw some definite conclusions about the electric field and unbalanced charge within the material of, and at or on the surface of, an ideal conductor. Please try to reason out the answers to the following questions:

1) Suppose you put a neutral ideal conducting solid sphere in a region of space in which there is, initially, a uniform electric field. Describe (as specifically as possible) the electric field inside the conductor and the electric field at the surface of the conductor. Describe the distribution of charge in and on the conductor.

2) Repeat question 1 for the case of a non-uniform field.

3) Suppose you put some charge on an initially-neutral, solid, perfectly-conducting sphere (where the sphere is not in a pre-existing electric field). Describe the electric field inside the conductor, at the surface of the conductor, and outside the conductor as a result of the unbalanced charge. Describe the distribution of the charge in and on the conductor.

4) Repeat questions 1-3 for the case of a hollow perfectly-conducting spherical shell (with the interior being vacuum).

---

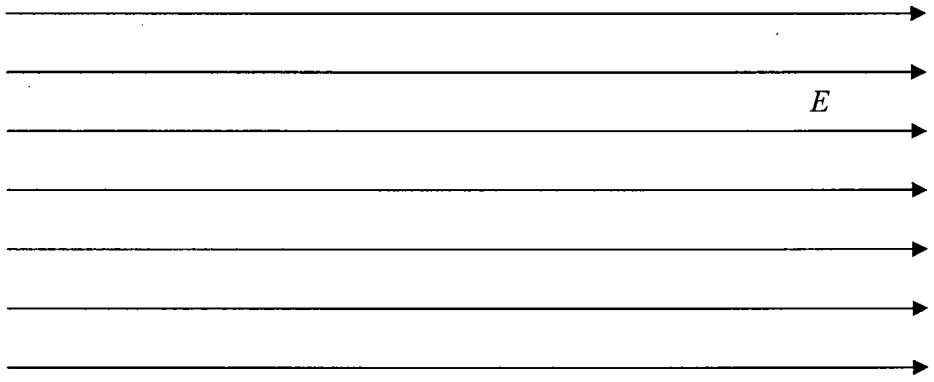
[1] We refer to this as the positive charge carrier model for charge movement. We thoroughly exploit it in our analysis of circuits (in later chapters). Such analysis leads to accurate results even though it is typically applied to circuits in which the nearly ideal conductors are metals, materials in which the charge carriers are electrons, which, as you know, are negatively charged.

5) How would your answers to questions 1-4 change if the conductor had some shape other than spherical?

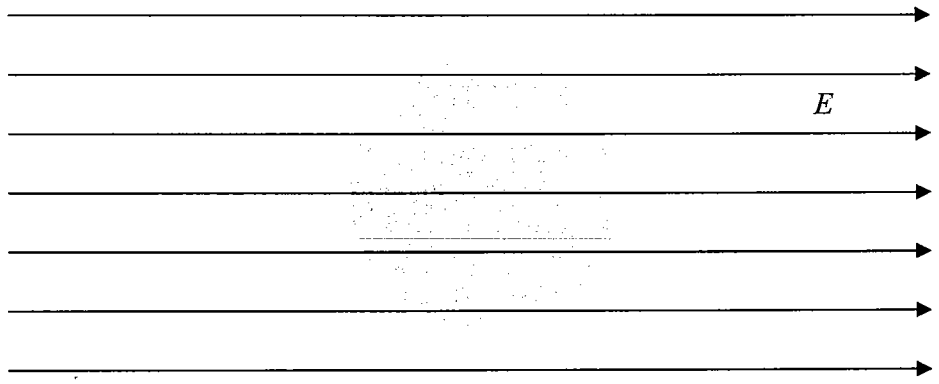
Here we provide the answers (preceded in each case, with the corresponding question).

1) Suppose you put a neutral ideal conducting solid sphere in a region of space in which there is, initially, a uniform electric field. Describe (as specifically as possible) the electric field inside the conductor and the electric field at the surface of the conductor. Describe the distribution of charge in and on the conductor.

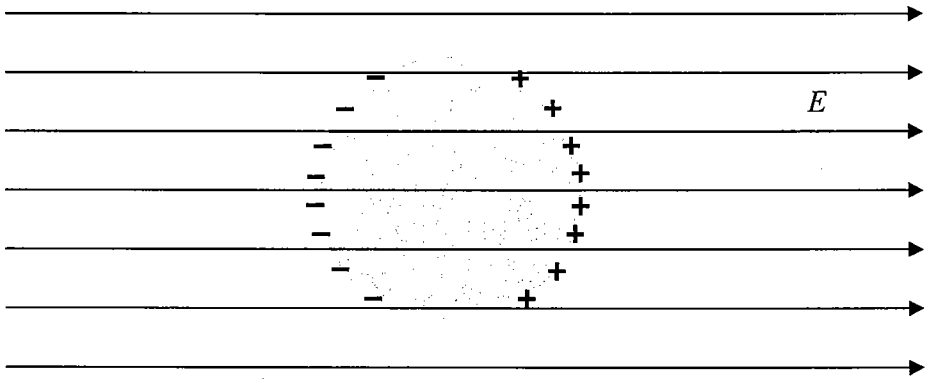
Answer: We start with a uniform electric field.



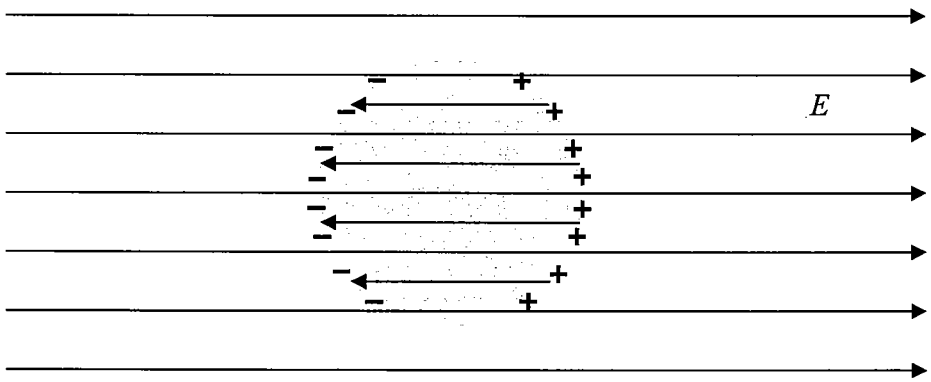
We put a solid, ideal conductor in it. The electric field permeates everything, including the conductor.



The charged particles in the conductor respond to the force exerted on them by the electric field. (The force causes acceleration, the acceleration of particles that are initially at rest causes them to acquire some velocity. In short, they move.) All this occurs in less than a microsecond. The net effect is a redistribution of the charged particles.

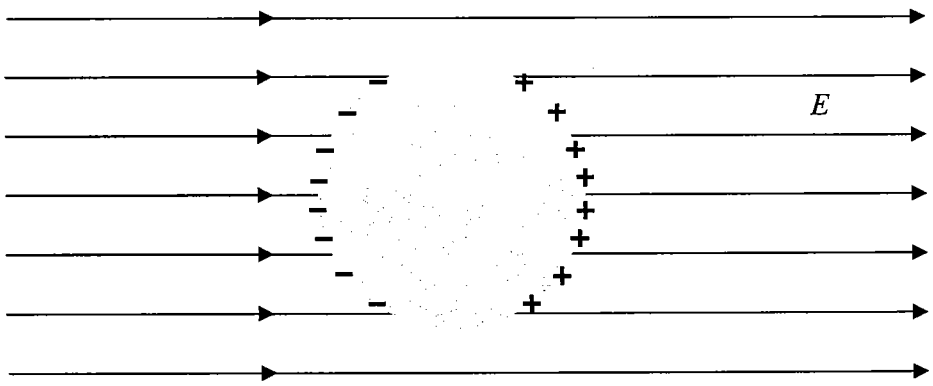


Now, get this! The charged particles create their own electric field.

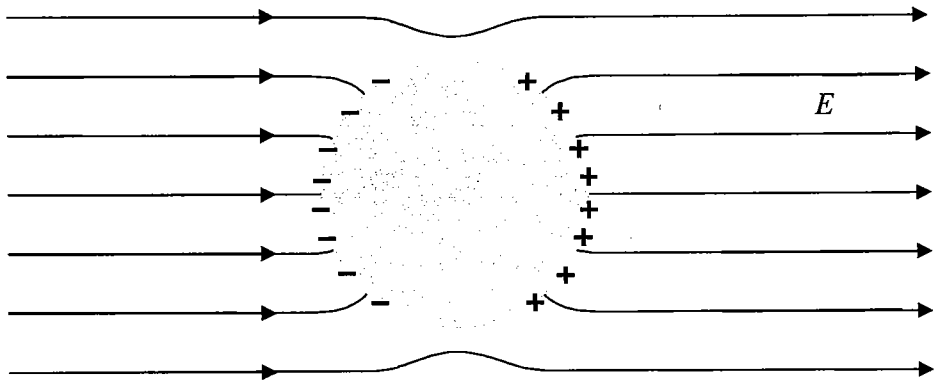


The total electric field at any point in the conductor is the vector sum of the original electric field and the electric field due to the redistributed charged particles. Since they are oppositely-directed, the two contributions to the electric field inside the conductor tend to cancel each other. Now comes the profound part of the argument: the two contributions to the electric field at any point in the conductor exactly cancel. We know they have to completely cancel because, if they didn't, the free-to-move-charge in the conductor would move as a result of the force exerted on it by the electric field. And the force on the charge is always in a direction that causes the charge to be redistributed to positions in which it will create its own electric field that tends to cancel the electric field that caused the charge to move. The point is that the charge will not stop responding to the electric field until the net electric field at every point in the conductor is zero.

So far, in answer to the question, we have: The electric field is zero at all points inside the conductor, and, while the total charge is still zero, the charge has been redistributed as in the following diagram:



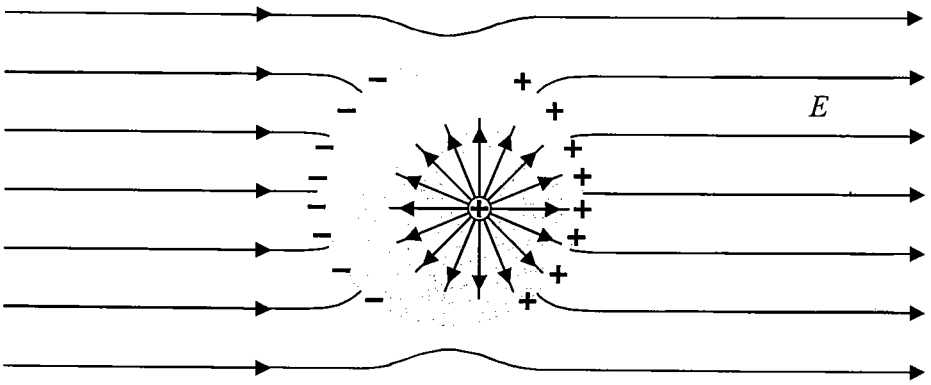
Recall that we were also called upon to describe the electric field at the surface of the conductor. Note that the charge on the surface of the sphere will not only contribute to the electric field inside the conductor, it will also contribute to the electric field outside. The net effect of all the contributions to the electric field in the near vicinity of the sphere is to cause the electric field to be normal to (perpendicular to) the surface of the sphere at all points where it meets the sphere.



How is it that we are able to assert this without doing any calculations? Here's the argument: If the electric field at the surface had a component parallel to the surface, then the charged particles on the surface of the conductor would experience a force directed along the surface. Since those particles are free to move anywhere in the conductor, they would be redistributed. In their new positions, they would make their own contribution to the electric field in the surface and their contribution would cancel the electric field that caused the charge redistribution.

About the charge distribution: The object started out neutral and no charge has left or entered the conductor from the outside world so it is still neutral. But we do see a separation of the two different kinds of charge. Something that we have depicted but not discussed is the assertion that all the charge resides on the surface. (In the picture above, there is positive charge on the right surface of the sphere and an equal amount of negative charge on the left side.) How do we know

that all charge must be on the surface? Assume that there was a positive point charge at some location within the conductor:

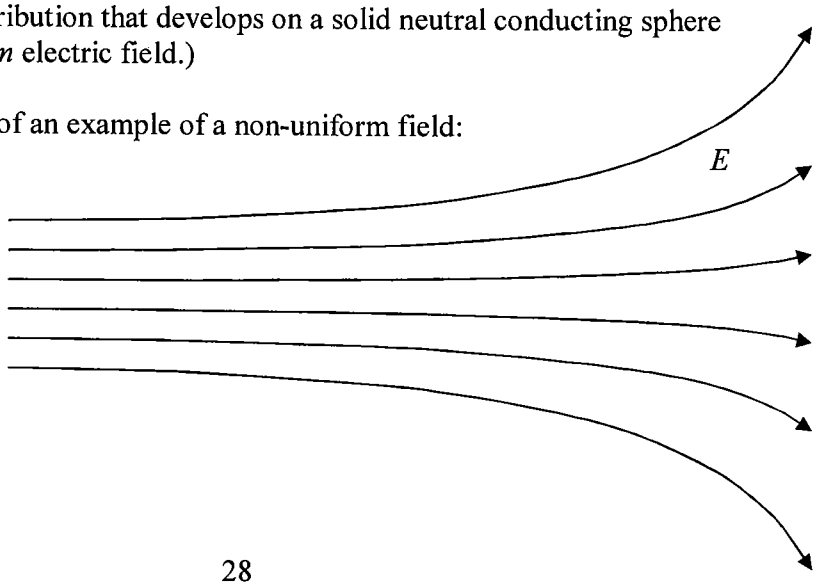


The electric field of that point charge would cause the free-to-move charge in the conductor to move, and it would keep moving as long as there was an electric field. So where would the charge move in order to cancel out the electric field of the positive point charge. You can try any arrangement of charge that you want to, around that positive point charge, but, if it is stipulated that there be a net positive charge at that location, there is no way to cancel out the electric field of that positive charge. So the situation doesn't even occur. If it did happen, the particle would repel the conductor's free-to-move-positive charge away from the stipulated positive charge, so that (excluding the stipulated positive charge under consideration) the conductor would have a net negative charge at that location, an amount of negative charge exactly equal to the originally-stipulated positive charge. Taking the positive charge into account as well, the point, after the redistribution of charge, would be neutral. The point of our argument is that, under static conditions, there can be no net charge inside the material of a perfect conductor. Even if you assume there to be some, it would soon be neutralized by the nearly instantaneous charge redistribution that it would cause.

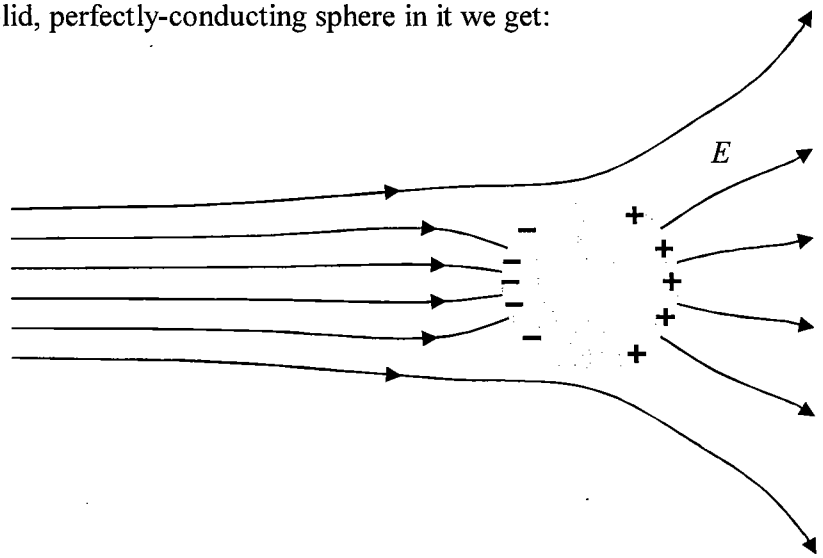
Next question:

2) Repeat question 1 for the case of a *non-uniform field.* (Question 1 asked for a description of the charge distribution that develops on a solid neutral conducting sphere when you place it in a *uniform* electric field.)

Answer: Here is a depiction of an example of a non-uniform field:



If we put a solid, perfectly-conducting sphere in it we get:



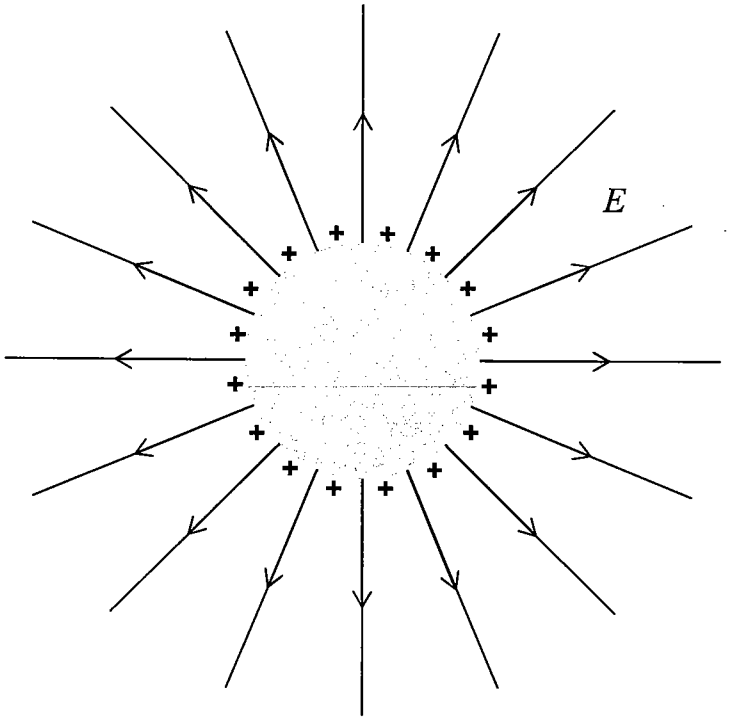
The same arguments lead to the same conclusions. When, after less than a microsecond, the new static conditions are achieved: There can be no electric field inside the conductor or else the free-to-move-charges in it would still be moving around within the volume of the conductor. There can be no unbalanced charge within the volume of the conductor or else there would be an electric field inside the conductor. Hence, any locally unbalanced charge (overall, the initially-neutral sphere remains neutral) must be on the surface. The electric field has to be normal to the surface of the sphere or else the free-to-move-charge at the surface would still be moving around on the surface. The only thing that is different in this case, as compared to the initially-uniform electric field case, is the way the charge is distributed on the surface. We see that the negative charge is more bunched up than the positive charge in the case at hand. In the initially-uniform electric field case, the positive charge distribution was the mirror image of the negative charge distribution.

Next Question:

3) Suppose you put some charge on an initially-neutral, solid, perfectly-conducting sphere (where the sphere is not in a pre-existing electric field). Describe the electric field inside the conductor, at the surface of the conductor, and outside the conductor as a result of the unbalanced charge. Describe the distribution of the charge in and on the conductor.

Again, we assume that we have waited long enough (less than a microsecond) for static conditions to have been achieved. There can be no charge within the bulk of the conductor or else there would be an electric field in the conductor and there can't be an electric field in the conductor or else the conductor's free-to-move charge would move and static conditions would

not be prevailing. So, all the unbalanced charge must be on the surface. It can't be bunched up more at any location on the surface than it is at any other location on the surface or else the charge on the edge of the bunch would be repelled by the bunch and it would move, again in violation of our stipulation that we have waited until charge stopped moving. So, the charge must be distributed uniformly over the surface of the sphere. Inside the sphere there is no electric field. Where the outside electric field meets the surface of the sphere, the electric field must be normal to the surface of the sphere. Otherwise, the electric field at the surface would have a vector component parallel to the surface which would cause charge to move along the surface, again in violation of our static conditions stipulations. Now, electric field lines that are perpendicular to the surface of a sphere lie on lines that pass through the center of the sphere. Hence, outside the sphere, the electric field lines form the same pattern as the pattern that would be formed by a point charge at the location of the center of the sphere (with the sphere gone). Furthermore, if you go so far away from the sphere that the sphere "looks like" a point, the electric field will be the same as that due to a point charge at the location of the center of the sphere. Given that outside the sphere, it has the same pattern as the field due to a point charge at the center of the sphere, the only way it can match up with the point charge field at a great distance from the sphere, is if it is identical to the point charge field everywhere that it exists. So, outside the sphere, the electric field is indistinguishable from the electric field due to the same amount of charge that you put on the sphere, all concentrated at the location of the center of the sphere (with the sphere gone).



Next question:

4) Repeat questions 1-3 for the case of a hollow perfectly-conducting spherical shell (with the interior being vacuum).

In all three cases we have considered so far, the interior of the sphere has played no role. It is initially neutral and it is neutral after the sphere is placed in a pre-existing electric field or some charge is placed on it. Nothing would change if we removed all that neutral material making up the bulk of the conductor, leaving nothing but a hollow shell of a sphere. Hence all the results that we found for the solid sphere apply to the hollow sphere. In particular, the electric field at all points inside an empty hollow perfectly-conducting spherical shell is, under all conditions, zero.

Last question:

5) How would your answers to questions 1-4 change if the conductor had some shape other than spherical?

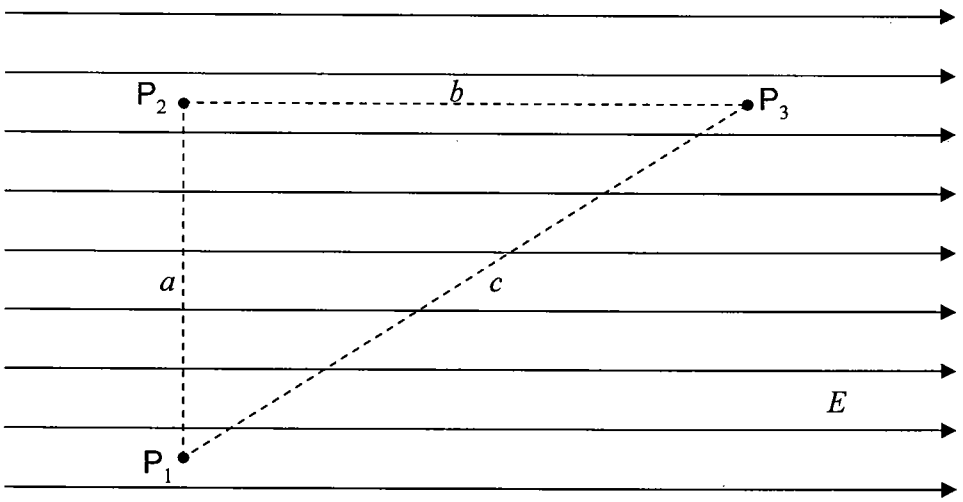
For a solid perfect conductor, the electric field and the charge everywhere inside would have to be zero for the same reasons discussed above. Furthermore, the electric field would have to be normal to the surface for the same reasons as before. Again, it would not make any difference if we hollow out the conductor by removing a bunch of neutral material. The only things that would be different for a non-spherical conductor are the way the charge would be distributed on the surface, and, the outside electric field. In particular, if you put some charge on a perfectly-conducting object that is not a sphere, the electric field in the vicinity of the object will not be the same as the electric field due to a point charge at the center of the object (although the difference would be negligible at great enough distances from the object).

5 Work Done by the Electric Field, and, the Electric Potential

When a charged particle moves from one position in an electric field to another position in that same electric field, the electric field does work on the particle. The work done is conservative; hence, we can define a potential energy for the case of the force exerted by an electric field. This allows us to use the concepts of work, energy, and the conservation of energy, in the analysis of physical processes involving charged particles and electric fields.

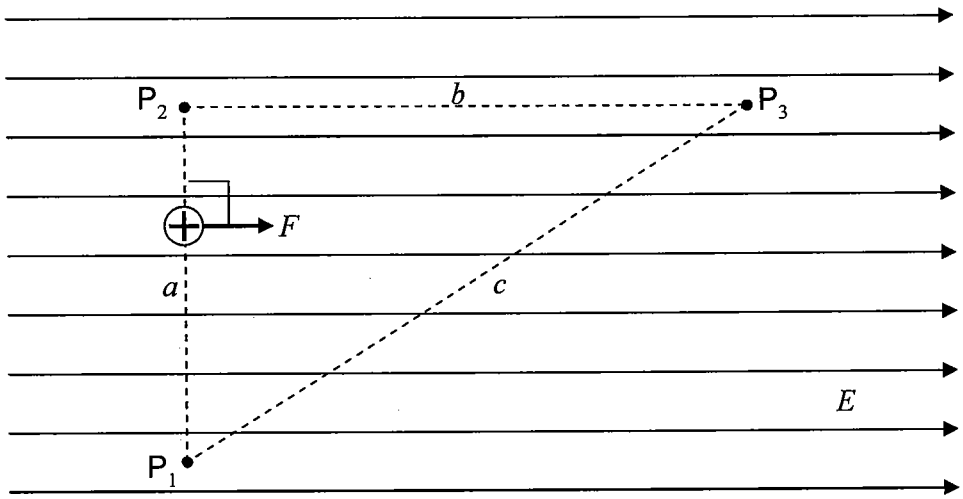
We have defined the work done on a particle by a force, to be the force-along-the-path times the length of the path, with the stipulation that when the component of the force along the path is different on different segments of the path, one has to divide up the path into segments on each of which the force-along-the-path has one value for the whole segment, calculate the work done on each segment, and add up the results.

Let's investigate the work done by the electric field on a charged particle as it moves in the electric field in the rather simple case of a uniform electric field. For instance, let's calculate the work done on a positively-charged particle of charge q as it moves from point P_1 to point P_3



along the path: "From P_1 straight to point P_2 and from there, straight to P_3." Note that we are not told what it is that makes the particle move. We don't care about that in this problem. Perhaps the charged particle is on the end of a quartz rod (quartz is a good insulator) and a person who is holding the rod by the other end moves the rod so the charged particle moves as specified.

Along the first part of the path, from P_1 to P_2, the force on the charged particle is perpendicular to the path.

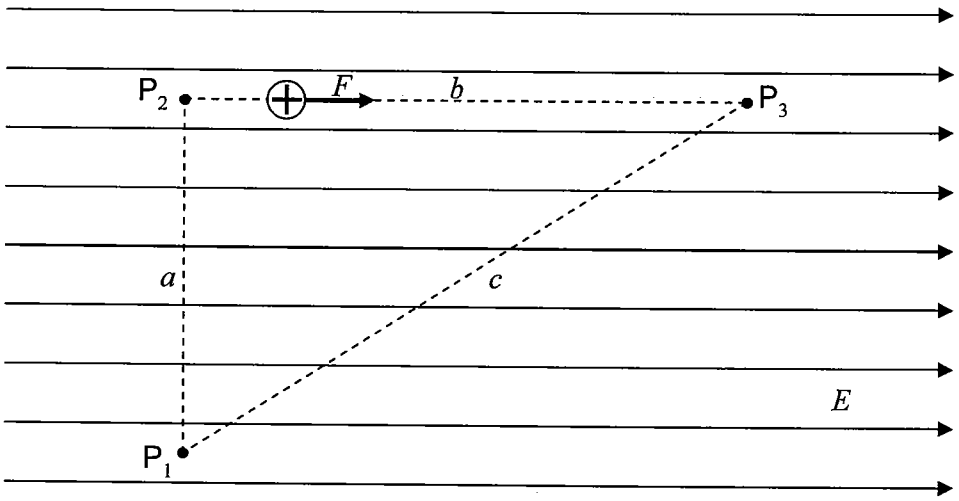


The force has no component along the path so it does no work on the charged particle at all as the charged particle moves from point P_1 to point P_2.

$$W_{12} = 0$$

From P_2, the particle goes straight to P_3.

On that segment of the path (from P_2 to P_3) the force is in exactly the same direction as the direction in which the particle is going.



As such, the work is just the magnitude of the force times the length of the path segment:

$$W_{23} = Fb$$

The magnitude of the force is the charge of the particle times the magnitude of the electric field $F=qE$, so,

$$W_{23} = qEb$$

Thus, the work done on the charged particle by the electric field, as the particle moves from point P_1 to P_3 along the specified path is

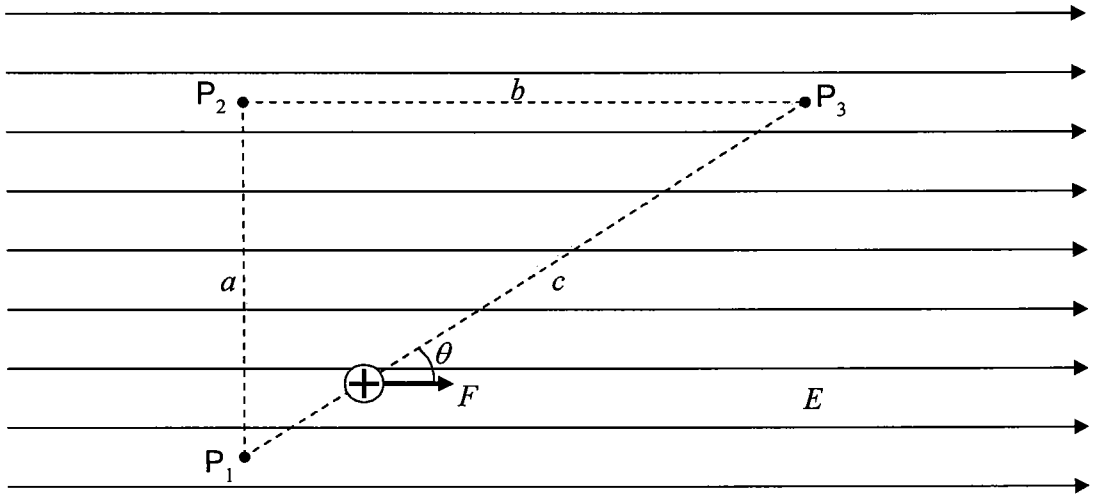
$$W_{123} = W_{12} + W_{23}$$

$$W_{123} = 0 + qEb$$

$$\boxed{W_{123} = qEb}$$

(This is just an answer to a sample problem. Don't use it as a starting point for the solution to a homework or test problem.)

Now let's calculate the work done on the charged particle if it undergoes the same displacement (from P_1 to P_3) but does so by moving along the direct path, straight from P_1 to P_3.



The force on a positively-charged particle being in the same direction as the electric field, the force vector makes an angle θ with the path direction and the expression

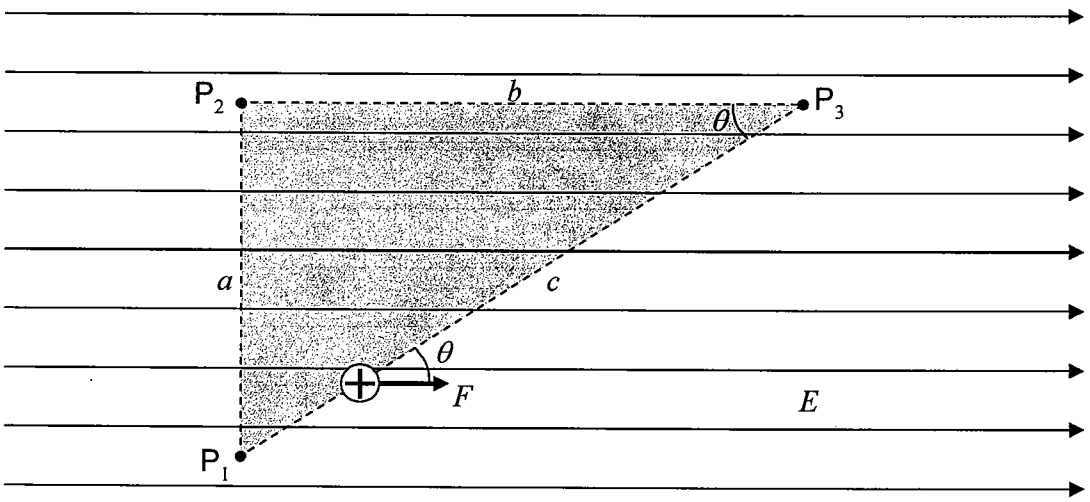
$$W = \vec{\mathbf{F}} \cdot \vec{\Delta \mathbf{r}}$$

for the work becomes

$$W_{13} = Fc\cos\theta$$

$$W_{13} = qEc\cos\theta$$

Analyzing the shaded triangle in the following diagram:



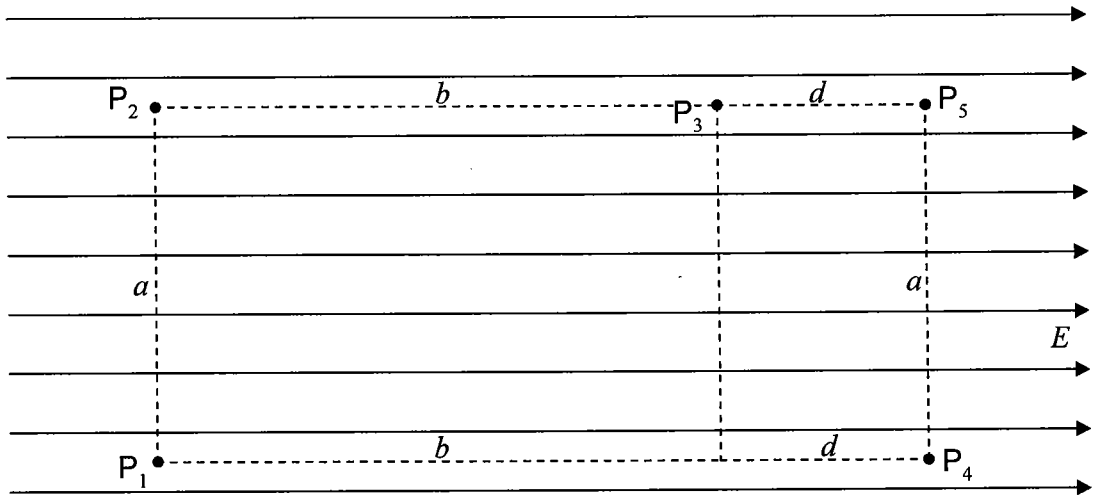
we find that $\cos\theta = \frac{b}{c}$. Substituting this into our expression for the work ($W_{13} = qEc\cos\theta$) yields

$$W_{13} = qEc\frac{b}{c}$$

$$\boxed{W_{13} = qEb}$$

(This is just an answer to a sample problem.)

This is the same result we got for the work done on the charged particle by the electric field as the particle moved between the same two points (from P_1 to P_3) along the other path (P_1 to P_2 to P_3). As it turns out, the work done is the same no matter what path the particle takes on its way from P_1 to P_3. I don't want to take the time to prove that here but I would like to investigate one more path (not so much to get the result, but rather, to review an important point about how to calculate work). Referring to the diagram:



Let's calculate the work done on a particle with charge q, by the electric field, as the particle moves from P_1 to P_3 along the path "from P_1 straight to P_4, from P_4 straight to P_5, and from P_5 straight to P_3." On P_1 to P_4, the force is in the exact same direction as the direction in which the particle moves along the path, so,

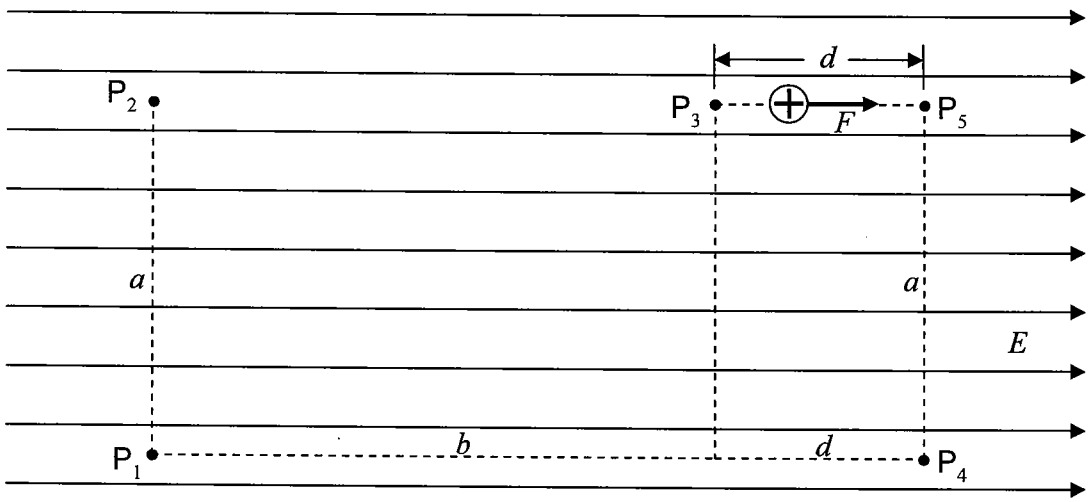
$$W_{14} = F(b+d)$$

$$W_{14} = qE(b+d)$$

From point P_4 to P_5, the force exerted on the charged particle by the electric field is at right angles to the path, so, the force does no work on the charged particle on segment P_4 to P_5.

$$W_{45} = 0$$

On the segment from P_5 to P_3,



the force is in the exact opposite direction to the direction in which the particle moves. This means that the work done by the force of the electric field on the charged particle as the particle moves form P_5 to P_3 is the *negative* of the magnitude of the force times the length of the path segment. Thus

$$\mathcal{W}_{53} = -Fd$$

$$\mathcal{W}_{53} = -qEd$$

and

$$\mathcal{W}_{1453} = \mathcal{W}_{14} + \mathcal{W}_{45} + \mathcal{W}_{53}$$

$$\mathcal{W}_{1453} = qE(b+d) + 0 + (-qEd)$$

$$\boxed{\mathcal{W}_{1453} = qEb}$$

(This is just an answer to a sample problem.)

As advertised, we obtain the same result for the work done on the particle as it moves from P_1 to P_3 along "P_1 to P_4 to P_5 to P_3" as we did on the other two paths.

Whenever the work done on a particle by a force acting on that particle, when that particle moves from point P_1 to point P_3, is the same no matter what path the particle takes on the way from P_1 to P_3, we can define a potential energy function for the force. The potential energy function is an assignment of a value of potential energy to every point in space. Such an assignment allows us to calculate the work done on the particle by the force when the particle moves from point P_1 to point P_3 simply by subtracting the value of the potential energy of the particle at P_1 from the value of the potential energy of the particle at P_3 and taking the negative of the result. In other words, the work done on the particle by the force of the electric field when the particle goes from one point to another is just the negative of the change in the potential energy of the particle.

In determining the potential energy function for the case of a particle of charge q in a uniform electric field $\vec{\mathbf{E}}$, (an infinite set of vectors, each pointing in one and the same direction and each having one and the same magnitude E) we rely heavily on your understanding of the near-earth's-surface gravitational potential energy. Near the surface of the earth, we said back in volume 1 of this book, there is a uniform gravitational field, (a force-per-mass vector field) in the downward direction. A particle of mass m in that field has a force "mg downward" exerted upon it at any location in the vicinity of the surface of the earth. For that case, the potential energy of a particle of mass m is given by mgy where mg is the magnitude of the downward force and y is the height that the particle is above an arbitrarily-chosen reference level. For ease of comparison with the case of the electric field, we now describe the reference level for gravitational potential energy as a plane, perpendicular to the gravitational field $\vec{\mathbf{g}}$, the force-per-mass vector field; and; we call the variable y the "upfield" distance (the distance in the direction opposite that of the gravitational field) that the particle is from the reference plane. (So, we're calling the direction in which the gravitational field points, the direction you know to be downward, the "downfield" direction.)

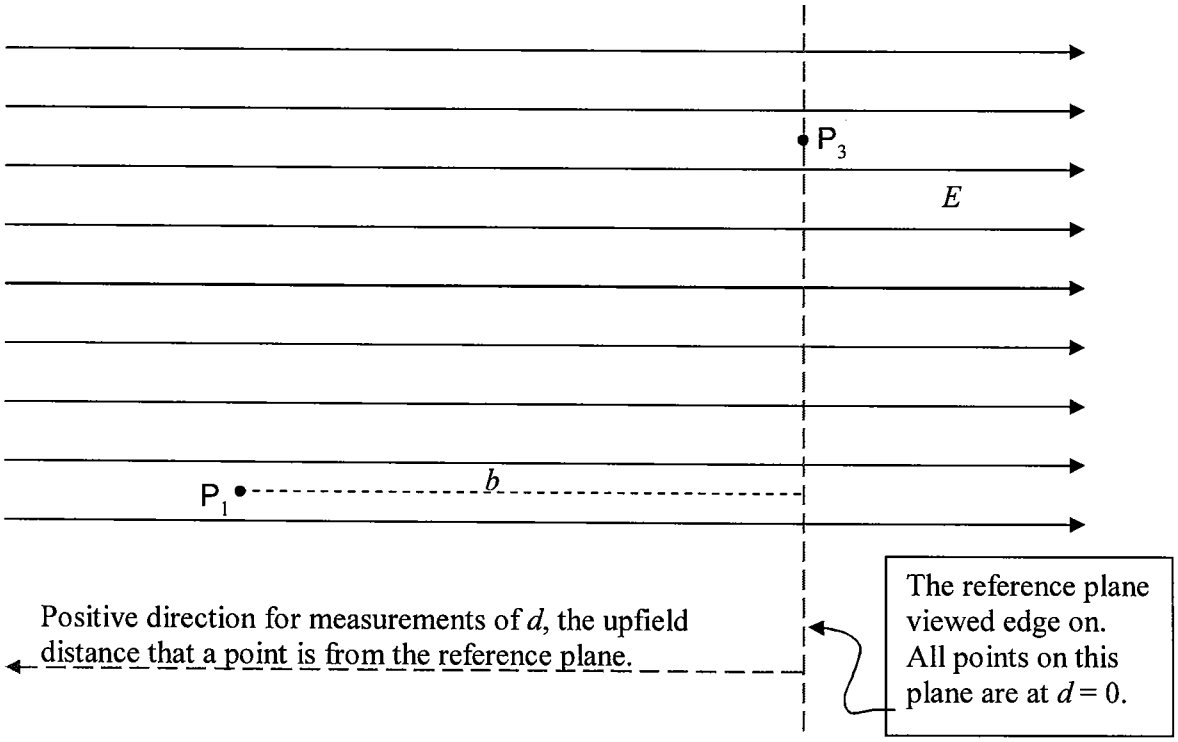
Now let's switch over to the case of the uniform electric field. As in the case of the near-earth's surface gravitational field, the force exerted on its victim by a uniform electric field has one and the same magnitude and direction at any point in space. Of course, in the electric field case, the force is qE rather than mg and the characteristic of the victim that matters is the charge q rather than the mass m. We call the direction in which the electric field points, the "downfield" direction, and the opposite direction, the "upfield" direction. Now we arbitrarily define a plane that is perpendicular to the electric field to be the reference plane for the electric potential energy of a particle of charge q in the electric field. If we call d the distance that the charged particle is away from the plane in the upfield direction, then the potential energy of the particle with charge q is given by

$$U = qEd$$

where:

U is the electric potential energy of the charged particle,
q is the charge of the particle,
E is the magnitude of every electric field vector making up the uniform electric field, and
d is the "upfield" distance that the particle is from the $U = 0$ reference plane.

Let's make sure this expression for the potential energy function gives the result we obtained previously for the work done on a particle with charge q, by the uniform electric field depicted in the following diagram, when the particle moves from P_1 to P_3



As you can see, I have chosen (for my own convenience) to define the reference plane to be at the most downfield position relevant to the problem. With that choice, the particle of charge q, when it is at P_1 has potential energy qEb (since point P_1 is a distance b "upfield" from the reference plane) and, when it is at P_3, the particle of charge q has potential energy 0 since P_3 is on the reference plane.

$$W_{13} = -\Delta U$$

$$W_{13} = -(U_3 - U_1)$$

$$W_{13} = -(0 - qEb)$$

$$\boxed{W_{13} = qEb}$$ *(This is just an answer to a sample problem.)*

This is indeed the result we got (for the work done by the electric field on the particle with charge q as that particle was moved from P_1 to P_3) the other three ways that we calculated this work.

The Electric Potential Energy per Charge

The expression for the work that we found above had the form "the charge of the victim times other stuff." Likewise, the potential energy of the victim (see above) has the form "the charge of the victim times other stuff." In both cases the "other stuff" consisted of quantities characterizing the electric field and positions in space. This turns out to be a general result: The electric potential energy of a charged particle (victim) in any electric field (not just a uniform electric field) can be expressed as the product of the charge of the victim, and, quantities used to characterize the electric field in the region of space in which the particle finds itself. As such, we can always divide the potential energy of the victim by the charge of the victim to obtain what can be called the electric potential energy per charge for the point in space at which the victim finds itself. No matter what the charge of the victim is, the potential energy of the victim divided by the charge of the victim always yields the same value for the potential-energy-per-charge-of-would-be-victim. This is because the potential-energy-per-charge-of-would-be-victim is a characteristic of the point in space at which the victim finds itself, not a characteristic of the victim. This means that we can specify values of potential-energy-per-charge-of-would-be-victim (which we will use the symbol φ to represent) for all the points in a region of space in which there is an electric field, without even having a victim in mind. Then, once you find a victim, the potential energy of the victim at a particular point in space is just

$$U = q\varphi \tag{5-1}$$

where:

- U is the electric potential energy of the victim (the charged particle in the electric field),
- q is the charge of the victim, and,
- φ is the electric-potential-energy-per-charge-of-would-be-victim (also known more simply as the *electric potential*) of the point in space at which the victim finds itself.

Okay, I spilled the beans in the variable list; "potential-energy-per-charge-of-would-be-victim" is just too much of a mouthful so we call it the *electric potential*. Now, for the potential energy U to come out in Joules in the expression $U = q\varphi$, with q having units of coulombs, the electric potential φ must have units of J/C. The concept of electric potential is such an important one that we give its combination unit (J/C) a name. The name of the unit is the volt, abbreviated V.

$$1 \text{ volt} = 1 \frac{\text{joule}}{\text{coulomb}} \quad \text{or} \quad 1\text{V} = 1 \frac{\text{J}}{\text{C}}$$

For the case of a uniform electric field, our expression $U = qEd$ for the electric potential energy of a victim with charge q, upon division by q, yields, for the electric potential at a point of interest in a uniform electric field,

$$\varphi = Ed \tag{5-2}$$

where:

- φ is the electric potential at the point of interest,
- E is the magnitude of every electric field vector in the region of space where the uniform electric field exists, and,
- d is the upfield distance that the point of interest is from the (arbitrarily-chosen) reference plane.

6 The Electric Potential Due to One or More Point Charges

The electric potential due to a point charge is given by

$$\varphi = \frac{kq}{r} \tag{6-1}$$

where

φ is the electric potential due to the point charge,

$k = 8.99\times10^9 \frac{\text{Nm}^2}{\text{C}^2}$ is the Coulomb constant,

q is the charge of the particle (the source charge, a.k.a. the point charge) causing the electric field for which the electric potential applies, and,

r is the distance that the point of interest is from the point charge.

In the case of a non-uniform electric field (such as the electric field due to a point charge), the electric potential method for calculating the work done on a charged particle is much easier than direct application of the force-along-the-path times the length of the path. Suppose, for instance, a particle of charge q' is fixed at the origin and we need to find the work done by the electric field of that particle on a victim of charge q as the victim moves along the x axis from x_1 to x_2. We can't simply calculate the work as

$$F\cdot(x_2 - x_1)$$

even though the force is in the same direction as the displacement, because the force F takes on a different value at every different point on the x axis from $x = x_1$ to $x = x_2$. So, we need to do an integral:

$$dW = F\,dx$$

$$dW = qE\,dx$$

$$dW = q\frac{kq'}{x^2}dx$$

$$\int dW = \int_{x_1}^{x_2} q\frac{kq'}{x^2}dx$$

$$W = kq'q\int_{x_1}^{x_2} x^{-2}dx$$

$$W = kq'q\frac{x^{-1}}{-1}\Bigg|_{x_1}^{x_2}$$

$$W = -kq'q\left(\frac{1}{x_2} - \frac{1}{x_1}\right)$$

$$\boxed{W = -\left(\frac{kq'q}{x_2} - \frac{kq'q}{x_1}\right)}$$

Compare this with the following solution to the same problem (a particle of charge q' is fixed at the origin and we need to find the work done by the electric field of that particle on a victim of charge q as the victim moves along the x axis from x_1 to x_2):

$$W = -\Delta U$$

$$W = -q\,\Delta\varphi$$

$$W = -q\,(\varphi_2 - \varphi_1)$$

$$W = -q\left(\frac{kq'}{x_2} - \frac{kq'}{x_1}\right)$$

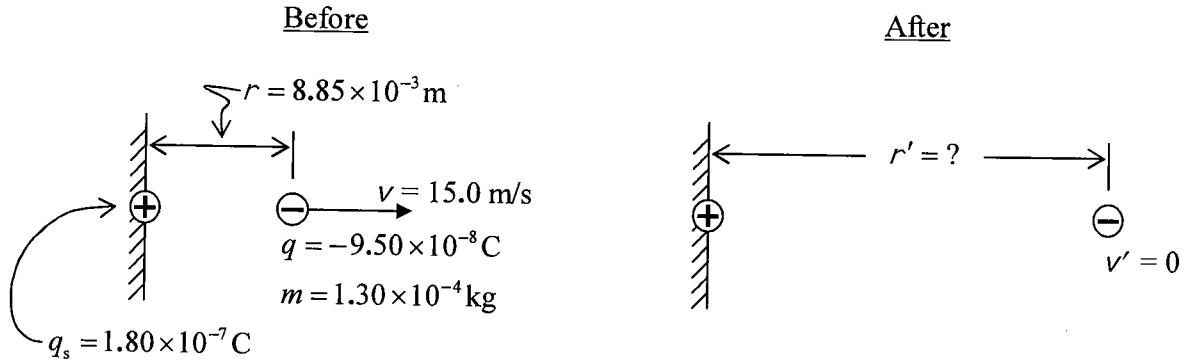
$$\boxed{W = -\left(\frac{kq'q}{x_2} - \frac{kq'q}{x_1}\right)}$$

The electric potential energy of a particle, used in conjunction with the principle of the conservation of mechanical energy, is a powerful problem-solving tool. The following example makes this evident:

Example

A particle of charge 0.180 μC is fixed in space by unspecified means. A particle of charge −0.0950 μC and mass 0.130 grams is 0.885 cm away from the first particle and moving directly away from the first particle with a speed of 15.0 m/s. How far away from the first particle does the second particle get?

This is a conservation of energy problem. As required for all conservation of energy problems, we start with a before and after diagram:



Energy Before = Energy After

$$K + U = K' + U'$$

$$K + q\varphi = \cancel{K'}^{0} + q\varphi'$$

$$\frac{1}{2}mv^2 + q\frac{kq_s}{r} = q\frac{kq_s}{r'}$$

$$\frac{1}{r'} = \frac{1}{r} + \frac{mv^2}{2kq_s q}$$

$$r' = \frac{1}{\frac{1}{r} + \frac{mv^2}{2kq_s q}}$$

$$r' = \frac{1}{\frac{1}{8.85\times10^{-3}\text{ m}} + \frac{1.30\times10^{-4}\text{ kg}\ (15.0\text{ m/s})^2}{2(8.99\times10^{9}\ \frac{\text{N}\cdot\text{m}^2}{\text{C}^2})\ 1.80\times10^{-7}\text{ C}\ (-9.50\times10^{-8}\text{ C})}}$$

$$r' = 0.05599\text{ m}$$

$$r' = 0.0560\text{ m}$$

$$\boxed{r' = 5.60\text{ cm}}$$

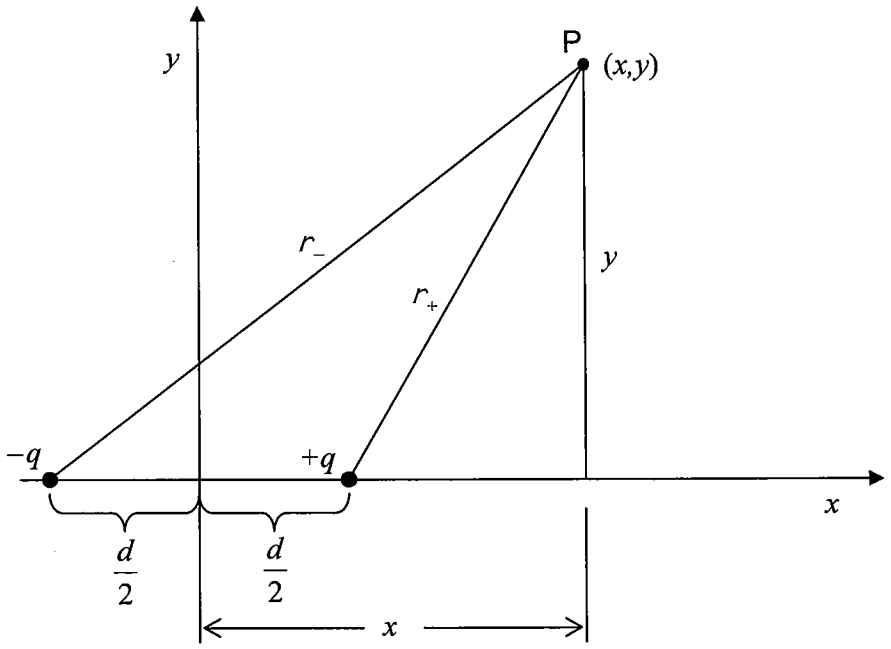
Superposition in the Case of the Electric Potential

When there is more than one charged particle contributing to the electric potential at a point in space, the electric potential at that point is the sum of the contributions due to the individual charged particles. The electric potential at a point in space, due to a set of several charged particles, is easier to calculate than the electric field due to the same set of charged particles is. This is true because the sum of electric potential contributions is an ordinary arithmetic sum, whereas, the sum of electric field contributions is a vector sum.

Example

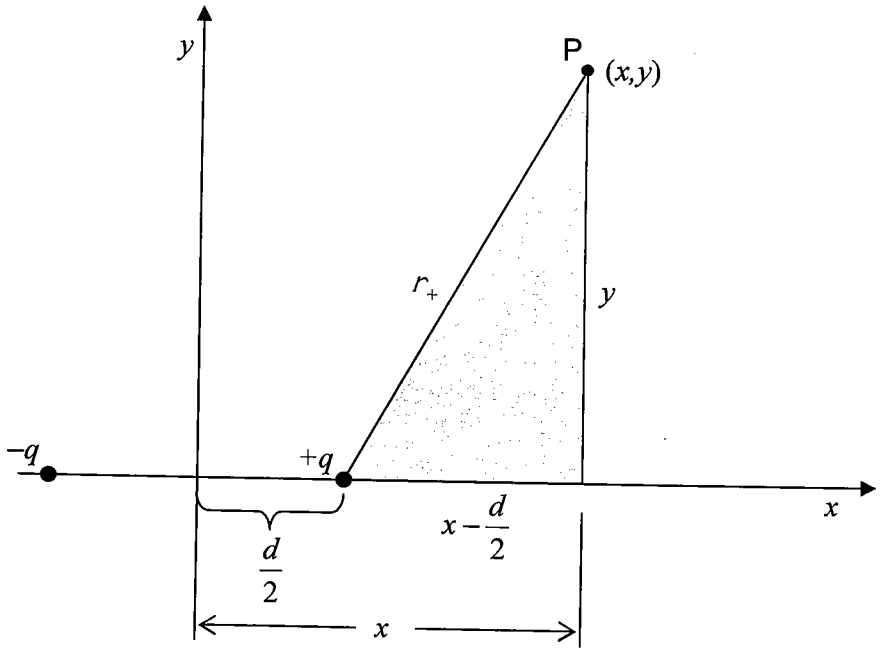
Find a formula that gives the electric potential at any point (x, y) on the x-y plane, due to a pair of particles: one of charge $-q$ at $\left(-\frac{d}{2}, 0\right)$ and the other of charge $+q$ at $\left(\frac{d}{2}, 0\right)$.

Solution: We establish a point P at an arbitrary position (x, y) on the x-y plain and determine the distance that point P is from each of the charged particles. In the following diagram, I use the symbol r_+ to represent the distance that point P is from the positively-charged particle, and r_- to represent the distance that point P is from the negatively-charged particle.



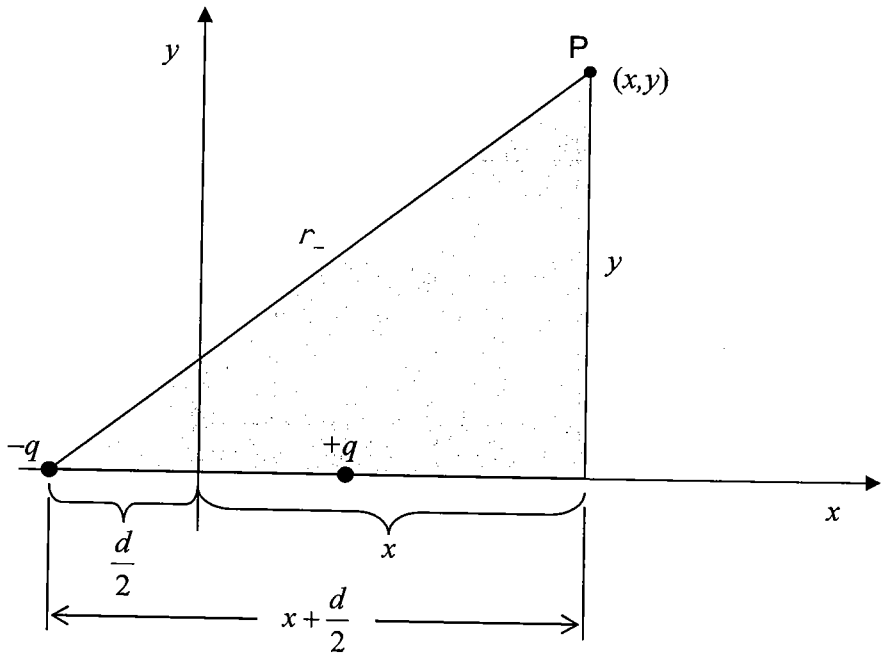
Analysis of the shaded triangle in the diagram at right gives us r_+.

$$r_+ = \sqrt{\left(x - \frac{d}{2}\right)^2 + y^2}$$



Analysis of the shaded triangle in the diagram at right gives us r_-.

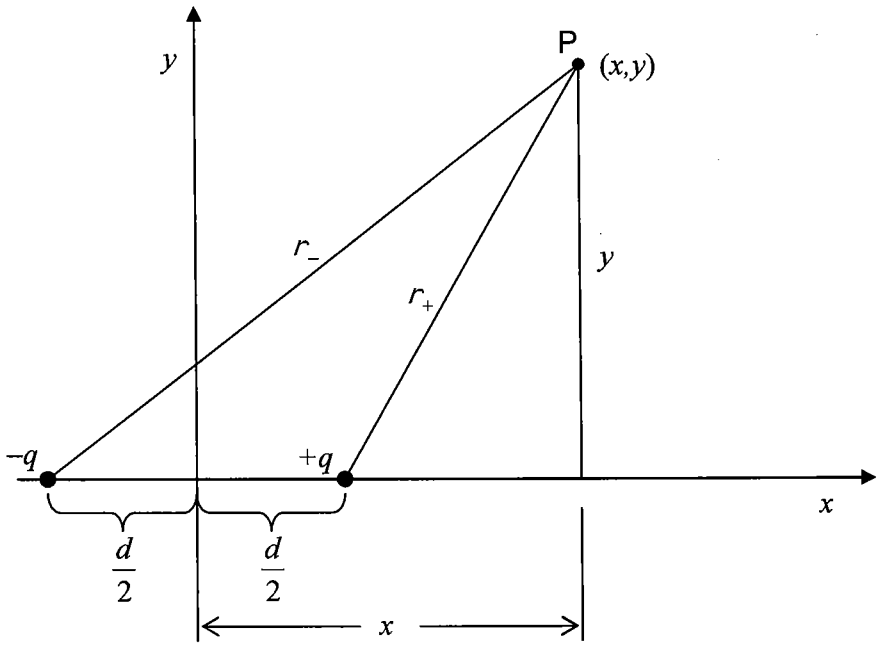
$$r_- = \sqrt{\left(x + \frac{d}{2}\right)^2 + y^2}$$



With the distances that point P is from each of the charged particles in hand, we are ready to determine the potential:

$$r_+ = \sqrt{\left(x - \frac{d}{2}\right)^2 + y^2}$$

$$r_- = \sqrt{\left(x + \frac{d}{2}\right)^2 + y^2}$$



$$\varphi(x,y) = \varphi_+ + \varphi_-$$

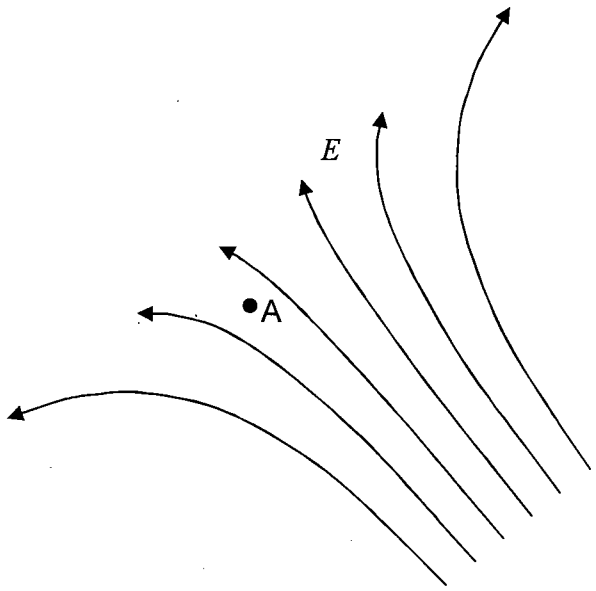
$$\varphi(x,y) = \frac{kq}{r_+} + \frac{k(-q)}{r_-}$$

$$\varphi(x,y) = \frac{kq}{r_+} - \frac{kq}{r_-}$$

$$\varphi(x,y) = \frac{kq}{\sqrt{\left(x - \frac{d}{2}\right)^2 + y^2}} - \frac{kq}{\sqrt{\left(x + \frac{d}{2}\right)^2 + y^2}}$$

7 Equipotential Surfaces, Conductors, and Voltage

Consider a region of space in which there exists an electric field. Focus your attention on a specific point in that electric field, call it point A.



Imagine placing a positive test charge at point A. (Assume that, by means not specified, you can move the test charge anywhere you want to.) Please think about the answer to the following question before reading on: Is it possible for you to move the test charge around in the electric field in such a manner that the electric field does no work on the test charge?

If we move the positive test charge in the "downfield" direction (toward the upper left corner of the diagram), there will be a positive amount of work (force-along-the-path times the length of the path) done on the test charge. And, if we move the positive test charge in the "upfield" direction there will be a negative amount of work done on it. But, if we move the positive test charge at right angles to the electric field, no work is done on it. That is, if we choose a path for the positive test charge such that every infinitesimal displacement of the particle is normal to the electric field at the location of the particle when it (the particle) undergoes said infinitesimal displacement, then the work done on the test charge, by the electric field, is zero. The set of all points that can be reached by such paths makes up an infinitesimally thin shell, a surface, which is everywhere perpendicular to the electric field. In moving a test charge along the surface from one point (call it point A) to another point (call it point B) on the surface, the work done is zero because the electric field is perpendicular to the path at all points along the path. Let's (momentarily) call the kind of surface we have been discussing a "zero-work surface." We have constructed the surface by means of force-along-the-path times the length-of-the-path work considerations. But the work done by the electric field when a test charge is moved from point A on the surface to point B on the surface must also turn out to be zero if we calculate it as the

negative of the change in the potential energy of the test charge. Let's do that and see where it leads us. We know that the work $W = 0$.

Also

$$W = -\Delta U$$

$$W = -(U_B - U_A)$$

In terms of the electric potential φ, $U = q\varphi$ so the work can be expressed as

$$W = -(q\varphi_B - q\varphi_A)$$

$$W = -q(\varphi_B - \varphi_A)$$

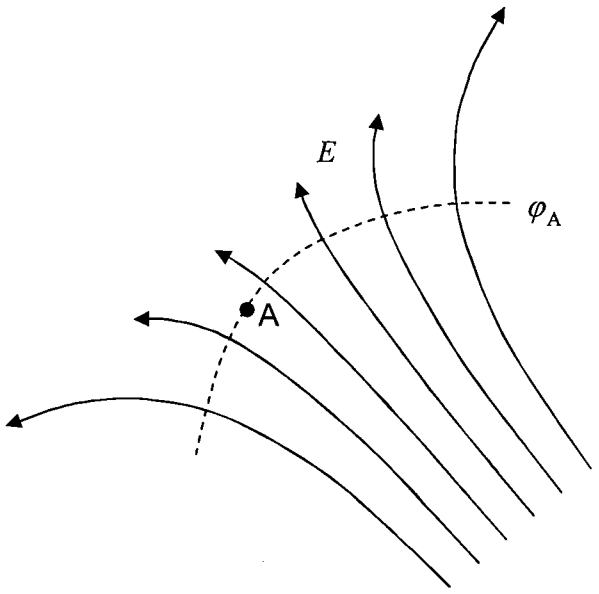
Given that $W = 0$, this means that

$$0 = -q(\varphi_B - \varphi_A)$$

$$\varphi_B - \varphi_A = 0$$

$$\varphi_B = \varphi_A$$

This is true for any point B on the entire "zero-work" surface. This means that every point on the entire surface is at the same value of electric potential. Thus a "zero-work" surface is also an *equipotential surface*. Indeed, this is the name (equipotential surface) that physicists use for such a surface. An equipotential surface is typically labeled with the corresponding potential value (φ_A in the case at hand). In the following diagram, the dashed curve represents the equipotential surface viewed edge on.



Summarizing:

- An equipotential surface is an imaginary surface on which every point has one and the same value of electric potential.
- An equipotential surface is everywhere perpendicular to the electric field that it characterizes.
- The work done by the electric field on a particle when it is moved from one point on an equipotential surface to another point on the same equipotential surface is always zero.

Perfect Conductors and the Electric Potential

Please recall what you know about perfect conductors and the electric *field*. Namely, that everywhere inside and on a perfect conductor, the electric field is zero. This goes for solid conductors as well as hollow, empty shells of perfectly conducting material. This means that the work done by the electric field on a test charge that is moved from one point in or on a perfect conductor (consider this to be a thought experiment), to another point in or on the same conductor, is zero. This means that the difference in the electric *potential* between any two points[1] in or on a perfect conductor must be zero. This means that the electric potential at every point in and on a perfect conductor must have one and the same value. Note that the value is *not*, in general, zero.

Some Electric Potential Jargon

When we talk about the electric potential in the context of a perfect conductor (or an object that approximates a perfect conductor), because every point in and on the conductor has the same value of electric potential, we typically call that value the electric potential *of the conductor*. We also use expressions such as, "the conductor is at a potential of 25 volts," meaning that the value of electric potential at every point in and on the conductor is 25 volts with respect to infinity (meaning that the zero of electric potential is at an infinite distance from the conductor) and/or with respect to "ground" (meaning that the potential of the earth is the zero of electric potential).

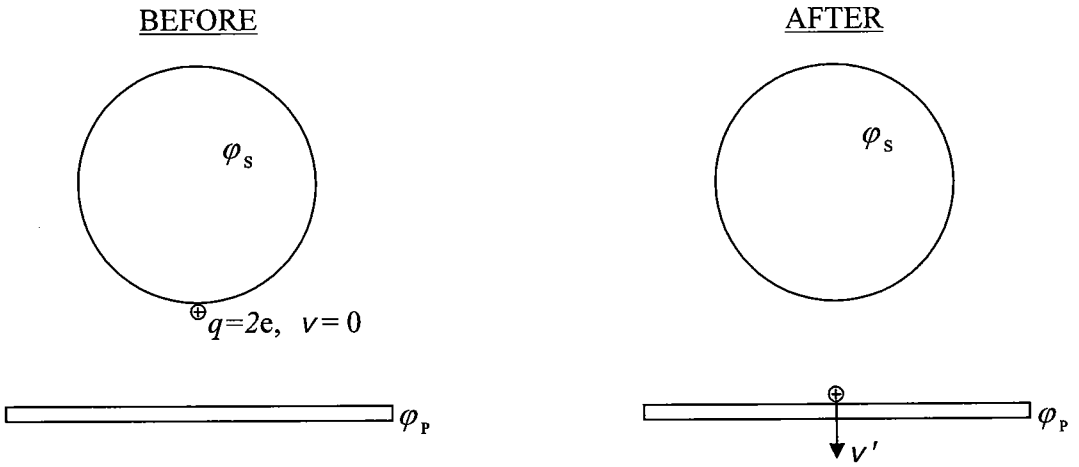
Electric Potential Difference, a.k.a. Voltage

In general, what is at issue when one talks about conductors and electric potential is not the value of the electric potential of a conductor, but rather, the electric potential difference between one conductor and another.

[1] The "difference in the electric potential between points A and B" is the value of the electric potential at B minus the value of the electric potential at point A.

Example 7-1

A hollow metal sphere is at a potential that is 472 volts higher than that of a nearby metal plate. A particle of charge 2e is released from rest at the surface of the sphere. It subsequently strikes the plate. With what kinetic energy does the charged particle strike the plate? (Assume that the only force acting on the particle is that due to the electric field corresponding to the given information.)



(Given $\varphi_S - \varphi_P = \Delta\varphi = 472$ volts.)

$$\textit{Energy Before} = \textit{Energy After}$$

$$\cancel{K}^{0} + U = K' + U'$$

$$q\varphi_S = K' + q\varphi_P$$

$$K' = q\varphi_S - q\varphi_P$$

$$K' = q(\varphi_S - \varphi_P)$$

$$K' = q\,\Delta\varphi$$

$$K' = 2e\ (472 \text{ volts})$$

$$\boxed{K' = 944 \text{ eV}}$$

Note that in the solution to the example problem, we never needed to know the value of the electric potential of either the sphere or the plate, only the difference between the two potentials. There is a device which can be used to measure the potential difference between two points in space. The device is called a voltmeter. A typical voltmeter consists of a box with two wires extending from it. On the end of each wire is a short metal wand called a probe. Each wire and each probe, except for the tip of the probe, is covered with insulating material. The box displays, either by means of a digital readout or the position of a needle, the potential difference between

the two wires. In typical use, one presses the metal tip of one probe against a conductor of interest and holds the tip there. That causes that probe and wire to be at the same potential as the conductor. One presses the tip of the other probe against another conductor. This causes that probe and wire to be at the potential of the second conductor. With each probe in contact with a conductor, the voltmeter continually displays the potential difference between the two conductors.

Based on the SI units of measurement, the electric potential difference between two points in space goes by another name, namely, *voltage*. Voltage means *electric potential difference* which means, the difference between the electric-potential-energy-per-charge-of-would-be-victim at one point in space and the electric-potential-energy-per-charge-of-would-be-victim at another point in space. While *voltage* literally means potential difference, the word is also, quite often used to mean electric potential itself, where, one particular conductor or point in space is defined to be the zero of potential. If no conductor or point in space has been defined to be the zero, then it is understood that "infinity" is considered to be at the zero of electric potential. So, if you read that a metal object is at a potential of 230 volts (when no conductor or point in space has been identified as the zero of electric potential), you can interpret the statement to mean the same thing as a statement that the electric potential of the metal object is 230 volts higher than the electric potential at any point that is an infinite distance away from the object.

As you move on in your study of physics, onward to your study and work with electric circuits, it is important to keep in mind that voltage, in a circuit, is the difference in the value of a characteristic (the electric potential) of one conductor, and the value of the same characteristic (electric potential) of another conductor.

Analogy Between Voltage and Altitude

One can draw a pretty good analogy between voltage (electric potential) and altitude. Consider a particular altitude above the surface of the earth (measured, for instance, from sea level). The value of the altitude characterizes a point in space or a set of points in space. In fact, the set of all points in space that are at the same altitude above the surface of the earth forms an "equi-altitude" surface. On a local scale, we can think of that "equi-altitude" surface as a plane. On a global scale, looking at the big picture, we recognize it to be a spheroidal shell. Flocks of birds can be at that altitude and when they are, we attribute the altitude to the flock of birds. We say that the flock of birds has such and such an altitude. But, whether or not the flock of birds is there, the altitude exists. Regarding a particular altitude, we can have birds and air and clouds moving or flowing through space at that altitude, but the altitude itself just exists—it doesn't flow or go anywhere. This is like the voltage in a circuit. The voltage in a circuit exists. The voltage characterizes a conductor in a circuit. Charged particles can move and flow in and through a conductor that is at that voltage, but, the voltage doesn't flow or go anywhere, any more than altitude flows or goes anywhere.

8 Capacitors, Dielectrics, and Energy in Capacitors

Capacitance is a characteristic of a conducting object. Capacitance is also a characteristic of a pair of conducting objects.

Let's start with the capacitance of a single conducting object, isolated from its surroundings. Assume the object to be neutral. Now put some positive charge on the object. The electric potential of the object is no longer zero. Put some more charge on the object and the object will have a higher value of electric potential. What's interesting is, no matter how much, or how little charge you put on the object, the ratio of the amount of charge q on the object to the resulting electric potential φ of the object has one and the same value.

$$\frac{q}{\varphi} \text{ has the same value for any value of } q.$$

You double the charge, and, the electric potential doubles. You reduce the amount of charge to one tenth of what it was, and, the electric potential becomes one tenth of what it was. The actual value of the unchanging ratio is called the capacitance C_{sc} of the object (where the subscript "sc" stands for "single conductor").

$$C_{sc} = \frac{q}{\varphi} \tag{8-1}$$

where:

C_{sc} is the capacitance of a single conductor, isolated (distant from) its surroundings,
q is the charge on the conductor, and,
φ is the electric potential of the conductor relative to the electric potential at infinity (the position defined for us to be our zero level of electric potential).

The capacitance of a conducting object is a property that an object has even if it has no charge at all. It depends on the size and shape of the object.

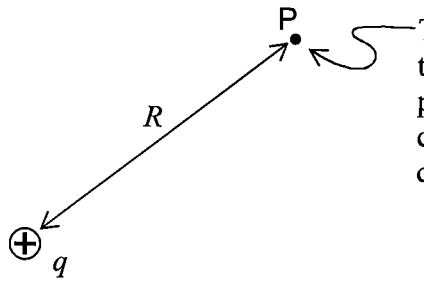
The more positive charge you need to add to an object to raise the potential of that object 1 volt, the greater the capacitance of the object. In fact, if you define q_1 to be the amount of charge you must add to a particular conducting object to increase the electric potential of that object by one volt, then the capacitance of the object is $\frac{q_1}{1 \text{ volt}}$.

The Capacitance of a Spherical Conductor

Consider a sphere (either an empty spherical shell or a solid sphere) of radius R made out of a perfectly-conducting material. Suppose that the sphere has a positive charge q and that it is isolated from its surroundings. We have already covered the fact that the electric field of the charged sphere, from an infinite distance away, all the way to the surface of the sphere, is

indistinguishable from the electric field due to a point charge q at the position of the center of the sphere; and; everywhere inside the surface of the sphere, the electric field is zero. Thus, outside the sphere, the electric *potential* must be identical to the electric potential due to a point charge at the center of the sphere (instead of the sphere). Working your way in from infinity, however, as you pass the surface of the sphere, the electric potential no longer changes. Whatever the value of electric potential at the surface of the sphere is, that is the value of electric potential at every point inside the sphere.

This means that the electric potential of the sphere is equal to the electric potential that would be caused by a point charge (all by itself) at a point in space a distance R from the point charge (where R is the radius of the sphere).



The value of the electric potential here is the same as the value of the electric potential at every point in and on a conducting sphere of radius R and charge q.

Thus, $\varphi = \frac{kq}{R}$ is the electric potential of a conducting sphere of radius R and charge q.

Solving this expression for $\frac{q}{\varphi}$ yields:

$$\frac{q}{\varphi} = \frac{R}{k}$$

Since, by definition, the capacitance $C_{sc} = \frac{q}{\varphi}$, we have:

$$C_{sc} = \frac{R}{k} \qquad (8\text{-}2)$$

The capacitance of a conducting sphere is directly proportional to the radius of the sphere. The bigger the sphere, the more charge you have to put on it to raise its potential one volt (in other words, the bigger the capacitance of the sphere). This is true of conducting objects in general. Since all the unbalanced charge on a conductor resides on the surface of the conductor, it really has to do with the amount of surface area of the object. The more surface area, the more room

the charge has to spread out and, therefore, the more charge you have to put on the object to raise its potential one volt (in other words, the bigger the capacitance of the object).

Consider, for instance, a typical paper clip. It only takes an amount of charge on the order of a pC (picocoulomb, 1×10^{-12} coulombs) to raise the potential of a paper clip 10 volts.

Units

The unit of capacitance is the coulomb-per-volt, $\frac{\text{C}}{\text{V}}$. That combination unit is given a name, the farad, abbreviated F.

$$1\,\text{F} = 1\frac{\text{C}}{\text{V}}$$

The Capacitance of a Pair of Conducting Objects

So far, we've been talking about the capacitance of a conducting object that is isolated from its surroundings. You put some charge on such an object, and, as a result, the object takes on a certain value of electric potential. The charge-to-potential ratio is called the capacitance of the object. But get this, if the conductor is near another conductor when you put the charge on it, the conductor takes on a different value of electric potential (compared to the value it takes on when it is far from all other conductors) for the exact same amount of charge. This means that just being in the vicinity of another conductor changes the effective capacitance[1] of the conductor in question. In fact, if you put some charge on an isolated conductor, and then bring another conductor into the vicinity of the first conductor, the electric potential of the first conductor will change, meaning, its effective capacitance changes. Let's investigate a particular case to see how this comes about.

Consider a conducting sphere with a certain amount of charge, q, on it. Suppose that, initially, the sphere is far from its surroundings and, as a result of the charge on it, it is at a potential φ.

Let's take a moment to review what we mean when we say that the sphere is at a potential φ. Imagine that you take a test charge q_T from a great distance away from the sphere and take it to the surface of the sphere. Then you will have changed the potential energy of the test charge from zero to $q_T\varphi$. To do that, you have to do an amount of work $q_T\varphi$ on the test charge. We're assuming that the test charge was initially at rest and is finally at rest. You have to push the

[1] By definition, the capacitance of a single conducting object is the charge-to-voltage ratio when the object is isolated (far away from) its surroundings. When it is near another conductor, we generally talk about the capacitance of the pair of conductors (as we do later in this chapter) rather than what I have been calling the "effective capacitance" of one of the conductors.

charge onto the sphere. You apply a force over a distance to give that particle the potential energy $q_T\varphi$. You do positive work on it. The electric field of the sphere exerts a force on the test charge in the opposite direction to the direction in which you are moving the test charge. The electric field does a negative amount of work on the test charge such that the total work, the work done by you plus the work done by the electric field, is zero (as it must be since the kinetic energy of the test charge does not change). But I want you to focus your attention on the amount of work that you must do, pushing the test charge in the same direction in which it is going, to bring the test charge from infinity to the surface of the sphere. That amount of work is $q_T\varphi$ because $q_T\varphi$ is the amount by which you increase the potential energy of the charged particle. If you were to repeat the experiment under different circumstances and you found that you did not have to do as much work to bring the test charge from infinity to the surface of the sphere, then you would know that the sphere is at a lower potential than it was the first time.

Now, we are ready to explore the case that will illustrate that the charge-to-voltage ratio of the conducting object depends on whether or not there is another conductor in the vicinity. Let's bring an identical conducting sphere near one side of the first sphere. The first sphere still has the same amount of charge q on it that it always had, and, *the second sphere is neutral.* The question is, "Is the potential of the original sphere still the same as what it was when it was all alone?" Let's test it by bringing a charge in from an infinite distance on the opposite side of the first sphere (as opposed to the side to which the second sphere now resides). Experimentally we find that it takes less work to bring the test charge to the original sphere than it did before, meaning that the original sphere now has a lower value of electric potential. How can that be? Well, when we brought the second sphere in close to the original sphere, the second sphere became polarized. (Despite the fact that it is neutral, it is a conductor so the balanced charge in it is free to move around.) The original sphere, having positive charge q, attracts the negative charge in the second sphere and repels the positive charge. The near side of the second sphere winds up with a negative charge and the far side, with the same amount of positive charge. (The second sphere remains neutral overall.) Now the negative charge on the near side of the second sphere attracts the (unbalanced) positive charge on the original sphere to it. So the charge on the original sphere, instead of being spread out uniformly over the surface as it was before the second sphere was introduced, is bunched up on the side of the original sphere that is closer to the second sphere. This leaves the other side of the original sphere, if not neutral, at least less charged than it was before. As a result, it takes less work to bring the positive test charge in from infinity to that side of the original sphere. As mentioned, this means that the electric potential of the original sphere must be lower than it was before the second sphere was brought into the picture. Since it still has the same charge that it always had, the new, lower potential, means that the original sphere has a greater charge-to-potential ratio, and hence a greater effective capacitance.

In practice, rather than call the charge-to-potential ratio of a conductor that is near another conductor, the "effective capacitance" of the first conductor, we define a capacitance for the pair of conductors. Consider a pair of conductors, separated by vacuum or insulating material, with a given position relative to each other. We call such a configuration a capacitor. Start with both conductors being neutral. Take some charge from one conductor and put it on the other. The amount of charge moved from one conductor to the other is called the charge of the capacitor. (Contrast this with the actual total charge of the device which is still zero.) As a result of the

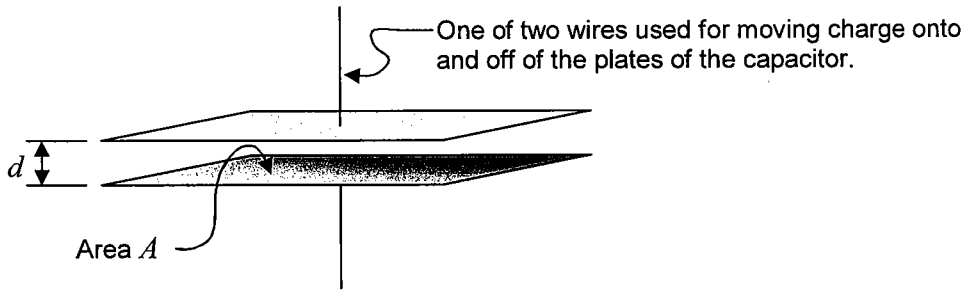
repositioning of the charge, there is a potential difference between the two conductors. This potential difference $\Delta\varphi$ is called the voltage of the capacitor or, more often, the voltage across the capacitor. We use the symbol V to represent the voltage across the capacitor. In other words, $V \equiv \Delta\varphi$. The ratio of the amount of charge moved from one conductor to the other, to, the resulting potential difference of the capacitor, is the capacitance of the capacitor (the pair of conductors separated by vacuum or insulator).

$$C = \frac{q}{V} \tag{8-3}$$

where:

- C is the capacitance of a capacitor, a pair of conductors separated by vacuum or an insulating material,
- q is the "charge on the capacitor," the amount of charge that has been moved from one initially neutral conductor to the other. One conductor of the capacitor actually has an amount of charge q on it and the other actually has an amount of charge –q on it.
- V is the electric potential difference $\Delta\varphi$ between the conductors. It is known as the voltage of the capacitor. It is also known as the voltage across the capacitor.

A two-conductor capacitor plays an important role as a component in electric circuits. The simplest kind of capacitor is the parallel-plate capacitor. It consists of two identical sheets of conducting material (called plates), arranged such that the two sheets are parallel to each other. In the simplest version of the parallel-plate capacitor, the two plates are separated by vacuum.



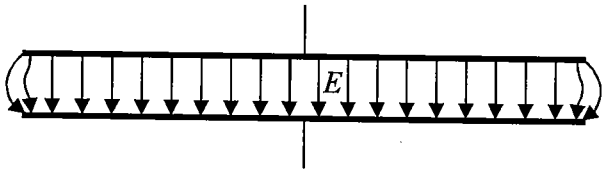
The capacitance of such a capacitor is given by

$$C = \epsilon_o \frac{A}{d}$$

where:

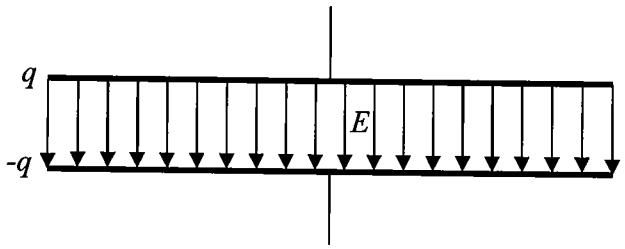
- C is the capacitance of the parallel-plate capacitor whose plates are separated by vacuum,
- d is the distance between the plates,
- A is the area of one face of one of the plates,
- ϵ_o is a universal constant called the permittivity of free space. ϵ_o is closely related to the Coulomb constant k. In fact, $k = \frac{1}{4\pi\epsilon_o}$. Thus, $\epsilon_o = 8.85\times10^{-12}\frac{C^2}{N\cdot m^2}$. Our equation for the capacitance can be expressed in terms of the Coulomb constant k as $C = \frac{1}{4\pi k}\frac{A}{d}$, but, it is more conventional to express the capacitance in terms of ϵ_o.

This equation for the capacitance is an approximate formula. It is a *good* approximation as long as the plate separation d is small compared to a representative plate dimension (the diameter in the case of circular plates, the smaller edge length in the case of rectangular plates). The derivation of the formula is based on the assumption that the electric field, in the region between the plates is uniform, and the electric field outside that region is zero. In fact, the electric field is not uniform in the vicinity of the edges of the plates. As long as the region in which the electric field is not well-approximated by a uniform electric field is small compared to the region in which it is, our formula for the capacitance is good.



The Effect of Insulating Material Between the Plates of a Capacitor

To get at the effect of insulating material, rather than vacuum, between the plates of a capacitor, I need to at least outline the derivation of the formula $C = \epsilon_o \frac{A}{d}$. Keep in mind that the capacitance is the charge-per-voltage of the capacitor. Suppose that we move charge q from one initially-neutral plate to the other. We assume that the electric field is uniform between the plates of the capacitor and zero elsewhere.



V (the potential difference between the plates, a.k.a. the voltage across the capacitor)

By means that you will learn about later in this book we establish that the value of the electric field (valid everywhere between the plates) is given by:

$$E = \frac{q}{A\,\epsilon_o} \tag{8-4}$$

Also, we know that the work done on a test charge q_T by the electric field when the test charge is moved from the higher-potential plate to the lower-potential plate is the same whether we calculate it as force-along the path times the length of the path, or, as the negative of the change

in the potential energy. This results in a relation between the electric field and the electric potential as follows:

W calculated as force times distance = W calculated as minus change in potential energy

$$F\Delta x = -\Delta U$$
$$q_T E d = -q_T \Delta\varphi$$
$$Ed = -(-V)$$
$$V = Ed$$

Using equation 8-4 ($E = \frac{q}{A\epsilon_o}$) to replace the E in $V = Ed$ with $\frac{q}{A\epsilon_o}$ gives us:

$$V = \frac{q}{A\epsilon_o} d$$

Solving this for q/V yields

$$\frac{q}{V} = \epsilon_o \frac{A}{d}$$

for the charge-to-voltage ratio. Since the capacitance is the charge-to-voltage ratio, this means

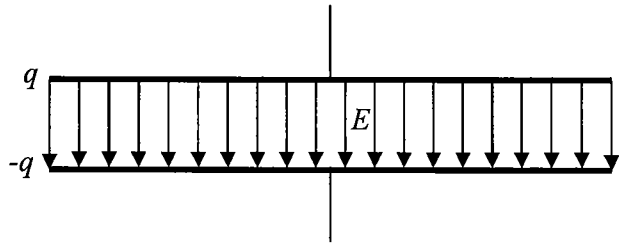
$$C = \epsilon_o \frac{A}{d}$$

which is what we set out to derive.

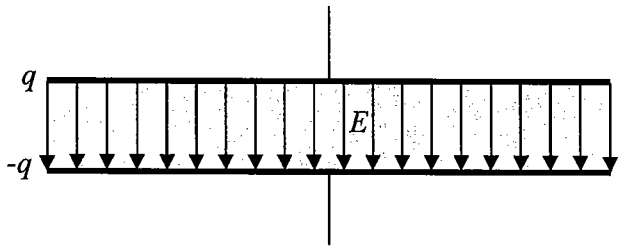
Okay now, here's the deal on having an insulator between the plates: Consider a capacitor that is identical in all respects to the one we just dealt with, except that there is an insulating material between the plates, rather than vacuum. Further suppose that the capacitor has the same amount of charge q on it as the vacuum-between-the-plates capacitor had on it. *The presence of the insulator between the plates results in a weaker electric field between the plates.* This means that a test charge moved from one plate to another would have less work done on it by the electric field, meaning that it would experience a smaller change in potential energy, meaning the electric potential difference between the plates is smaller. So, with the same charge, but a smaller potential difference, the charge-to-voltage ratio (that is, the capacitance of the capacitor) must be *bigger*.

The presence of the insulating material makes the capacitance bigger. The part of the preceding argument that still needs explaining is that part about the insulating material weakening the electric field. Why does the insulating material make the field weaker? Here's the answer:

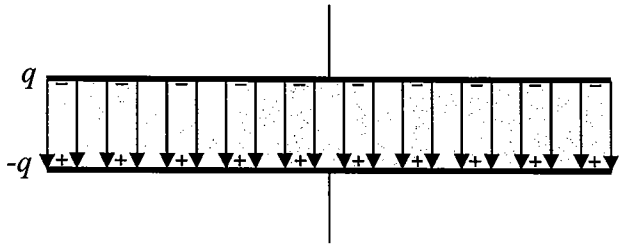
Starting with vacuum between the plates,



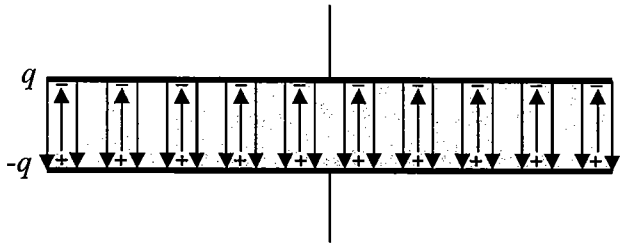
we insert some insulating material:



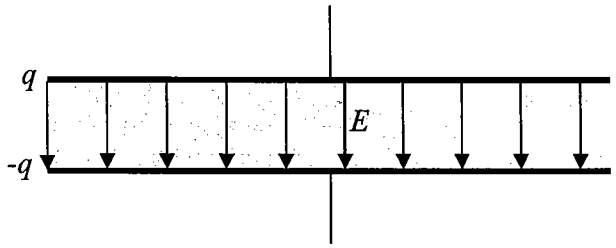
The original electric field polarizes the insulating material:



The displaced charge creates an electric field of its own, in the direction opposite that of the original electric field:



The net electric field, being at each point in space, the vector sum of the two contributions to it, is in the same direction as the original electric field, but weaker than the original electric field:



This is what we wanted to show. The presence of the insulating material makes for a weaker electric field (for the same charge on the capacitor), meaning a smaller potential difference, meaning a bigger charge-to-voltage ratio, meaning a bigger capacitance. How much bigger depends on how much the insulator is polarized which depends on what kind of material the insulator consists of. An insulating material, when placed between the plates of a capacitor is called a *dielectric*. The net effect of using a dielectric instead of vacuum between the plates is to multiply the capacitance by a factor known as the dielectric constant. Each dielectric is characterized by a unitless dielectric constant specific to the material of which the dielectric is made. The capacitance of a parallel-plate capacitor which has a dielectric in between the plates, rather than vacuum, is just the dielectric constant κ times the capacitance of the same capacitor with vacuum in between the plates.

$$C = \kappa \epsilon_0 \frac{A}{d} \tag{8-5}$$

where:

C is the capacitance of the parallel-plate capacitor whose plates are separated by an insulating material,
κ is the dielectric constant characterizing the insulating material between the plates,
d is the distance between the plates,
A is the area of one face of one of the plates, and
ϵ_0 is a universal constant called the permittivity of free space.

Calling the dielectric constant for vacuum 1 (exactly one), we can consider this equation to apply to all parallel-plate capacitors. Some dielectric constants of materials used in manufactured capacitors are provided in the following table:

Substance	**Dielectric Constant**
Air	1.00
Aluminum Oxide (a corrosion product found in many electrolytic capacitors)	7
Mica	3-8
Titanium Dioxide	114
Vacuum	1 (exactly)
Waxed Paper	2.5-3.5

Energy Stored in a Capacitor

Moving charge from one initially-neutral capacitor plate to the other is called *charging* the capacitor. When you charge a capacitor, you are storing energy in that capacitor. Providing a conducting path for the charge to go back to the plate it came from is called discharging the capacitor. If you discharge the capacitor through an electric motor, you can definitely have that charge do some work on the surroundings. So, how much energy is stored in a charged capacitor? Imagine the charging process. You use some force to move some charge over a distance from one plate to another. At first, it doesn't take much force because both plates are neutral. But the more charge that you have already relocated, the harder it is to move more charge. Think about it. If you are moving positive charge, you are pulling positive charge from a negatively charged plate and pushing it onto a positively charged plate. The total amount of work you do in moving the charge is the amount of energy you store in the capacitor. Let's calculate that amount of work.

In this derivation, I am going to use a lower case q to represent the variable amount of charge on the capacitor plate (it increases as we charge the capacitor), and an upper case Q to represent the final amount of charge. Similarly, I choose to use a lower case v to represent the variable amount of voltage across the capacitor (it too increases as we charge the capacitor), and the upper case V to represent the final voltage across the capacitor. Let U represent the energy stored in the capacitor:

$$dU = v\,dq$$

but the voltage across the capacitor is related to the charge of the capacitor by $C = q/v$ which, solved for v is $v = q/C$, so:

$$dU = \frac{q}{C}dq$$

$$\int dU = \frac{1}{C}\int_0^Q q\,dq$$

$$U = \frac{1}{C}\frac{q^2}{2}\Big|_0^Q$$

$$U = \frac{1}{C}\left(\frac{Q^2}{2} - \frac{0^2}{2}\right)$$

$$U = \frac{1}{2}\frac{1}{C}Q^2$$

Using $C = Q/V$, we can also express the energy stored in the capacitor as $U = \frac{1}{2}QV$, or

$$U = \frac{1}{2}CV^2 \qquad (8\text{-}6)$$

9 Electric Current, EMF, Ohm's Law

We now begin our study of electric circuits. A circuit is a closed conducting path through which charge flows. In circuits, charge goes around in loops. The charge flow rate is called electric current. A circuit consists of circuit elements connected together by wires. A capacitor is an example of a circuit element with which you are already familiar. We introduce some more circuit elements in this chapter. In analyzing circuits, we treat the wires as perfect conductors and the circuit elements as ideal circuit elements. There is a great deal of variety in the complexity of circuits. A computer is a complicated circuit. A flashlight is a simple circuit.

The kind of circuit elements that you will be dealing with in this course are two-terminal circuit elements. There are several different kinds of two-terminal circuit elements but they all have some things in common. A two-terminal circuit element is a device with two ends, each of which is a conductor. The two conductors are called terminals. The terminals can have many different forms. Some are wires, some are metal plates, some are metal buttons, and some are metal posts. One connects wires to the terminals to make a circuit element part of a circuit.

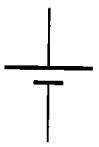
An important two-terminal circuit element is a seat of EMF[1]. You can think of a seat of EMF as an ideal battery or as an ideal power supply. What it does is to maintain a constant potential difference (a.k.a. a constant voltage) between its terminals. One uses either the constant name $\mathcal{E}$ (script E) or the constant name V to represent that potential difference.

To achieve a potential difference $\mathcal{E}$ between its terminals, a seat of EMF, when it first comes into existence, has to move some charge (we treat the movement of charge as the movement of *positive* charge) from one terminal to the other. The "one terminal" is left with a net negative charge and "the other" acquires a net positive charge. The seat of EMF moves charge until the positive terminal is at a potential $\mathcal{E}$ higher than the negative terminal. Note that the seat of EMF does not produce charge; it just pushes existing charge around. If you connect an isolated wire to the positive terminal, then it is going to be at the same potential as the positive terminal, and, because the charge on the positive terminal will spread out over the wire, the seat of EMF is going to have to move some more charge from the lower-potential terminal to maintain the potential difference. One rarely talks about the charge on either terminal of a seat of EMF or on a wire connected to either terminal. A typical seat of EMF maintains a potential difference between its terminals on the order of 10 volts and the amount of charge that has to be moved, from one wire whose dimensions are similar to that of a paper clip, to another of the same sort, is on the order of a pC (1×10^{-12} C). Also, the charge pileup is almost instantaneous, so, by the time you finish connecting a wire to a terminal, that wire already has the charge we are talking about. In general, we don't know how much charge is on the positive terminal and whatever wire might be connected to it, and we don't care. It is minuscule. But, it is enough for the potential difference between the terminals to be the rated voltage of the seat of EMF.

[1] The reason for the name "seat of EMF" is of historical interest only. EMF stands for electromotive force. You would be better off calling it "ee em eff" and thinking of a so-called seat of EMF as a "maintainer of a constant potential difference".

You'll recall that electric potential is something that is used to characterize an electric field. In causing there to be a potential difference between its terminals and between any pair of wires that might be connected to its terminals, the seat of EMF creates an electric field. The electric field depends on the arrangement of the wires that are connected to the terminals of the seat of EMF. The electric field is another quantity that we rarely discuss in analyzing circuits. We can typically find out what we need to find out from the value of the potential difference $\mathcal{E}$ that the seat of EMF maintains between its terminals. But, the electric field does exist, and, in circuits, the electric field of the charge on the wires connected to the seat of EMF is what causes charge to flow in a circuit, and charge flow in a circuit is a huge part of what a circuit is all about.

We use the symbol

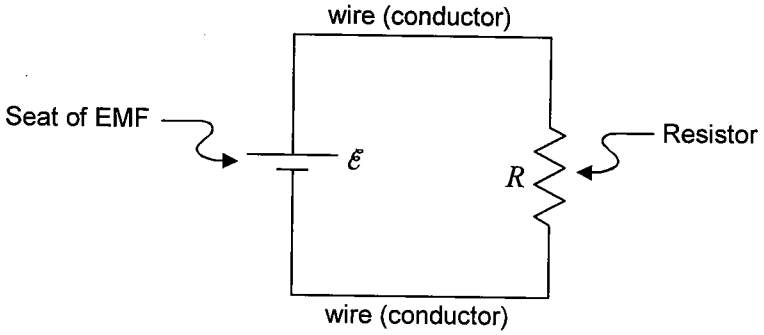


to represent a seat of EMF in a circuit diagram (a.k.a. a schematic diagram of a circuit) where the two collinear line segments represent the terminals of the seat of EMF, the one connected to the shorter of the parallel line segments being the negative, lower-potential, terminal; and; the one connected to the longer of the parallel line segments being the positive, higher-potential, terminal.

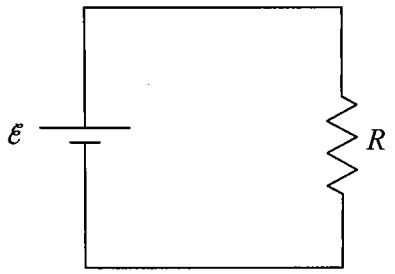
The other circuit element that I want to introduce in this chapter is the *resistor*. A resistor is a poor conductor. The resistance of a resistor is a measure of how poor a conductor the resistor is. The bigger the value of resistance, the more poorly the circuit element allows charge to flow through itself. Resistors come in many forms. The filament of a light bulb is a resistor. A toaster element (the part that glows red when the toaster is on) is a resistor. Humans manufacture small ceramic cylinders (with a coating of carbon and a wire sticking out each end) to have certain values of resistance. Each one has its value of resistance indicated on the resistor itself. The symbol

is used to represent a resistor in a circuit diagram. The symbol R is typically used to represent the value of the resistance of a resistor.

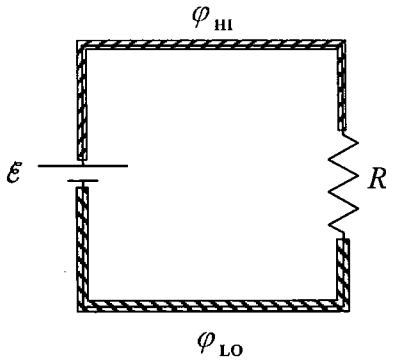
We are now ready to consider the following simple circuit:



Here it is again without so many labels:

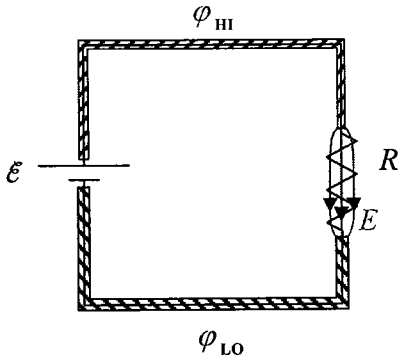


The upper wire (conductor) has one value of electric potential (call it φ_{HI}) and the lower wire has another value of electric potential (call it φ_{LOW}) such that the difference $\varphi_{HI} - \varphi_{LOW}$ is $\mathcal{E}$.



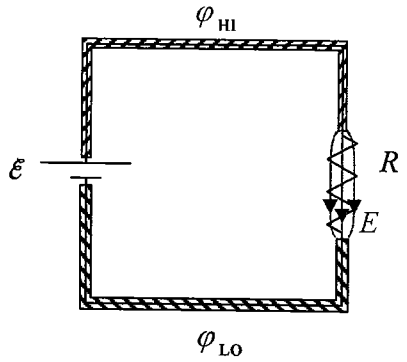
$$\varphi_{HI} - \varphi_{LOW} = \mathcal{E}$$

In order to maintain the potential difference $\mathcal{E}$ between the two conductors, the seat of EMF causes there to be a minuscule amount of positive charge on the upper wire and the same amount of negative charge on the lower wire. This charge separation causes an electric field in the resistor.



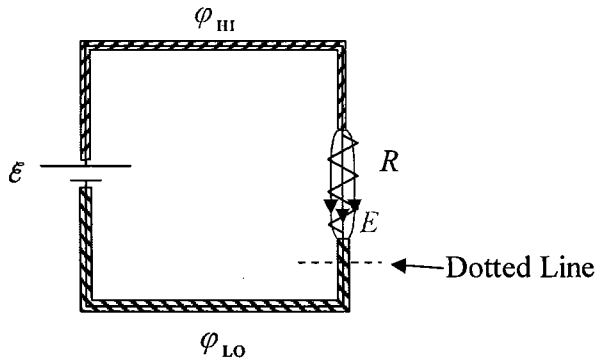
(We carry out this argument in the positive charge carrier model. While it makes no difference for the circuit, as a point of fact, it is actually negatively charged particles moving in the opposite direction. The effect is the same.)

It is important to realize that every part of the circuit is chock full of both kinds of charge. The wire, the resistor, everything is incredibly crowded with both positive and negative charge. One kind of charge can move against the background of the other. Now the electric field in the resistor pushes the positive charge in the resistor in the direction from the higher-potential terminal toward the lower-potential terminal.



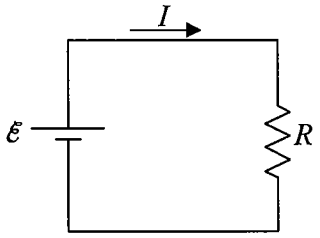
Pushing positive charge onto the lower-potential wire would tend to raise the potential of the lower-potential wire and leave the upper end of the resistor with a negative charge. I say "would" because any tendency for a change in the relative potential of the two wires is immediately compensated for by the seat of EMF. Remember, that's what the seat of EMF does, it maintains a constant potential difference between the wires. To do so in the case at hand, the seat of EMF must pull some positive charges from the lower-potential wire and push them onto the higher-potential wire. Also, any tendency of the upper end of the resistor to become negative immediately results in an attractive force on the positive charge in the higher-potential wire. This causes that positive charge to move down into the resistor in the place of the charge that just moved along the resistor toward the lower-potential wire. The net effect is a continual movement of charge, clockwise around the loop, as we view it in the diagram, with the net amount of charge in any short section of the circuit never changing. Pick a spot anywhere in the circuit. Just as fast as positive charge moves out of that spot, more positive charge from a neighboring spot moves in. What we have is this whole crowded mass of positive charge carriers moving clockwise around the loop, all because of the electric field in the resistor, and the EMF's "insistence" on maintaining a constant potential difference between the wires.

Now draw a dotted line across the path of the circuit, at any point in the circuit, as indicated below.



The rate at which charge crosses that line is the charge flow rate at that point (the point at which you drew the dotted line) in the circuit. The charge flow rate, how many coulombs-of-charge-per-second are crossing that line is called the *electric current* at that point. In the case at hand, because the whole circuit consists of a single loop, the current is the same at every point in the circuit—it doesn't matter where you "draw the line." The symbol that one typically uses to represent the value of the current is *I*.

In analyzing a circuit, if the current variable is not already defined for you, you should define it by drawing an arrow on the circuit and labeling it *I* or I with a subscript.



The units for current are coulombs per second (C/s). That combination of units is given a name: the ampere, abbreviated A.

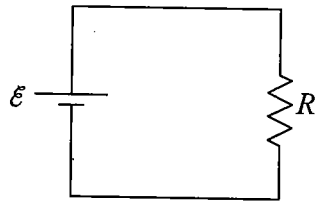
$$1\,\text{A} = 1\frac{\text{C}}{\text{s}}$$

Now about that resistor: In our positive charge carrier model, the charged particles that are free to move in the resistor experience a force exerted on them by the electric field, in the direction of the electric field. As a result, they experience acceleration. But, the background material making up the substance of which the charge carriers are a part, exerts a velocity-dependent retarding force on the charge carriers. The faster they go, the bigger the retarding force. Upon completion of the circuit (making that final wire-to-terminal connection), the charge carriers in the resistor, almost instantaneously, reach a terminal velocity at which the retarding force on a given charge carrier is

just as great as the force exerted by the electric field on that charge carrier. The value of the terminal velocity, along with the number-of-charge-carriers-per-volume in the resistor, and the cross-sectional area of the poorly-conducting material making up the resistor, determine the charge flow rate, the *current*, in the resistor. In the simple circuit under consideration, the charge flow rate in the resistor is the charge flow rate everywhere in the circuit.

The value of the terminal velocity itself depends on how strong the electric field is, and, on the nature of the retarding force. The nature of the retarding force depends on what kind of material the resistor is made of. One kind of material will result in a bigger terminal velocity for the same electric field as another kind of material. Even with one kind of material, there's the question of how the retarding force depends on the velocity. Is it proportional to the square of the velocity, the log of the velocity, or what? Experiment shows that in an important subset of materials, over certain ranges of the terminal velocity, the retarding force is proportional to the velocity itself. Such materials are said to obey Ohm's law and are referred to as ohmic materials.

Consider the resistor in the simple circuit we have been dealing with.



If you double the voltage across the resistor (by using a seat of EMF that maintains twice the potential difference between its terminals as the original seat of EMF) then you double the electric field in the resistor. This doubles the force exerted on each charge carrier. This means that, at the terminal velocity of any charge carrier, the retarding force has to be twice as great. (Since, upon making that final circuit connection, the velocity of the charge carriers increases until the retarding force on each charge carrier is *equal* in magnitude to the applied force.) In an ohmic material, if the retarding force is twice as great, then the velocity is twice as great. If the velocity is twice as great, then the charge flow rate, the electric current, is twice as great. So, doubling the voltage across the resistor doubles the current. Indeed, for a resistor that obeys Ohm's Law, the current in a resistor is directly proportional to the voltage across the resistor.

Summarizing: When you put a voltage across a resistor, there is a current in that resistor. The ratio of the voltage to the current is called the resistance of the resistor.

$$R = \frac{V}{I}$$

This definition of resistance is consistent with our understanding that the resistance of a resistor is a measure of how lousy a conductor it is. Check it out. If, for a given voltage across the resistor, you get a tiny little current (meaning the resistor is a very poor conductor), the value of resistance $R = \frac{V}{I}$ with that small value of current in the denominator, is very big. If, on the

other hand, for the same voltage, you get a big current (meaning the resistor is a good conductor), then the value of resistance $R = \frac{V}{I}$ is small.

If the material of which the resistor is made obeys Ohm's Law, then the resistance R is a constant, meaning that its value is the same for different voltages. The relation $R = \frac{V}{I}$ is typically written in the form $V = IR$.

> Ohm's Law: The resistance R, in the expression $V = IR$, is a constant.
>
> Ohm's Law is good for resistors made of certain materials (called ohmic materials) over a limited range of voltages.

Units of Resistance

Given that the resistance of a resistor is defined as the ratio of the voltage across that resistor to the resulting current in that resistor,

$$R = \frac{V}{I}$$

it is evident that the unit of resistance is the volt per ampere, $\frac{\text{V}}{\text{A}}$. This combination unit is given a name. We call it the ohm, abbreviated Ω, the Greek letter upper-case omega.

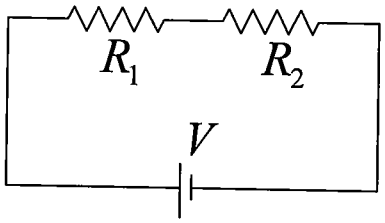
$$1\,\Omega = 1\,\frac{\text{volt}}{\text{ampere}}$$

10 Resistors in Series and Parallel; Measuring *I* & *V*

The analysis of a circuit involves the determination of the voltage across, and the current through, circuit elements in that circuit. A method that I call "the method of ever simpler circuits" can be used to simplify the analysis of many circuits that have more than one resistor. The method involves the replacement of a combination of resistors with a single resistor, carefully chosen so that the replacement does not change the voltage across, nor the current through, the other circuit elements in the circuit. The resulting circuit is easier to analyze, and, the results of its analysis apply to the original circuit. Because the single carefully-chosen resistor has the same effect on the rest of the circuit as the original combination of resistors, we call the single resistor the equivalent resistance of the combination, or, simply, the equivalent resistor.

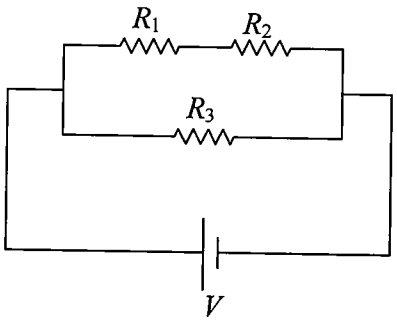
Resistors in Series

One combination of resistors that can be replaced with a single effective resistor is a series combination of resistors. Two two-terminal circuit elements in a circuit are in series with each other when one end of one is connected with one end of the other with nothing else connected to the connection[1]. For instance, R_1 and R_2 in the following circuit are in series with each other.



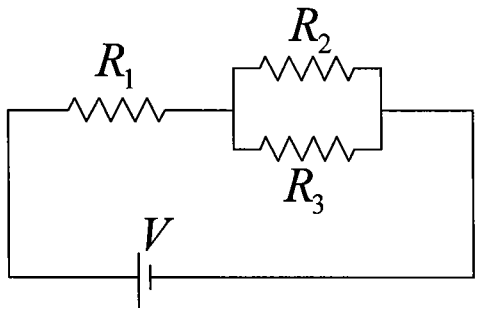
From our viewpoint, the right end of R_1 is connected to the left end of R_2 and nothing else is connected to the point in the circuit where they are connected.

R_1 and R_2 in the following circuit are also in series with each other:



[1] Here we have described adjacent resistors that are in series. Non-adjacent two-terminal circuit elements are also in series with each other if each is in series with a third two-terminal circuit element. In this definition, in addition to an ordinary two-terminal circuit element such as a seat of EMF or a resistor, a two-terminal combination of circuit elements is considered to be a two-terminal circuit element.

But, R_1 and R_2 in the following circuit are *not* in series with each other:



While it is true that the right end of R_1 is connected to the left end of R_2, it is not true that "nothing else is connected to the connection." Indeed, the left end of R_3 is connected to the point in the circuit at which R_1 and R_2 are connected to each other.

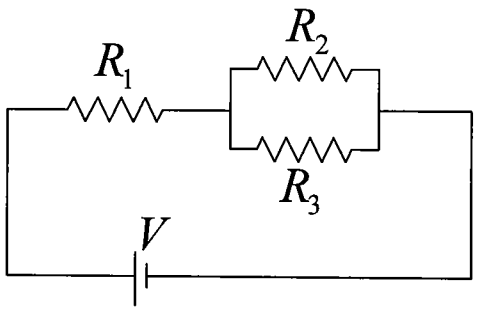
In implementing the method of ever simpler circuits, the plan is to replace a combination of resistors that are in series with each other with a single, well-chosen *equivalent* resistor. The question is, what value must the resistance of the single resistor be in order for it to be equivalent to the set of series resistors it replaces? For now, we simply give you the result. The derivation will be provided in the next chapter.

The equivalent resistance of resistors in series is simply the sum of the resistances.

$$R_S = R_1 + R_2 + R_3 + \ldots$$

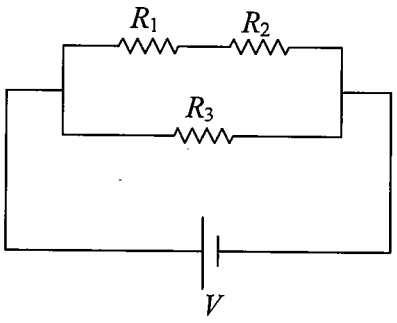
Resistors in Parallel

Circuit elements are in parallel with each other if they are connected together (by nothing but "perfect" conductor) at both ends. So, for instance, R_2 and R_3 in the following circuit:



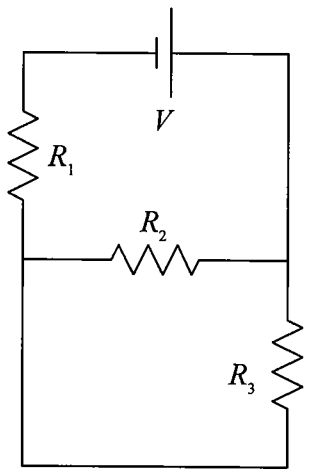
are in parallel with each other.

On the other hand, R_1 and R_3 in the following circuit

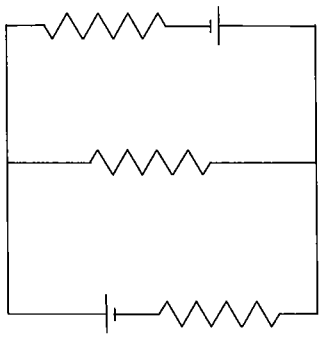


are *not* in parallel with each other.

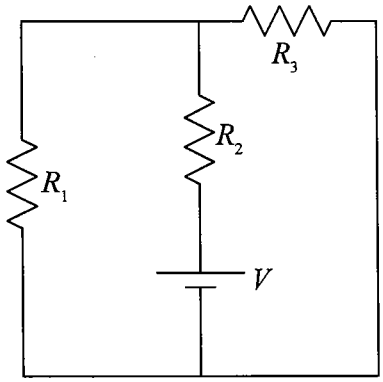
Resistors R_2 and R_3 in the following circuit are in parallel with each other:



But, none of the resistors in the following circuit are in parallel with each other:



whereas R_1 and R_3 in the following circuit are in parallel with each other:



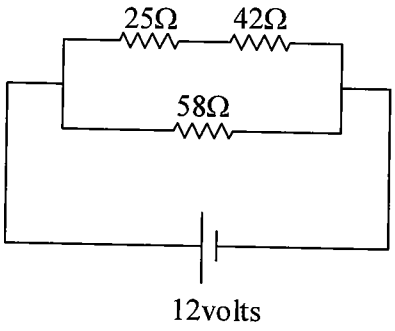
So what is the equivalent resistor for resistors in parallel? Here we provide the result. We save the derivation for the next chapter.

The equivalent resistance of resistors in parallel is the reciprocal of the sum of the reciprocals of the resistances of the resistors making up the parallel combination:

$$R_P = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots}$$

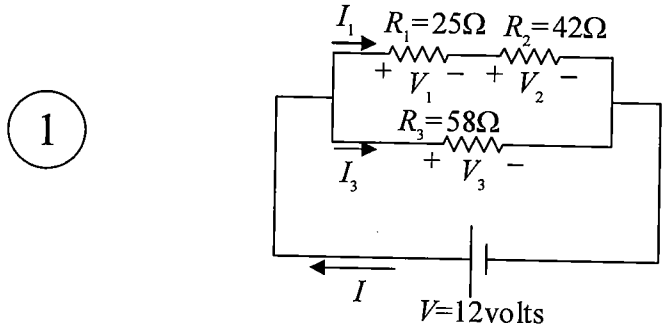
Example

Find the voltage across, and the current through, each of the circuit elements in the diagram below.



Solution

First we add some notation to the diagram to define our variables (do not omit this step):

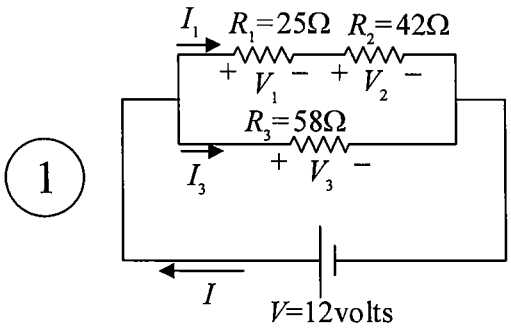


The + and – signs on the resistors (indicating the high potential side and the low potential side of each resistor), are an important part of the definition of the voltages. If you are given values, and the value you calculate for V_1 turns out to be positive, e.g. +5.0 volts, then the reader of your solution knows that the potential of the left end of R_1 is 5.0 volts higher than that of the right end. But, if the value that you calculate for V_1 is negative, e.g. –5.0 volts, then the reader knows that the potential of the left end of R_1 is 5.0 volts *lower* than that of the right end.

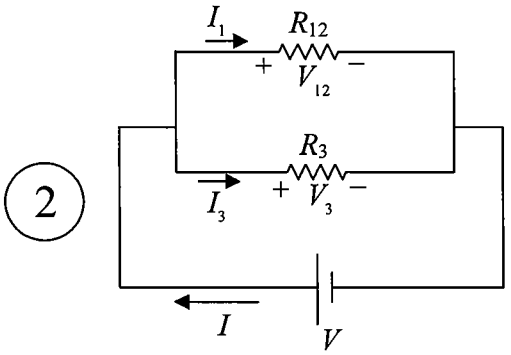
The "+" and "–" labels on the resistors must be consistent with the current direction. In fact, one first draws and labels the current arrows, and then puts the "+" on the end of the resister that the current enters (and the "–" on the other end).

Next we draw a sequence of circuits. Each new diagram includes an equivalent resistor in place of *one* series combination or *one* parallel combination. (Do not omit any diagrams, and, do not replace anything more than a single series combination or a single parallel combination of resistors in any one step.) As you draw each circuit, calculate the value of the equivalent resistance.

First, we copy the diagram from the preceding page.



Next, we replace the series combination of R_1 and R_2 with the equivalent resistor R_{12}.

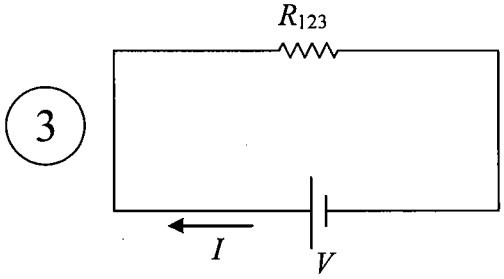


$$R_{12} = R_1 + R_2$$

$$R_{12} = 25\Omega + 42\Omega$$

$$\boxed{R_{12} = 67\Omega}$$

Finally, we replace the parallel combination of R_{12} and R_3 with the equivalent resistor R_{123}.

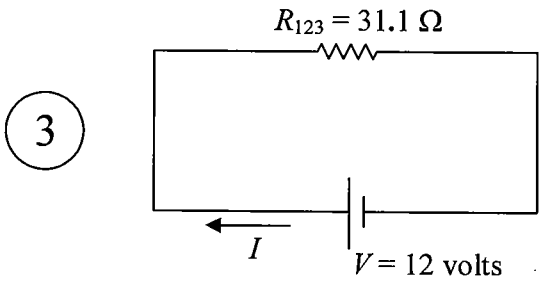


$$R_{123} = \frac{1}{\frac{1}{R_{12}} + \frac{1}{R_3}}$$

$$R_{123} = \frac{1}{\frac{1}{67\Omega} + \frac{1}{58\Omega}}$$

$$\boxed{R_{123} = 31.1\Omega}$$

Now we analyze the simplest circuit, the one I have labeled "3" above.



One of the most common mistakes that folks make in analyzing circuits is using any old voltage in $V=IR$. You have to use the voltage across the resistor. In analyzing circuit 3, however, we can use the one voltage in the diagram because the voltage across the seat of EMF *is* the voltage across the resistor. The terminals of the resistor are connected to the same two conductors that the terminals of the seat of EMF are connected to. Thus,

$$V = IR_{123}$$

$$I = \frac{V}{R_{123}}$$

$$I = \frac{12\ \text{volts}}{31.1\ \Omega}$$

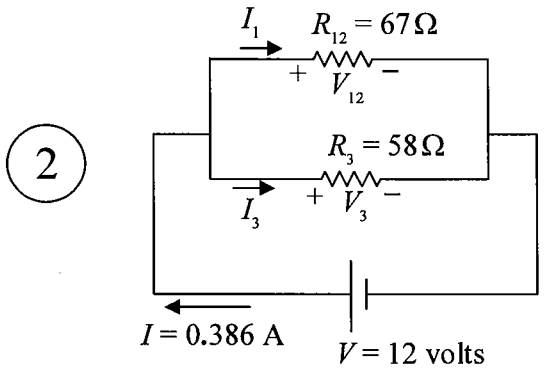
$$I = 0.386\ \text{A}$$

At this point, we've got two of the answers. The voltage across the seat of EMF was asked for, but it is also given, so we don't have to show any work for it. And now we have the current through the seat of EMF.

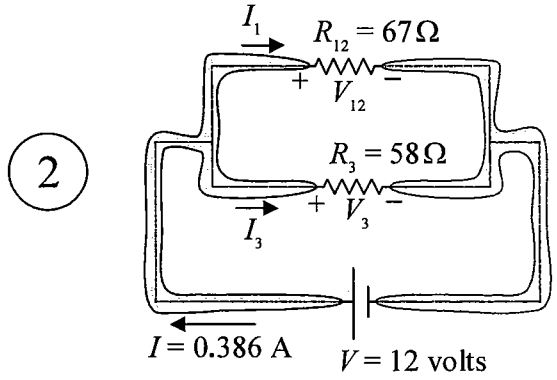
$V = 12$ volts
$I = 0.39$ amperes

Note that the arrow labeled I in our diagram is part of our answer. It tells the reader what I means, including the direction of charge flow for a positive value of I.

Our next step is to take the information that we have learned here, to our next more complicated circuit. That would be the one I labeled "2" above.



There are only two wires (conductors) in this circuit. I am going to highlight them in order to make my next point:



Highlighting the conductors makes it obvious that the voltage across R_{12} is the same as the voltage across the seat of EMF because, in both cases, the voltage is the potential difference between one and the same pair of conductors. Likewise, the voltage across R_3 is the same as the voltage across the seat of EMF. Hence, we have,

$$V_{12} = V$$

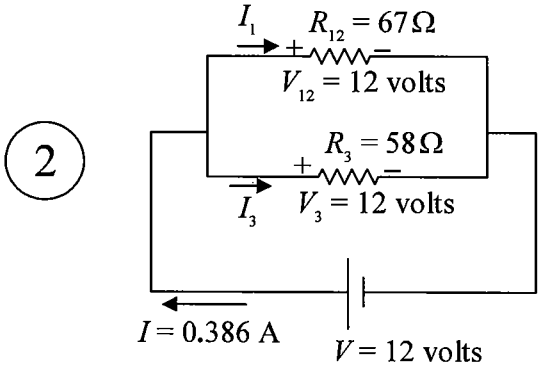
$$V_{12} = 12 \text{ volts}$$

and,

$$V_3 = V$$

$$V_3 = 12 \text{ volts.}$$

The last value is one of our answers. We were supposed to find V_3. Now that we know the voltage across R_3, we can use it in $V=IR$ to get I_3.



For resistor R_3, we have:

$$V_3 = I_3 R_3$$

$$I_3 = \frac{V_3}{R_3}$$

$$I_3 = \frac{12 \text{ volts}}{58\ \Omega}$$

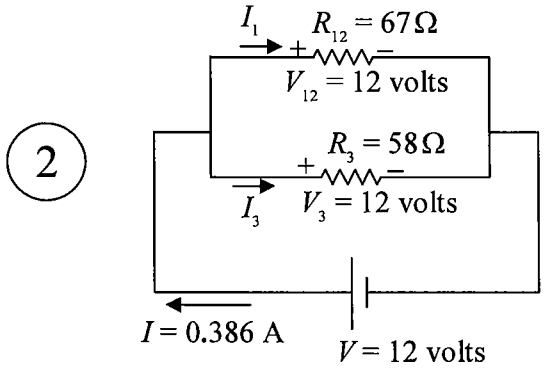
$$I_3 = \frac{12 \text{ volts}}{58\ \Omega}$$

$$I_3 = 0.207 \text{ A}$$

The voltage and current through resistor R_3 are answers to the problem:

$V_3 = 12$ volts
$I_3 = 0.21$ amperes

Now let's get the current through R_{12}. I've labeled that current I_1 in diagram 2.



For resistor R_{12}, we have:

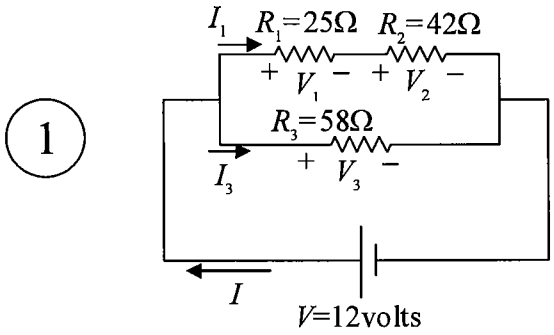
$$V_{12} = I_1 R_{12}$$

$$I_1 = \frac{V_{12}}{R_{12}}$$

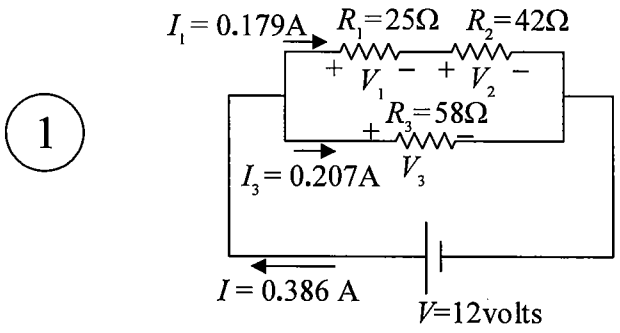
$$I_1 = \frac{12 \text{ volts}}{67\ \Omega}$$

$$I_1 = 0.179 \text{ A}$$

Now it is time to take what we have learned here up to the next more complicated circuit (which is the original circuit).



I copy that here with the values of the current included:



It is clear from this diagram that the current I_1 that we just found (the current through R_{12}) is the current through R_1, and, it is the current through R_2.

$$I_2 = I_1$$

$$I_2 = 0.179\text{A}$$

These are answers to the problem.

With the current through R_1 known, we can now solve for V_1:

$$V_1 = I_1 R_1$$

$$V_1 = 0.179\text{ A}\ (25\Omega)$$

$$V_1 = 4.5\text{ volts}$$

Thus, our answers for resistor R_1 are:

$V_1 = 4.5$ volts
$I_1 = 0.18$ amperes

And, with the current through R_2 known, we can solve for V_2:

$$V_2 = I_2 R_2$$

$$V_2 = 0.179\text{ A}\ (42\Omega)$$

$$V_2 = 7.5\text{ volts}$$

Thus, our answers for resistor R_2 are:

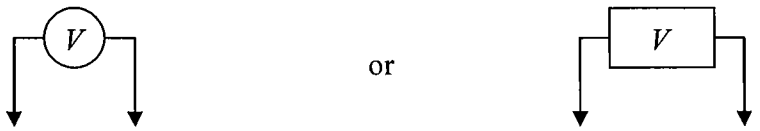
$V_2 = 7.5$ volts
$I_2 = 0.18$ amperes

How to Connect a Voltmeter in a Circuit

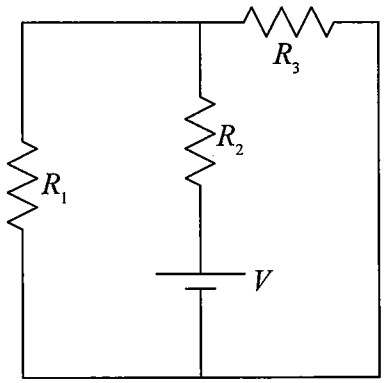
As discussed earlier in this book, a voltmeter is a device used for measuring the potential difference between two different points in space. In a circuit, we use it to measure the potential difference between two conductors (wires) in the circuit. When you do that, the voltmeter becomes a two-terminal circuit element of the circuit. The ideal voltmeter, as a circuit element, can be considered to be a resistor with infinite resistance. As such, it has no effect on the circuit. This is good. We don't want the measuring device to change the value of that which you are trying to measure.

A voltmeter consists of a box with two wires coming out of it. Typically, each wire ends in a metal-tipped wand (called a probe) or some kind of metal clip. The box has a gauge on it which displays the potential difference between the two wires. Touch the tip of one wire to one point in the circuit and the tip of the other wire to another point in the circuit (being sure to establish good metal-to-metal contact at both points) and the voltmeter will display the potential difference (the voltage) between those two points in the circuit.

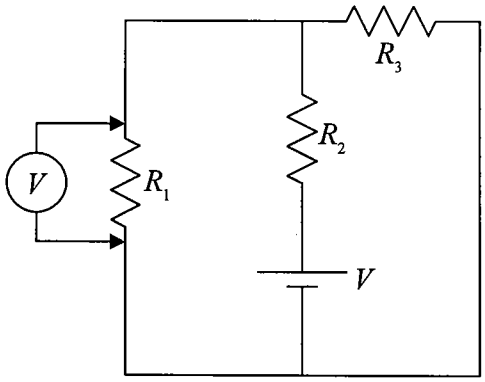
A typical manner of depicting a voltmeter in a circuit is to draw it as



To connect a voltmeter to measure the voltage across R_1 in the following circuit:



hook it up as indicated in the following diagram.

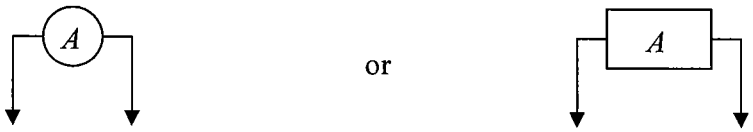


As far as its role as a circuit element (a side effect), the ideal voltmeter has as much effect on the circuit it is used on, as the air around the circuit has.

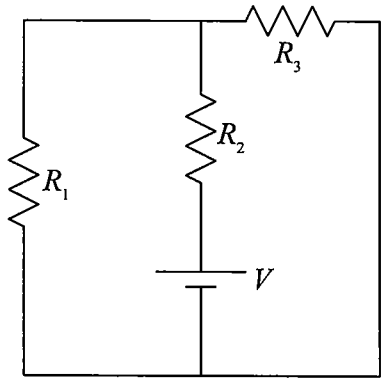
How to Connect an Ammeter in a Circuit

The ammeter, a device used to measure current, is a totally different beast. The ideal ammeter acts like a perfectly-conducting piece of wire that monitors the charge flow through itself. Connecting it in a circuit as you would a voltmeter (don't do it!) will drastically change the circuit (and could cause damage to the meter).

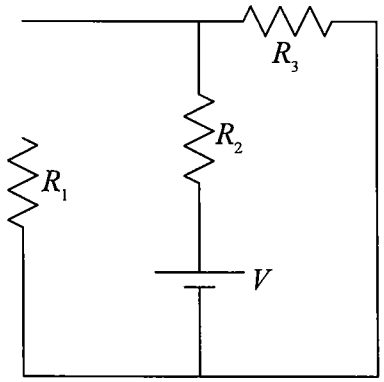
A typical manner of depicting an ammeter in a circuit is to draw it as



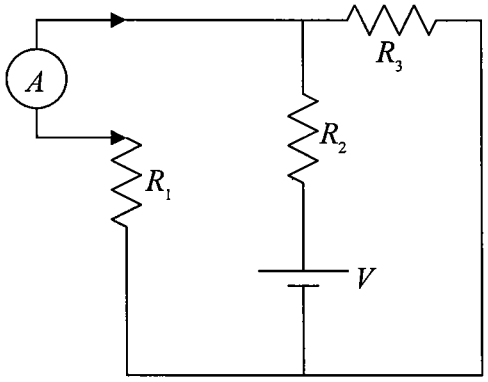
To connect an ammeter to measure the current in R_1 in the following circuit:



You have to first break the circuit,



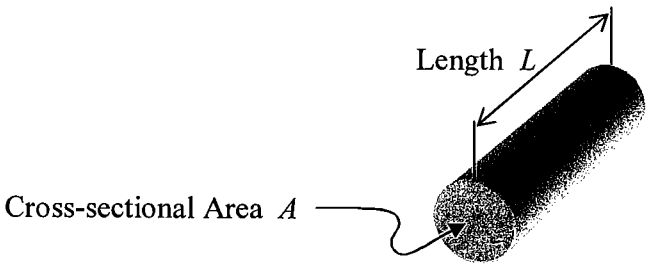
and then connect the ammeter in series with the circuit element whose current you wish to measure.



Remember, to measure current with an ammeter, *some disassembly is required*!

11 Resistivity, Power

In chapter 9 we discussed resistors that conform to Ohm's Law. From the discussion, one could deduce that the resistance of such a resistor depends on the nature of the material of which the resistor is made and on the size and shape of the resistor. In fact, for resistors made out of a single kind of material, in the shape of a wire[1] with a terminal at each end,



the resistance is given by:

$$R = \rho \frac{L}{A} \tag{11-1}$$

where:

- R is the resistance of the resistor as measured between the ends,
- ρ is the resistivity of the substance of which the resistor is made,
- A is the cross-sectional area of the wire-shaped resistor, and
- L is the length of the resistor.

The values of resistivity for several common materials are provided in the following table:

Material	Resistivity ρ
Silver	1.6×10^{-8} Ω·m
Copper	1.7×10^{-8} Ω·m
Gold	2.4×10^{-8} Ω·m
Aluminum	3×10^{-8} Ω·m
Tungsten	5.6×10^{-8} Ω·m
Nichrome	1.0×10^{-6} Ω·m
Seawater	0.25 Ω·m
Rubber	1×10^{13} Ω·m
Glass	1×10^{10} to 1×10^{14} Ω·m
Quartz	5×10^{15} to 7.5×10^{17} Ω·m

[1] The resistor can have any shape such that one linear dimension can be identified as the length of the resistor, and, such that the intersection of a plane perpendicular to the length of the resistor, at any position along the length of the resistor, has one and the same area (the cross-sectional area of the resistor). I am calling the shape "the shape of a wire" for ease in identification of what we mean by the "along-the-length" dimension.

In the expression $R = \rho\frac{L}{A}$, the resistivity ρ depends on the charge carrier[2] density, that is, the number-of-charge-carriers-per-volume. The more charge carriers per volume, the smaller the resistance since, for a given velocity of the charge carriers, more of them will be passing any point along the length of the resistor every second for a given voltage across the resistor. The resistivity also depends on the retarding force factor. We said that the retarding force on each charge carrier is proportional to the velocity of that charge carrier.

Retarding Force = – (factor) times (charge carrier velocity)

(The minus sign is there because the retarding force is in the direction opposite that of the charge-carrier velocity.) The bigger the retarding force factor, the greater the resistivity of the material for which the factor applies.

The charge carrier density and the retarding force factor determine the value of ρ. The effect of ρ on the resistance is evident in the expression $R = \rho\frac{L}{A}$. The bigger ρ is, the greater the resistance is.

Why the factor of L in $R = \rho\frac{L}{A}$? It's saying that the greater the length of the single-substance resistor in the shape of a wire, the greater the resistance of the resistor, all other things being equal (same substance, same cross-sectional area). It means, for instance, that if you have two resistors, identical in all respects except that one is twice as long as the other, and you put the same voltage across each of the resistors, you'll get half the current in the longer resistor. Why is that?

To get at the answer, we need to consider the electric field inside the wire-shaped resistor when we have a voltage V across the resistor. The thing is, the electric field inside the resistor is directed along the length of the resistor, and, it has the same magnitude everywhere along the length of the resistor. Evidence for this can be obtained by means of simple voltage measurements. Use a voltmeter to measure the potential difference $\Delta\varphi$ between two points on the resistor that are separated by a certain distance Δx, say 2 mm (measured along the length of the resistor) for instance. It turns out that no matter where along the length you pick the pair of points (separated from each other by the Δx), you always get the same voltage reading. Imagine (this part is a thought experiment) moving a positive test charge q_T that distance Δx along the resistor from high potential toward low potential. No matter where along the length of the resistor you do that, the work done (by the electric field characterized by the potential) $q_T\Delta\varphi$ (calculated as the negative of the change of the potential energy of the test charge) is the same. The work, calculated as force times distance, is $q_T E\Delta x$. For that to be the same at every point along the length of the resistor, the electric field E has to have the same value everywhere along the length of the resistor. Furthermore, setting the two expressions for the work equal to each other yields:

[2] A charge carrier is a particle that has charge and is free to move about within the material of which it is a part.

$$q_T E \Delta x = q_T \Delta \varphi$$
$$E = \frac{\Delta \varphi}{\Delta x}$$

E being constant thus means that $\frac{\Delta \varphi}{\Delta x}$ is constant which means that a graph of φ vs. x is a straight line with slope $\frac{\Delta V}{\Delta x}$. But, in calculating that slope, since it is a straight line, we don't have to use a tiny little Δx. We can use the entire length of the resistor and the corresponding potential difference, which is the voltage V across the resistor. Thus,

$$E = \frac{V}{L}$$

where:

- E is the magnitude of the electric field everywhere in the single-substance wire-shaped resistor,
- V is the voltage across the resistor, and
- L is the length of the resistor.

This result ($E = \frac{V}{L}$) is profound in and of itself, but, if you recall, we were working on answering the question about why the resistance R, of a single-substance wire-shaped resistor, is proportional to the length of the resistor. We are almost there. The resistance is the ratio of the voltage across the resistor to the current in it. According to $E = \frac{V}{L}$, the longer the resistor, the weaker the electric field in the resistor is for a given voltage across it. A weaker E results in a smaller terminal velocity for the charge carriers in the resistor, which results in a smaller current. Thus, the longer the resistor, the smaller the current is; and; the smaller the current, the greater the voltage-to-current ratio is; meaning, the greater the resistance.

The next resistance-affecting characteristic in $R = \rho \frac{L}{A}$ that I want to discuss is the area A. Why should that affect the resistance the way it does? Its presence in the denominator means that the bigger the cross-sectional area of the wire-shaped resistor, the *smaller* the resistance. Why is that?

If we compare two different resistors made of the same material and having the same length (but different cross-sectional areas) both having the same voltage across them, they will have the same electric field $E = \frac{V}{L}$ in them. As a result, the charge carriers will have the same velocity v. In an amount of time Δt,

$$L = v \Delta t$$
$$\Delta t = \frac{L}{v}$$

all the free-to-move charge carriers in either resistor will flow out the lower potential end of the resistor (while the same amount of charge flows in the higher potential end). This time Δt is the same for the two different resistors because both resistors have the same length, and the charge carriers in them have the same v. The number of charge carriers in either resistor is proportional to the volume of the resistor. Since the volume is given by volume = LA, the number of charge carriers in either resistor is proportional to the cross-sectional area A of the resistor. Since the number of charge carriers in either resistor, divided by the time Δt is the current in that resistor, this means that the current is proportional to the area.

If the current is proportional to the area, then the resistance, being the ratio of the voltage to the current, must be inversely proportional to the area. And so ends our explanation regarding the presence of the A in the denominator in the expression

$$R = \rho \frac{L}{A}$$

Power

You were introduced to power in Volume I of this book. It is the rate at which work is done. It is the rate at which energy is transferred. And, it is the rate at which energy is transformed from one form of energy into another form of energy. The unit of power is the watt, W.

$$1\,\mathrm{W} = 1\,\frac{\mathrm{J}}{\mathrm{s}}$$

In a case in which the power is the rate that energy is transformed from one form to another, the amount of energy that is transformed from time *0* to time t:

- if the power is constant, is simply the power times the duration of the time interval:

$$\text{Energy} = Pt$$

- if the power is a function of time, letting t' be the time variable that changes from 0 to t, is:

$$\text{Energy} = \int_0^t P(t')dt'$$

The Power of a Resistor

In a resistor across which there is a voltage V, energy is transformed from electric potential energy into thermal energy. A particle of charge q, passing through the resistor, loses an amount of potential energy qV but it does not gain any kinetic energy. As it passes through the resistor, the electric field in the resistor does an amount of work qV on the charged particle, but, at a same time, the retarding force exerted on the charged particle by the background material of the resistor, does the negative of that same amount of work. The retarding force, like friction, is a non-conservative force. It is exerted on the charge carrier when the charge carrier collides with impurities and ions (especially at sites of defects and imperfections in the structure of the material). During those collisions, the charge carriers impart energy to the ions with which they collide. This gives the ions vibrational energy which manifests itself, on a macroscopic scale, (early in the process) as an increase in temperature. Some of the thermal energy is continually transferred to the surroundings. Under steady state conditions, arrived at after the resistor has warmed up, thermal energy is transferred to the surroundings at the same rate that it is being transformed from electrical potential energy in the resistor.

The rate at which electric potential energy is converted to thermal energy in the resistor is the power of the resistor (a.k.a. the power dissipated[3] by the resistor). It is the rate at which the energy is being delivered to the resistor. The energy conversion that occurs in the resistor is sometimes referred to as the dissipation of energy. One says that the resistor power is the rate at which energy is dissipated in the resistor. It's pretty easy to arrive at an expression for the power of a resistor in terms of circuit quantities. Each time a coulomb of charge passes through a resistor that has a voltage V across it, an amount of energy equal to one coulomb times V is converted to thermal energy. The current I is the number of coulombs-per-second passing through the resistor. Hence V times I is the number of joules-per-second converted to thermal energy. That's the power of the resistor. In short,

$$P = IV$$

where:

- P is the power of the resistor. It is the rate at which the resistor is converting electrical potential energy into thermal energy. The unit of power is the watt. $1\ \text{W} = 1\frac{\text{J}}{\text{s}}$.
- I is the current in the resistor. It is the rate at which charge is flowing through the resistor. The unit of current is the ampere. $1\ \text{A} = 1\frac{\text{C}}{\text{s}}$.
- V is the voltage across the resistor. It is the amount by which the value of electric potential (the electric potential energy per charge) at one terminal of the resistor exceeds that at the other terminal. The unit of voltage is the volt. $1\ \text{volt} = 1\ \frac{\text{J}}{\text{C}}$.

[3] To be dissipated means to be dispersed or broken up and sent in all different directions.

The Power of a Seat of EMF

In a typical circuit, a seat of EMF causes positive charge carriers (in our positive-charge-carrier model) to go from a lower-potential conductor, through itself, to a higher-potential conductor. The electric field of the conductors exerts a force on the charge carriers inside the seat of EMF in the direction opposite to the direction in which the charge carriers are going. The charged particles gain electric potential energy in moving from the lower-potential terminal of the seat of EMF to the higher-potential terminal. Where does that energy come from?

In the case of a battery, the energy comes from chemical potential energy stored in the battery and released in chemical reactions that occur as the battery moves charge from one terminal to the other. In the case of a power supply, the power supply, when plugged into a wall outlet and turned on, becomes part of a huge circuit including transmission wires extending all the way back to a power plant. At the power plant, depending on the kind of power plant, kinetic energy of moving water, or thermal energy used to make steam to turn turbines, or chemical potential energy stored in wood, coal, or oil; is converted to electric potential energy. Whether it is part of a battery, or a part of a power supply, the seat of EMF converts energy into electric potential energy. It keeps one of its terminals at a potential $\mathcal{E}$ higher than the other terminal. Each time it moves a coulomb of charge from the lower potential terminal to the higher potential terminal, it increases the potential energy of that charge by one coulomb times $\mathcal{E}$. Since the current I is the number of coulombs per second that the seat of EMF moves from one terminal to the other, the power, the rate at which the seat of EMF delivers energy to the circuit, is given by:

$$P = I\mathcal{E}$$

Recall that it is common to use the symbol V (as well as $\mathcal{E}$) to represent the voltage across a seat of EMF. If you use V, then the power of the seat of EMF is given by:

$$P = IV$$

where:

- P is the rate at which a seat of EMF delivers energy to a circuit,
- I is the current in the seat of EMF (the rate at which charge flows through the seat of EMF), and
- V is the voltage across the seat of EMF.

This is the same expression as the expression for the power of a resistor.

12 Kirchhoff's Rules, Terminal Voltage

There are two circuit-analysis laws that are so simple that you may consider them "statements of the obvious" and yet so powerful as to facilitate the analysis of circuits of great complexity. The laws are known as Kirchhoff's Laws. The first one, known both as "Kirchhoff's Voltage Law" and "The Loop Rule" states that, starting on a conductor[1], if you drag the tip of your finger around any loop in the circuit back to the original conductor, the sum of the voltage changes experienced by your fingertip will be zero. (To avoid electrocution, please think of the finger dragging in an actual circuit as a thought experiment.)

Kirchhoff's Voltage Law (a.k.a. the Loop Rule)

To convey the idea behind Kirchhoff's Voltage Law, I provide an analogy. Imagine that you are exploring a six-story mansion that has 20 staircases. Suppose that you start out on the first floor. As you wander around the mansion, you sometimes go up stairs and sometimes go down stairs. Each time you go up stairs, you experience a positive change in your elevation. Each time you go down stairs, you experience a negative change in your elevation. No matter how convoluted the path of your explorations might be, if you again find yourself on the first floor of the mansion, you can rest assured that the algebraic sum of all your elevation changes is zero.

To relate the analogy to a circuit, it is best to view the circuit as a bunch of conductors connected by circuit elements (rather than the other way around as we usually view a circuit). Each conductor in the circuit is at a different value of electric potential (just as each floor in the mansion is at a different value of elevation). You start with your fingertip on a particular conductor in the circuit, analogous to starting on a particular floor of the mansion. The conductor is at a particular potential. You probably don't know the value of that potential any more than you know the elevation that the first floor of the mansion is above sea level. You don't need that information. Now, as you drag your finger around the loop, as long as you stay on the same conductor, your fingertip will stay at the same potential. But, as you drag your fingertip from that conductor, through a circuit element, to the next conductor on your path, the potential of your fingertip will change by an amount equal to the voltage across the circuit element (the potential difference between the two conductors). This is analogous to climbing or descending a flight of stairs and experiencing a change in elevation equal to the elevation difference between the two floors.

If you drag your fingertip around the circuit in a loop, back to the original conductor, your finger is again at the potential of that conductor. As such, the sum of the changes in electric potential experienced by your finger on its traversal of the loop must be zero. This is analogous to stating that if you start on one floor of the mansion, and, after wandering through the mansion, up and

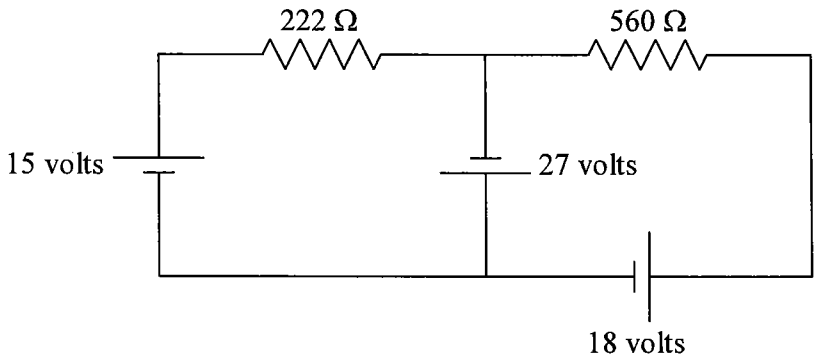
[1] Circuits consist of circuit elements and wires, I am calling the wires "conductors." More specifically, a conductor in a circuit is any wire segment, together will all other wire segments connected directly to the wire segment (with no intervening circuit elements).

down staircases, you end up on the same floor of the mansion, your total elevation change is zero.

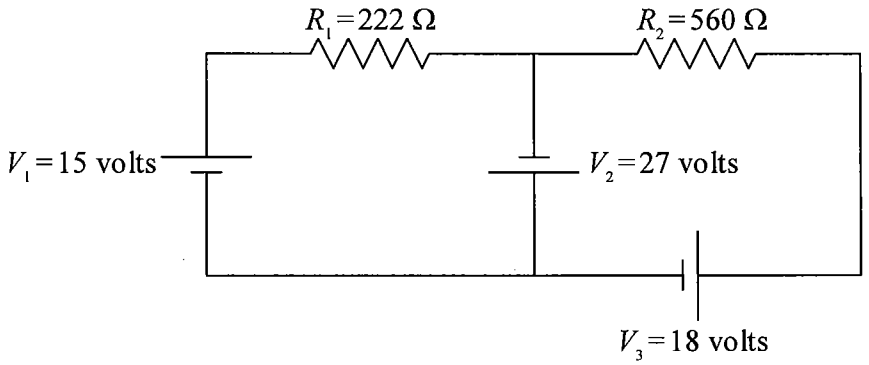
In dragging your finger around a closed loop of a circuit (in any direction you want, regardless of the current direction) and adding each of the voltage changes to a running total, the critical issue is the algebraic sign of each voltage change. In the following example we show the steps that you need to take to get those signs right, and to prove to the reader of your solution that they are correct.

Example

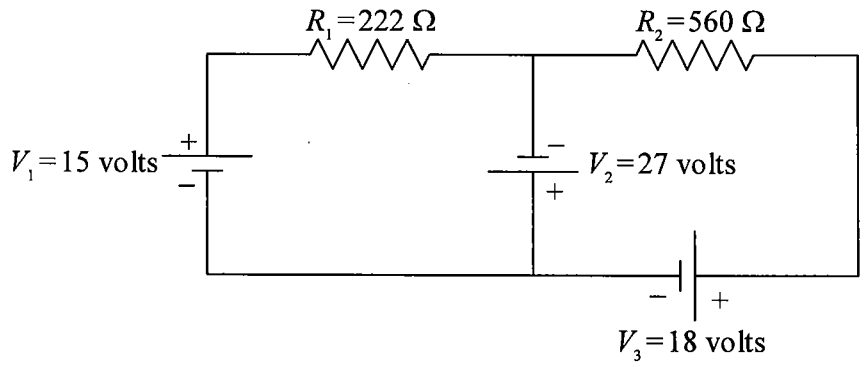
Find the current through each of the resistors in the following circuit.



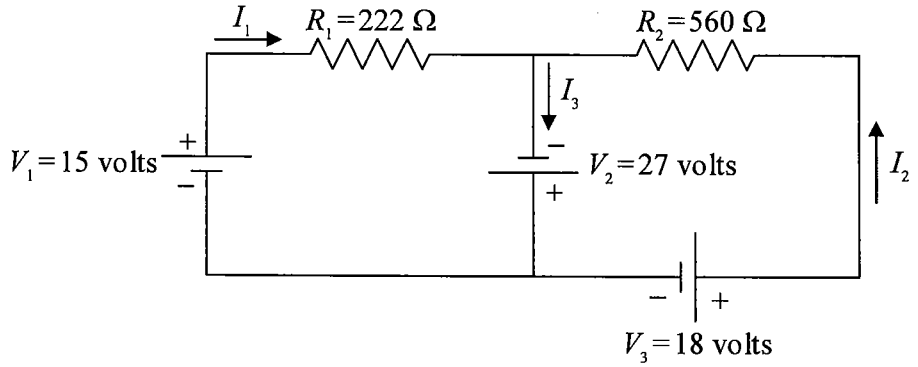
Before we get started, let's define some names for the given quantities:



Each two-terminal circuit element has one terminal that is at a higher potential than the other terminal. The next thing we want to do is to label each higher potential terminal with a "+" and each lower-potential terminal with a "–". We start with the seats of EMF. They are trivial. By definition, the longer parallel line segment, in the symbol used to depict a seat of EMF, is at the higher potential.

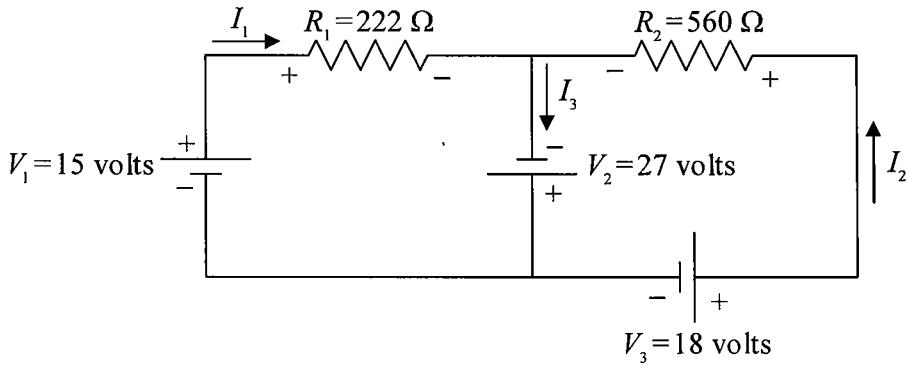


Next we define a current variable for each "leg" of the circuit. A "leg" of the circuit extends from a point in the circuit where three or more wires are joined (called a junction) to the next junction. All the circuit elements in any one leg of the circuit are in series with each other, so, they all have the same current through them.

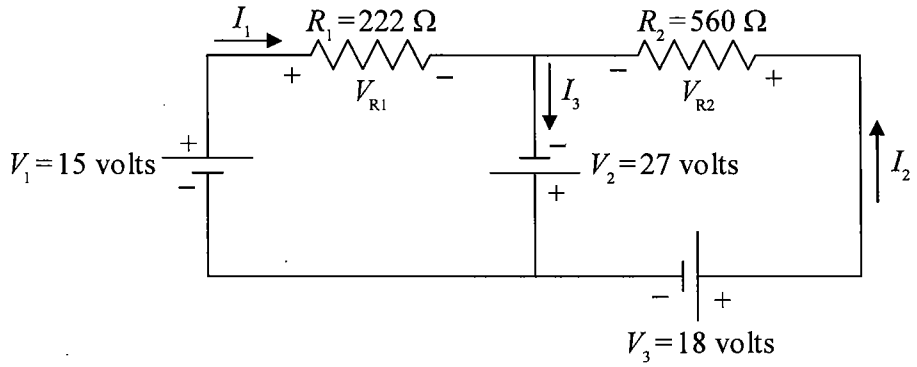


Note: In defining your current variables, the direction in which you draw the arrow in a particular leg of the circuit, is just a guess. Don't spend a lot of time on your guess. It doesn't matter. If the current is actually in the direction opposite that in which your arrow points, you will simply get a negative value for the current variable. The reader of your solution is responsible for looking at your diagram to see how you have defined the current direction and for interpreting the algebraic sign of the current value accordingly.

Now, by definition, the current is the direction in which positive charge carriers are flowing. The charge carriers *lose* electric potential energy when they go through a resistor, so, they go from a higher-potential conductor, to a lower-potential conductor when they go through a resistor. That means that the end of the resistor at which the current enters the resistor is the higher potential terminal (+), and, the end at which the current exits the resistor is the lower-potential terminal (–) of the resistor.

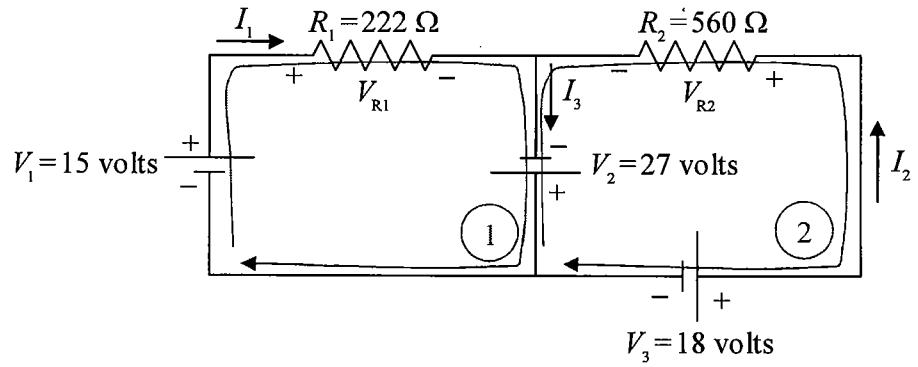


Now let's define some variable names for the resistor voltages:



Note that the + and – signs on the resistors are important parts of our definitions of V_{R1} and V_{R2}. If, for instance, we calculate V_{R1} to have a positive value, then, that means that the left (as we view it) end of V_{R1} is at a higher potential than the right end (as indicated in our diagram). If V_{R1} turns out to be negative, that means that the left end of R_1 is actually at a lower potential than the right end. We do not have to do any more work if V_{R1} turns out to be negative. It is incumbent upon the reader of our solution to look at our circuit diagram to see what the algebraic sign of our value for V_{R1} means.

With all the circuit-element terminals labeled "+" for "higher potential" or "–" for "lower potential," we are now ready to apply the Loop Rule. I'm going to draw two loops with arrowheads. The loop that one draws is not supposed to be a vague indicator of direction but a specific statement that says, "Start at this point in the circuit. Go around this loop in this direction, and, end at this point in the circuit." Also, the starting point and the ending point should be the same. In particular, they must be on the same conductor. (Never start the loop on a circuit element.) In the following diagram are the two loops, one labeled ① and the other labeled ②.



Now we write KVL ① to tell the reader that we are applying the Loop Rule (Kirchhoff's Voltage Law) using loop ①, and transcribe the loop equation from the circuit diagram:

KVL ①

$$+V_1 - V_{R1} + V_2 = 0$$

The equation is obtained by dragging your fingertip around the exact loop indicated and recording the voltage changes experienced by your fingertip, and then, remembering to write "= 0." Starting at the point on the circuit closest to the tail of the loop 1 arrow, as we drag our finger around the loop, we first traverse the seat of EMF, V_1. In traversing V_1 we go from lower potential (–) to higher potential (+). That means that the finger experiences a positive change in potential, hence, V_1 enters the equation with a positive sign. Next we come to resistor R_1. In traversing R_1 we go from higher potential (+) to lower potential (–). That's a negative change in potential. Hence, V_{R1} enters our loop equation with a negative sign. As we continue our way about the loop we come to the seat of EMF V_2 and go from lower potential (–) to higher potential (+) as we traverse it. Thus, V_2 enters the loop equation with a positive sign. Finally, we arrive back at the starting point. That means that it is time to write "= 0."

We transcribe the second loop equation in the same fashion:

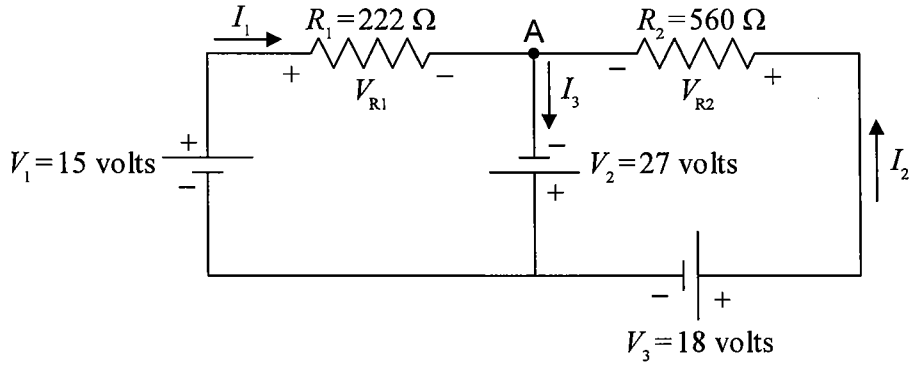
KVL ②

$$-V_2 + V_{R2} - V_3 = 0$$

With these two equations in hand, and knowing that $V_{R1} = I_1 R_1$ and $V_{R2} = I_2 R_2$, the solution to the example problem is straightforward. (We leave it as an exercise for the reader.) It is now time to move on to Kirchhoff's other law.

Kirchhoff's Current Law (a.k.a. the Junction Rule)

Kirchhoff's junction rule is a simple statement of the fact that charge does not pile up at a junction. (Recall that a junction is a point in a circuit where three or more wires are joined together.) I'm going to state it two ways and ask you to pick the one you prefer and use that one. One way of stating it is to say that the net current into a junction is zero. Check out the circuit from the example problem:



In this copy of the diagram of that circuit, I put a dot at the junction at which I wish to apply Kirchhoff's Current Law, and, I labeled that junction "A."

Note that there are three legs of the circuit attached to junction A. In one of them, current I_1 flows toward the junction. In another, current I_2 flows toward the junction. In the third leg, current I_3 flows away from the junction. A current away from the junction counts as the negative of that value of current, toward the junction. So, applying Kirchhoff's Current Law in the form, "The net current into any junction is zero," to junction A yields:

KCL A

$$I_1 + I_2 - I_3 = 0$$

Note the negative sign in front of I_3. A current of $-I_3$ into junction A is the same thing as a current of I_3 out of that junction, which is exactly what we have.

The other way of stating Kirchhoff's Current Law is, "The current into a junction is equal to the current out of that junction." In this form, in applying Kirchhoff's Current Law to junction A in the circuit above, one would write:

KCL A

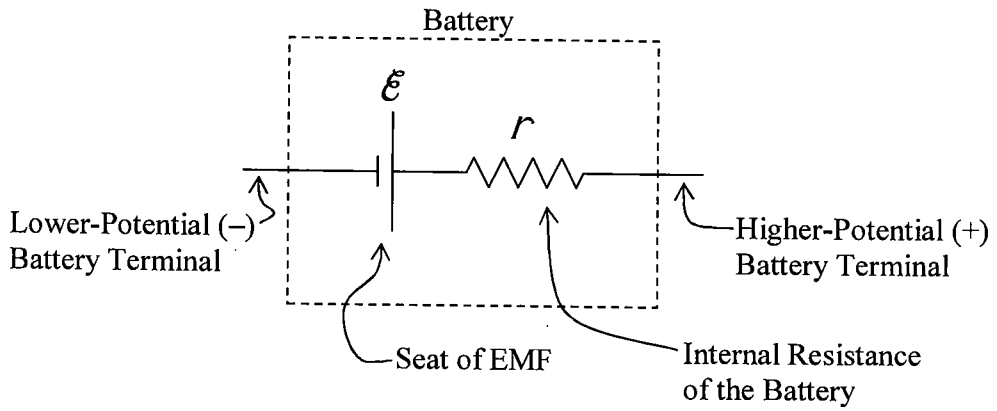
$$I_1 + I_2 = I_3$$

Obviously, the two results are the same.

Terminal Voltage – A More Realistic Model for a Battery or DC Electrical Power Source

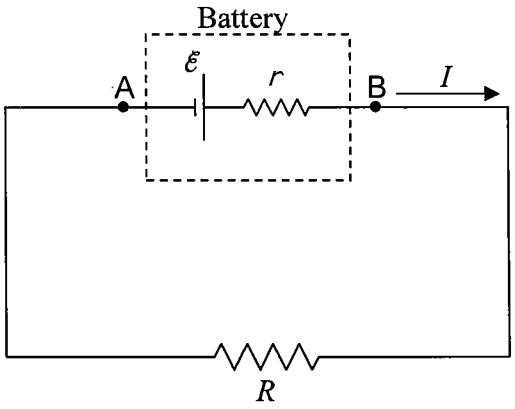
Our model for a battery up to this point has been a seat of EMF. I said that a seat of EMF can be considered to be an ideal battery. This model for a battery is good as long as the battery is fairly new and unused and the current through it is small. Small compared to what? How small? Well, small enough so that the voltage across the battery when it is in the circuit is about the same as it is when it is not in any circuit. How close to being the same? That depends on how accurate you want your results to be. The voltage across a battery decreases when you connect the battery in a circuit. If it decreases by five percent and you calculate values based on the voltage across the battery when it is in no circuit, your results will probably be about 5% off.

A better model for a battery is an ideal seat of EMF in series with a resistor. A battery behaves very much as if it consisted of a seat of EMF in series with a resistor, but, you can never separate the seat of EMF from the resistor, and if you open up a battery you will never find a resistor in there. Think of a battery as a black box containing a seat of EMF and a resistor. The resistor in this model is called the internal resistance of the battery.



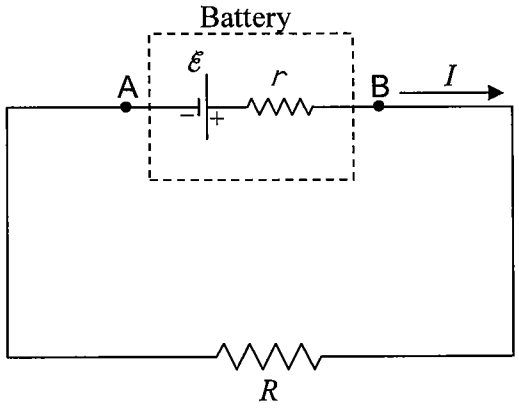
The point at which the seat of EMF is connected to the internal resistance of the battery is inaccessible. The potential difference between the terminals of the battery is called the terminal voltage of the battery. When the battery is not part of a circuit, the terminal voltage is equal to the EMF. You can deduce this from the fact that when the battery is not part of a circuit, there can be no current through the resistor. If there is no current through the resistor than the two terminals of the resistor must be at one and the same value of electric potential. Thus, in the diagram above, the right end of the resistor is at the same potential as the high-potential terminal of the seat of EMF.

Now, let's put the battery in a circuit:

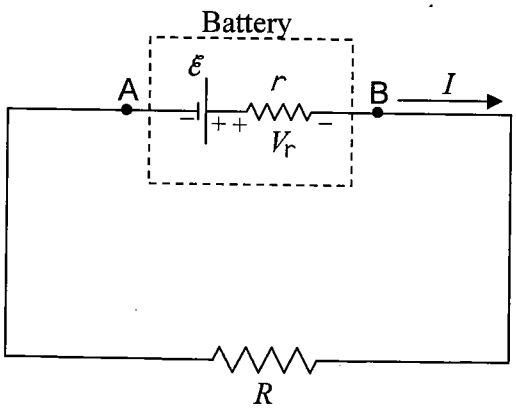


I've indicated the two points A and B on the circuit for communication purposes. The terminal voltage is the voltage from A to B (V_{AB}). If you trace the circuit, with your fingertip, from A to B, the terminal voltage (how much higher the potential is at B than it is at A) is just the sum of the voltage changes your finger experiences along the path. (Note that this time, we are *not* going all the way around a loop. We do *not* wind up on the same conductor upon which we started. So, the sum of the voltage changes from A to B is *not* zero.) To sum the voltage changes from A to B, I will mark the terminals of the components between A and B with "+" for higher potential and "–" for lower potential.

First the seat of EMF: That's trivial. The shorter side of the EMF symbol is the lower potential (–) side and the longer side is the higher potential (+) side.



Now, for the internal resistance of the battery: The end of the internal resistance r that the current enters is the higher-potential (+) end, and, the end that it exits is the lower-potential (–) end.

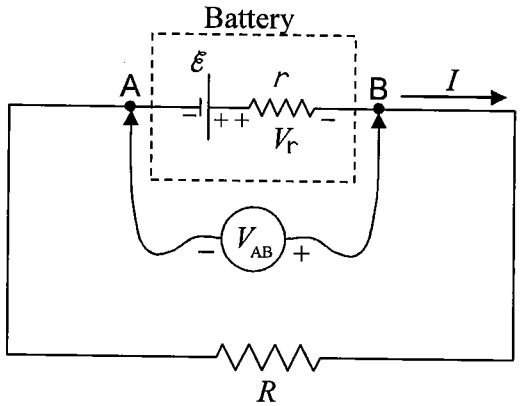


Note that I have also defined, in the preceding diagram, the variable V_r for the voltage across the internal resistance of the battery. Remember, to get the terminal voltage V_{AB} of the battery, all we have to do is to sum the potential changes that our fingertip would experience if we were to drag it from A to B in the circuit. (This is definitely a thought experiment because we can't get our fingertip inside the battery.)

$$V_{AB} = \mathcal{E} - V_r$$
$$V_{AB} = \mathcal{E} - Ir$$

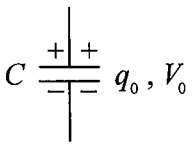
Note that, in the second line, I used the definition of resistance ($V=IR$) in the form $V_r = Ir$, to replace V_r with Ir.

We have been consistent, in this book, with the convention that a double subscript such as AB can be read "A to B" meaning, in the case at hand, that V_{AB} is the sum of the potential changes from A to B (rather than the other way around), in other words, that V_{AB} is how much higher the electric potential at point B is than the electric potential at point A. Still, there are some books out there that take V_{AB} (all by itself) to mean the voltage of A with respect to B (which is the negative of what we mean by it). So, for folks that may have used a different convention than you use, it is a good idea to diagrammatically define exactly what you mean by V_{AB}. Putting a voltmeter, labeled to indicate that it reads V_{AB}, and labeled to indicate which terminal is its "+" terminal and which is its "–" terminal is a good way to do this.

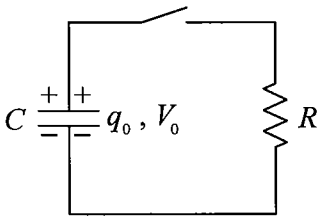


13 RC Circuits

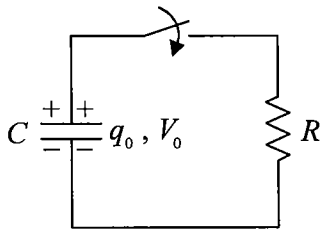
Suppose you connect a capacitor across a battery, and wait until the capacitor is charged to the extent that the voltage across the capacitor is equal to the EMF V_0 of the battery. Further suppose that you remove the capacitor from the battery. You now have a capacitor with voltage V_0 and charge q_0, where $q_0 = C\,V_0$.



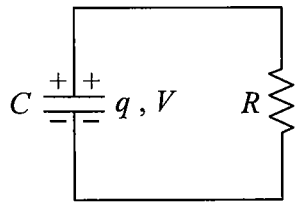
The capacitor is said to be charged. Now suppose that you connect the capacitor in series with an open switch and a resistor as depicted below.



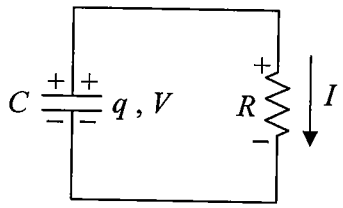
The capacitor remains charged as long as the switch remains open. Now suppose that, at a clock reading we shall call time zero, you close the switch.



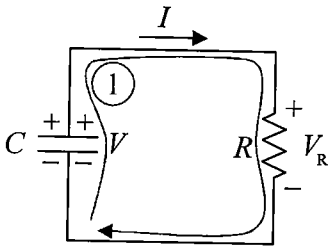
From time 0 on, the circuit is:



The potential across the resistor is now the same as the potential across the capacitor. This results in current through the resistor:



Positive charge flows from the upper plate of the capacitor, down through the resistor to the lower plate of the capacitor. The capacitor is said to be discharging. As the charge on the capacitor decreases; according to $q = CV$, which can be written $V = q/C$, the voltage across the capacitor decreases. But, as is clear from the diagram, the voltage across the capacitor is the voltage across the resistor. What we are saying is that the voltage across the resistor decreases. According to $V = IR$, which can be written as $I = V/R$, this means that the current through the resistor decreases. So, the capacitor continues to discharge at an ever-decreasing rate. Eventually, the charge on the capacitor decreases to a negligible value, essentially zero, and the current dies down to a negligible value, essentially zero. Of interest is how the various quantities, the voltage across both circuit elements, the charge on the capacitor, and the current through the resistor depend on the time t. Let's apply the loop rule to the circuit while the capacitor is discharging:



KVL (1)

$$+V - V_R = 0$$

Using $q = CV$ expressed as $V = \frac{q}{C}$ and $V_R = IR$, we obtain

$$\frac{q}{C} - IR = 0 \ .$$

I is the charge flow rate through the resistor, which is equivalent to the rate at which charge is being depleted from the capacitor (since the charge flowing through the resistor comes from the capacitor). Thus I is the negative of the rate of change of the charge on the capacitor:

$$I = -\frac{dq}{dt}$$

Substituting this ($I = -\frac{dq}{dt}$) into our loop rule equation ($\frac{q}{C} - IR = 0$) yields:

$$\frac{q}{C} + \frac{dq}{dt}R = 0$$

$$\frac{dq}{dt} = -\frac{1}{RC}q$$

Thus $q(t)$ is a function whose derivative with respect to time is itself, times the constant $-\frac{1}{RC}$. The function is essentially its own derivative. This brings the exponential function e^t to mind. The way to get that constant ($-\frac{1}{RC}$) to appear when we take the derivative of $q(t)$ with respect to t is to include it in the exponent. Try $q(t) = q_o e^{-\frac{1}{RC}t}$. Now, when you apply the chain rule for the function of a function you get $\frac{dq}{dt} = -\frac{1}{RC}q_o e^{-\frac{1}{RC}t}$ meaning that $\frac{dq}{dt} = -\frac{1}{RC}q$ which is just what we wanted. Let's check the units. R was defined as $\frac{V}{I}$ meaning the ohm is a volt per ampere. And C was defined as $\frac{q}{V}$ meaning that the farad is a coulomb per volt. So the units of the product RC are:

$$[RC] = \frac{\text{V}}{\text{A}}\frac{\text{coulombs}}{\text{V}} = \frac{\text{coulombs}}{\text{A}} = \frac{\text{coulombs}}{\text{coulombs/s}} = \text{s}$$

So the exponent in $e^{-\frac{1}{RC}t}$ is unitless. That works. We can't raise e to something that has units. Now, about that q_o out front in $q = q_o e^{-\frac{1}{RC}t}$. The exponential evaluates to a unitless quantity. So we need to put the q_o there to get units of charge. If you plug the value 0 in for the time in $q = q_o e^{-\frac{1}{RC}t}$ you get $q = q_o$. Thus, q_o is the initial value of the charge on the capacitor.

One final point: The product RC is called the "RC time constant." The symbol τ is often used to represent that time constant. In other words,

$$\tau = RC \qquad (13\text{-}1)$$

where τ is *also* referred to as the RC time constant. In terms of τ, our expression for q becomes:

$$q = q_o e^{-\frac{t}{\tau}}$$

which we copy here for your convenience:

$$q = q_o e^{-\frac{t}{\tau}}$$

Note that when t = τ, we have

$$q = q_o e^{-1}$$

$$q = \frac{1}{e} q_o$$

$\frac{1}{e}$ is .368 so τ is the time it takes for q to become 36.8% of its original value.

With our expression for q in hand, it is easy to get the expression for the voltage across the capacitor (which is the same as the voltage across the resistor, $V_C = V_R$) which we have been calling V. Substituting our expression $q = q_o e^{-\frac{1}{RC}t}$ into the defining equation for capacitance $q = CV$ solved for V,

$$V = \frac{q}{C}$$

yields:

$$V = \frac{q_o}{C} e^{-\frac{1}{RC}t}$$

But if q_0 is the charge on the capacitor at time 0, then $q_o = CV_0$ where V_0 is the voltage across the capacitor at time 0 or:

$$\frac{q_0}{C} = V_0 .$$

Substituting V_0 for $\frac{q_o}{C}$ in $V = \frac{q_o}{C} e^{-\frac{1}{RC}t}$ above yields:

$$V = V_0 e^{-\frac{t}{RC}} \qquad (13\text{-}2)$$

for both the voltage across the capacitor and the voltage across the resistor. From, the defining equation for resistance:

$$V = IR ,$$

we can write:

$$I = \frac{V}{R}$$

Substituting our expression $V_0 e^{-\frac{t}{RC}}$ in for V turns this equation $\left(I = \frac{V}{R}\right)$ into:

$$I = \frac{V_0 e^{-\frac{t}{RC}}}{R}$$

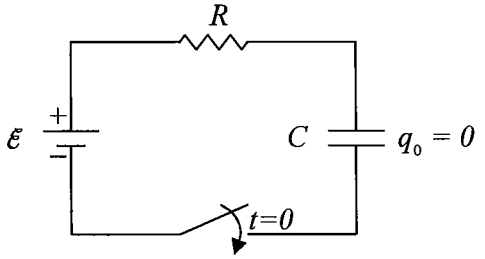
But, $\frac{V_0}{R}$ is just I_0 (from $V_0 = I_0 R$ solved for I_0), the current at the time 0, so:

$$I = I_0 e^{-\frac{t}{RC}} \tag{13-3}$$

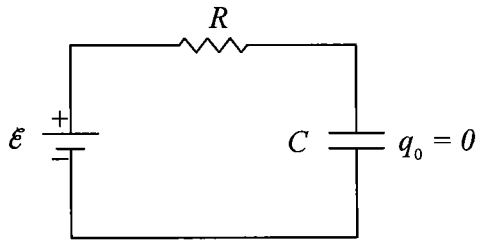
Summarizing, we note that all three of the quantities, V, I, and q decrease exponentially with time.

Charging Circuit

Consider the following circuit, containing an initially-uncharged capacitor, and

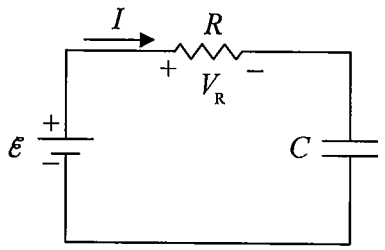


annotated to indicate that the switch is closed at time 0 at which point the circuit becomes:

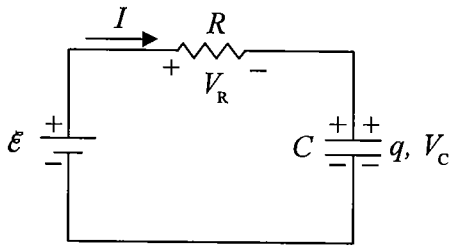


Let's think about what will happen as time elapses. With no charge on the capacitor, the voltage across it is zero, meaning the potential of the right terminal of the resistor is the same as the potential of the lower-potential terminal of the seat of EMF. Since the left end of the resistor is connected to the higher-potential terminal of the seat of EMF, this means that at time 0, the

voltage across the resistor is equivalent to the EMF $\mathcal{E}$ of the seat of EMF. Thus, there will be a rightward current through the resistor.



The positive charge flowing through the resistor has to come from someplace. Where does it come from? Answer: The bottom plate of the capacitor. Also, charge can't flow through an ideal capacitor. So where does it go? It piles up on the top plate of the capacitor.



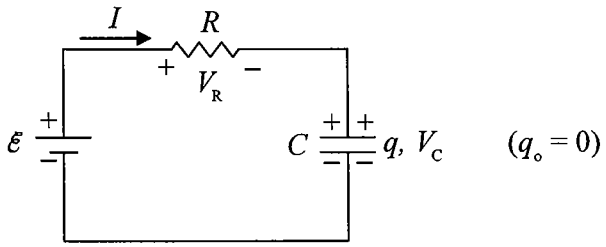
The capacitor is becoming charged. As it does, the voltage across the capacitor increases, meaning the potential of the right terminal of the resistor (relative to the potential of the lower-potential terminal of the seat of EMF) increases. The potential of the left terminal of the resistor remains constant, as dictated by the seat of EMF. This means that the voltage across the resistor continually decreases. This, in turn; from $V_R = IR$, written as $I = V_R/R$, means that the current continually decreases. This occurs until there is so much charge on the capacitor that $V_c = \mathcal{E}$, meaning that $V_R = 0$ so $I = 0$.

Recapping our conceptual discussion:

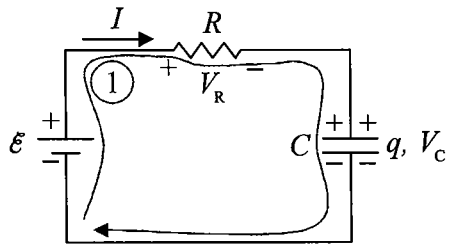
At time 0, we close the switch:

- The charge on the capacitor starts off at 0 and builds up to $q = C\mathcal{E}$ where $\mathcal{E}$ is the EMF voltage.
- The capacitor voltage starts off at 0 and builds up to the EMF voltage $\mathcal{E}$.
- The current starts off at $I_0 = \frac{\mathcal{E}}{R}$ and decreases to 0.

Okay, we have a qualitative understanding of what happens. Let's see if we can obtain formulas for V_R, I, V_C, and q as functions of time. Here's the circuit:



We apply the loop rule:



KVL (1)

$$+\mathcal{E} - V_R - V_C = 0$$

and the definitions of resistance and capacitance:

$$V_R = IR \qquad q = CV_C \qquad V_C = \frac{q}{C}$$

to obtain:

$$\mathcal{E} - IR - \frac{q}{C} = 0$$

$$IR + \frac{q}{C} = \mathcal{E}$$

Then we use the fact that the current is equal to the rate at which charge is building up on the capacitor, $I = \frac{dq}{dt}$, to get:

$$\frac{dq}{dt}R + \frac{q}{C} = \mathcal{E}$$

$$\frac{dq}{dt} + \frac{q}{RC} = \frac{\mathcal{E}}{R}$$

This is interesting. This is the same equation that we had before, except that we have the constant $\mathcal{E}$/R on the right instead of 0.

For this equation, I'm simply going to provide and discuss the solution, rather than show you how to solve the differential equation. The charge function of time that solves this equation is:

$$q = C\mathcal{E}\left(1 - e^{-\frac{t}{RC}}\right)$$

Please substitute it into the differential equation $\left(\frac{dq}{dt} + \frac{q}{RC} = \frac{\mathcal{E}}{R}\right)$ and verify that it leads to an identity.

Now let's check to make sure that $q = C\mathcal{E}\left(1 - e^{-\frac{t}{RC}}\right)$ is consistent with our conceptual understanding. At time zero ($t = 0$), our expression $q(t) = C\mathcal{E}\left(1 - e^{-\frac{t}{RC}}\right)$ evaluates to:

$$\begin{aligned} q(0) &= C\mathcal{E}\left(1 - e^{-\frac{0}{RC}}\right) \\ &= C\mathcal{E}\left(1 - e^{0}\right) \\ &= C\mathcal{E}\left(1 - 1\right) \\ q(0) &= 0 \end{aligned}$$

Excellent. This is consistent with the fact that the capacitor starts out uncharged.

Now, what does our charge function $q(t) = C\mathcal{E}\left(1 - e^{-\frac{t}{RC}}\right)$ say about what happens to the charge of the capacitor in the limit as t goes to infinity?

$$\begin{aligned}\lim_{t\to\infty} q(t) &= \lim_{t\to\infty} C\mathcal{E}\left(1-e^{-\frac{t}{RC}}\right)\\ &= C\mathcal{E}\lim_{x\to\infty}\left(1-e^{-x}\right)\\ &= C\mathcal{E}\lim_{x\to\infty}\left(1-\frac{1}{e^{x}}\right)\\ &= C\mathcal{E}\lim_{y\to\infty}\left(1-\frac{1}{y}\right)\\ &= C\mathcal{E}\left(1-\lim_{y\to\infty}\frac{1}{y}\right)\\ &= C\mathcal{E}\left(1-0\right)\\ \lim_{t\to\infty} q(t) &= C\mathcal{E}\end{aligned}$$

Well, this makes sense. Our conceptual understanding was that the capacitor would keep charging until the voltage across the capacitor was equal to the voltage across the seat of EMF. From the definition of capacitance, when the capacitor voltage is $\mathcal{E}$, its charge is indeed $C\mathcal{E}$. The formula yields the expected result for $\lim_{t\to\infty} q(t)$.

Once we have $q(t)$ it is pretty easy to get the other circuit quantities. For instance, from the definition of capacitance:

$$q = CV_{\mathrm{C}},$$

we have $V_{\mathrm{C}} = q/C$ which, with $q = C\mathcal{E}\left(1-e^{-\frac{t}{RC}}\right)$ evaluates to:

$$V_{\mathrm{C}} = \mathcal{E}\left(1-e^{-\frac{t}{RC}}\right) \tag{13-4}$$

Our original loop equation read:

$$\mathcal{E} - V_{\mathrm{R}} - V_{\mathrm{C}} = 0$$

So:

$$V_{\mathrm{R}} = \mathcal{E} - V_{\mathrm{C}}$$

which, with $V_{\mathrm{C}} = \mathcal{E}\left(1-e^{-\frac{t}{RC}}\right)$ can be written as:

$$V_{\mathrm{R}} = \mathcal{E} - \mathcal{E}\left(1-e^{-\frac{t}{RC}}\right)$$

$$V_R = \mathcal{E} - \mathcal{E} + \mathcal{E} e^{-\frac{t}{RC}}$$

$$V_R = \mathcal{E} e^{-\frac{t}{RC}}$$

From our definition of resistance:

$$V_R = IR$$

$$I = \frac{V_R}{R}$$

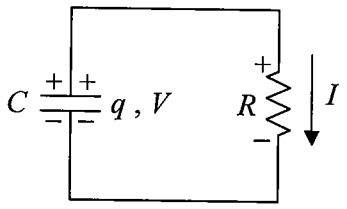
with $V_R = \mathcal{E} e^{-\frac{t}{RC}}$, this can be expressed as:

$$I = \frac{\mathcal{E}}{R} e^{-\frac{t}{RC}}$$

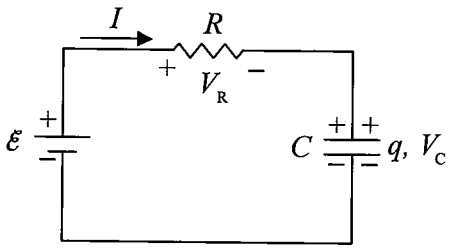
At time 0, this evaluates to $\mathcal{E}/R$ meaning that $\mathcal{E}/R$ can be interpreted as the current at time 0 allowing us to write our function $I(t)$ as

$$I = I_0 e^{-\frac{t}{RC}}$$

Our formula has the current starting out at its maximum value and decreasing exponentially with time, as anticipated based on our conceptual understanding of the circuit. Note that this is the same formula that we got for the current in the discharging-capacitor circuit. In both cases, the current dies off exponentially. The reasons differ, but the effect ($I = I_0 e^{-\frac{t}{RC}}$) is the same:



In the discharging-capacitor circuit, the current dies off because the capacitor runs out of charge.



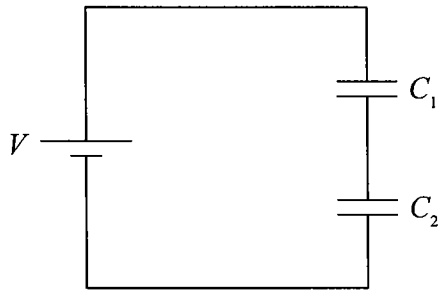
In the charging-capacitor circuit, the current dies off because the capacitor voltage, which counteracts the EMF, builds up to $\mathcal{E}$ as the capacitor charges.

14 Capacitors in Series & Parallel

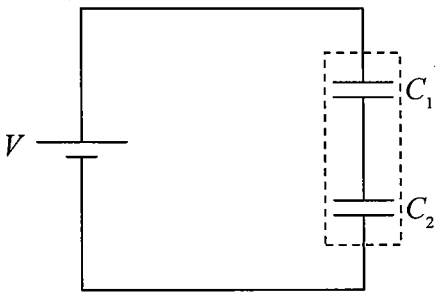
The method of ever-simpler circuits that we used for circuits with more than one resistor can also be used for circuits having more than one capacitor. The idea is to replace a combination circuit element consisting of more than one capacitor with a single equivalent capacitor. The equivalent capacitor should be equivalent in the sense that, with the same potential across it, it will have the same charge as the combination circuit element.

Capacitors in Series

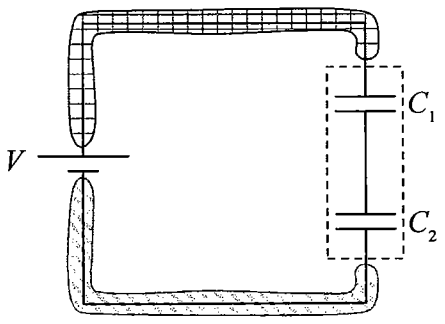
Let's start with a case in which the combination circuit element consists of two capacitors in series with each other:



We consider the two capacitors to be a two-terminal combination circuit element:



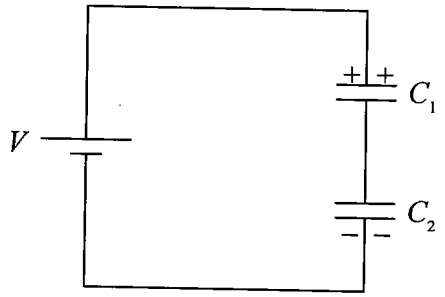
The voltage across the combination circuit element is clearly the EMF voltage V since, for both the seat of EMF and the combination circuit element, we're talking about the potential difference between the same two conductors:



The voltage across each individual capacitor is, however, not known.

But consider this: After that last wire is connected in the circuit, the charging process (which takes essentially no time at all) can be understood to proceed as follows (where, for ease of understanding, we describe things that occur simultaneously as if they occurred sequentially):

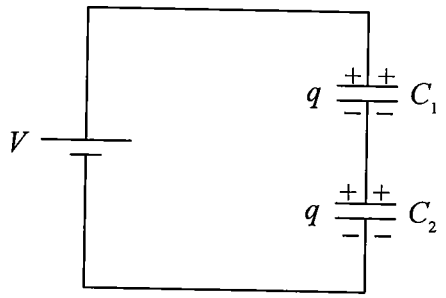
The seat of EMF pulls some positive charge from the bottom plate of the lower capacitor and pushes it onto the top plate of the upper capacitor.



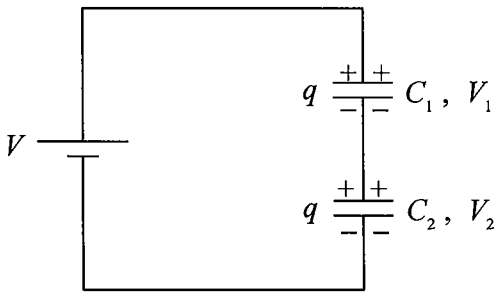
The key point about this movement of charge is that the amount of positive charge on the top plate of the upper capacitor is exactly equal to the amount of negative charge on the bottom plate of the lower capacitor (because that's where the positive charge came from!)

Now, the positive charge on the upper plate of the top capacitor repels the positive charge (remember, every neutral object consists of huge amounts of both kinds of charge, and, in our positive-charge-carrier convention, the positive charges are free to move) on the bottom plate of the upper capacitor and that charge has a conducting path to the top plate of the lower capacitor, to which it (the positive charge) is attracted by the negative charge on the bottom plate of the lower capacitor.

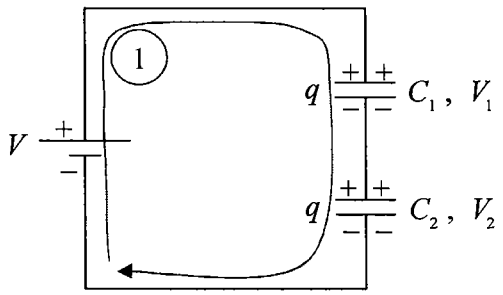
The final result is that both capacitors have one and the same charge q:



which in turn causes capacitor C_1 to have voltage $V_1 = \frac{q}{C_1}$ and capacitor C_2 to have voltage $V_2 = \frac{q}{C_2}$.



By the loop rule,



KVL (1)

$$V - V_1 - V_2 = 0$$

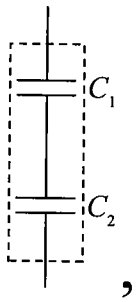
$$V = V_1 + V_2$$

$$V = \frac{q}{C_1} + \frac{q}{C_2}$$

$$V = q\left(\frac{1}{C_1} + \frac{1}{C_2}\right)$$

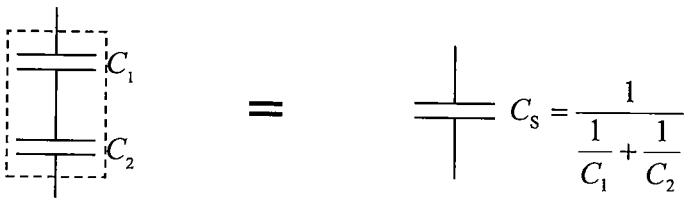
$$q = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2}} V$$

So, what we're saying is, that when you put a voltage V across the two-terminal circuit element



,

an amount of charge $q = \dfrac{1}{\dfrac{1}{C_1}+\dfrac{1}{C_2}}V$ is moved from the bottom terminal of the combination circuit element, around the circuit, to the top terminal. Then charge stops moving. Recall that we defined the capacitance of a capacitor to be the ratio $\frac{q}{V}$ of the charge on the capacitor to the corresponding voltage across the capacitor. $\frac{q}{V}$ for our two-terminal combination circuit element is thus the equivalent capacitance of the two terminal circuit element. Solving $q = \dfrac{1}{\dfrac{1}{C_1}+\dfrac{1}{C_2}}V$ for the ratio $\frac{q}{V}$ yields $\dfrac{q}{V} = \dfrac{1}{\dfrac{1}{C_1}+\dfrac{1}{C_2}}$ so our equivalent capacitance for two capacitors in series is $C_S = \dfrac{1}{\dfrac{1}{C_1}+\dfrac{1}{C_2}}$



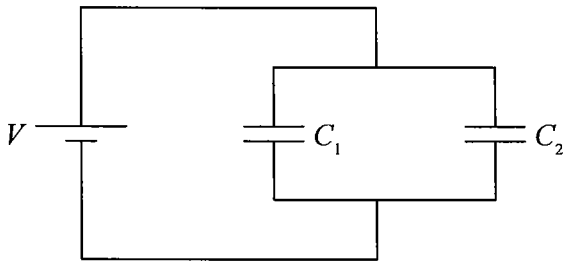
By logical induction, we can extend this argument to cover any number of capacitors in series with each other, obtaining:

$$C_S = \frac{1}{\dfrac{1}{C_1}+\dfrac{1}{C_2}+\dfrac{1}{C_3}+\ldots} \qquad (14\text{-}1)$$

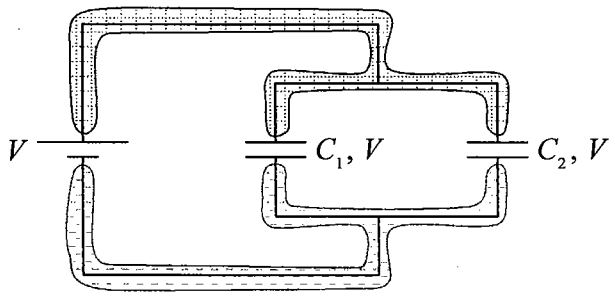
As far as making things easy to remember, it's just too bad the way things work out sometimes. This expression is mathematically identical to the expression for <u>resistors</u> in *parallel*. But, *this* expression is for <u>capacitors</u> in *series*.

Capacitors in Parallel

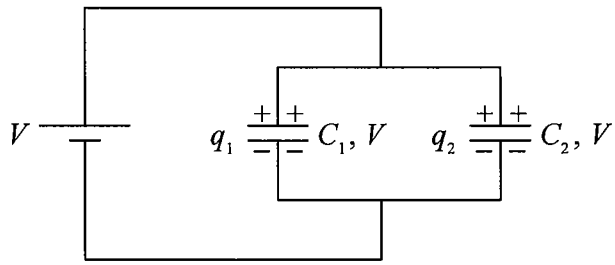
Suppose we put a voltage V across a combination circuit element consisting of a pair of capacitors in parallel with each other:



It is clear from the diagram that the voltage across each capacitor is just the EMF V since the voltage across every component in the circuit is the potential difference between the same two conductors.



So what happens (almost instantaneously) when we make that final connection? Answer: The seat of EMF pulls charge off the bottom plates of the two capacitors and pushes it onto the top plates until the charge on C_1 is $q_1 = C_1 V$ and the charge on C_2 is $q_2 = C_2 V$.



To do that, the seat of EMF has to move a total charge of

$$q = q_1 + q_2$$

$$q = C_1 V + C_2 V$$

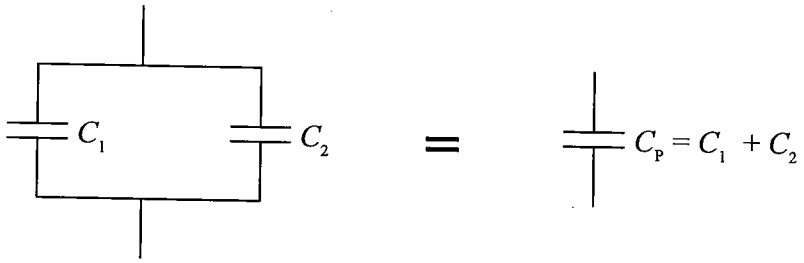
$$q = (C_1 + C_2) V$$

Solving the last equation, $q = (C_1 + C_2)\,V$, for the equivalent capacitance C_P, defined as q/V, yields:

$$\frac{q}{V} = C_1 + C_2$$

$$C_P = C_1 + C_2$$

In other words:



So, the equivalent capacitance of capacitors in parallel is simply the sum of the individual capacitances. (This is the way resistors in *series* combine.) By means of inductive reasoning, the result can be extended to any number of capacitors, yielding:

$$C_P = C_1 + C_2 + C_3 + \ldots \qquad (14\text{-}2)$$

Concluding Remarks

The facts that the voltage is the same for capacitors in parallel and the charge is the same for capacitors in series are important, but, if you look at these as two more things that you have to commit to memory then you are not going about your study of physics the right way. You need to be able to "see" that the charge on capacitors in series has to be the same because the charge on one capacitor comes from its (originally-neutral) neighbor. You need to be able to "see" that the voltage across capacitors in parallel has to be the same because, for each capacitor, the voltage is the potential difference between the *same* two conductors.

15 Magnetic Field Intro: Effects

We now begin our study of magnetism, and, analogous to the way in which we began our study of electricity, we start by discussing the effect of a given magnetic field without first explaining how such a magnetic field might be caused to exist. We delve into the causes of magnetic fields in subsequent chapters.

A magnetic field is a vector field. That is, it is an infinite set of vectors, one at each point in the region of space where the magnetic field exists. We use the expression "magnetic field" to designate both the infinite set of vectors, and, when one is talking about the magnetic field at a point in space, the one magnetic field vector at that point in space. We use the symbol $\vec{\mathbf{B}}$ to represent the magnetic field. The most basic effect of a magnetic field is to exert a torque on an object that has a property known as *magnetic dipole moment*, and, that finds itself in the magnetic field. A particle or object that has a non-zero value of magnetic dipole moment is called a magnetic dipole. A magnetic dipole is a bar magnet. The value of the magnitude of the magnetic dipole moment of an object is a measure of how strong a bar magnet it is. A magnetic dipole has two ends, known as poles—a north pole and a south pole. Magnetic dipole moment is a property of matter which has direction. We can define the direction, of the magnetic dipole moment of an object, by considering the object to be an arrow whose north pole is the arrowhead and whose south pole is the tail. The direction in which the arrow is pointing is the direction of the magnetic dipole moment of the object. The unit of magnetic dipole moment is the $\text{A}\cdot\text{m}^2$ (ampere meter-squared)[1]. While magnetic compass needles come in a variety of magnetic dipole moments, a representative value for the magnetic dipole moment of a compass needle is $.1\ \text{A}\cdot\text{m}^2$.

Again, the most basic effect of a magnetic field is to exert a torque on a magnetic dipole that finds itself in the magnetic field. The magnetic field vector, at a given point in space, is the maximum possible torque-per-magnetic-dipole-moment-of-would-be-victim that the magnetic field would/will exert on any magnetic dipole (victim) that might find itself at the point in question. I have to say "maximum possible" because the torque exerted on the magnetic dipole depends not only on the magnitude of the magnetic field at the point in space and the magnitude of the magnetic dipole moment of the victim, but it also depends on the orientation of the magnetic dipole relative to the direction of the magnetic field vector. In fact:

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{\mu}} \times \vec{\mathbf{B}} \tag{15-1}$$

where:

- $\vec{\boldsymbol{\tau}}$ is the torque exerted on the magnetic dipole (the bar magnet) by the magnetic field,
- $\vec{\boldsymbol{\mu}}$ is the magnetic dipole moment of the magnetic dipole (the bar magnet, the victim), and
- $\vec{\mathbf{B}}$ is the magnetic field vector at the location in space at which the magnetic dipole is.

---

[1] Magnetic dipole moment magnitude μ is a fundamental property of matter, as fundamental as mass m and charge q. For the elementary particle known as the electron: $m = 9.11 \times 10^{-31}$ kg, $q = 1.60 \times 10^{-19}$ C, and $\mu = 9.27 \times 10^{-24}\ \text{A}^2\cdot\text{m}$. Calling the unit of magnetic dipole moment the $\text{A}\cdot\text{m}^2$ is about as illuminating as calling the unit of mass the $\text{N}\cdot\text{s}/\text{m}^2$, or calling the unit of charge the $\text{A}\cdot\text{s}$ (both of which are correct). It would be nice if there was a name for the unit of magnetic dipole moment in the SI system of units, but there isn't. There is a non-SI unit of magnetic dipole moment. It is called the Bohr magneton, abbreviated μ_B. In units of Bohr magnetons, the magnetic moment of the electron is $1\,\mu_B$.

For the cross product of any two vectors, the magnitude of the cross product is the product of the magnitudes of the two vectors, times the sine of the angle the two vectors form when placed tail to tail. In the case of $\vec{\tau} = \vec{\mu} \times \vec{B}$, this means:

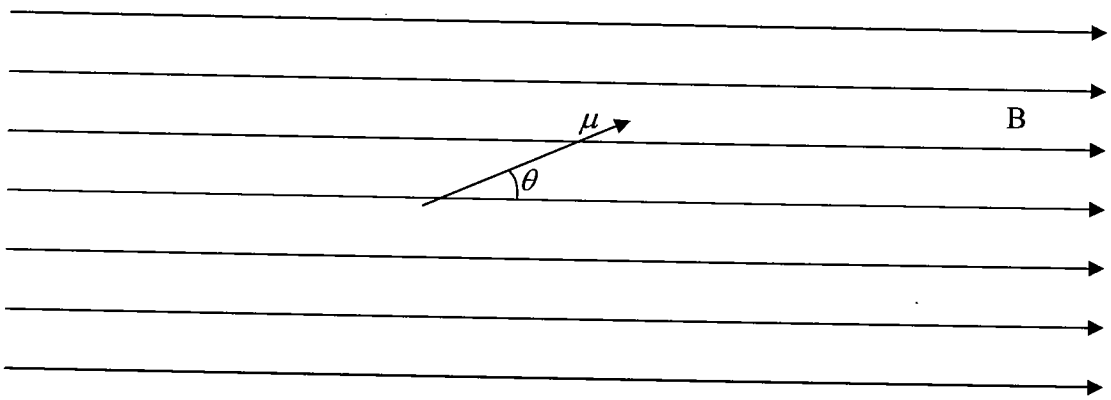
$$\tau = \mu B \sin\theta$$

In the SI system of units, torque has units of N·m (newton-meters). For the units on the right side of $\tau = \mu B \sin\theta$ to work out to be N·m, what with μ having units of electric dipole moment ($A \cdot m^2$) and $\sin\theta$ having no units at all, B must have units of torque-per-magnetic-dipole-moment, namely, $\frac{N \cdot m}{A \cdot m^2}$. That combination unit is given a name. It is called the tesla, abbreviated T.

$$1\ T = 1\frac{N \cdot m}{A \cdot m^2}$$

Example 15-1

Consider a magnetic dipole having a magnetic dipole moment $\mu = 0.045\ A \cdot m^2$, oriented so that it makes an angle of 23° with the direction of a uniform magnetic field of magnitude 5.0×10^{-5} T as depicted below. Find the torque exerted on the magnetic dipole, by the magnetic field.



Recall that the arrowhead represents the north pole of the bar magnet that a magnetic dipole is. The direction of the torque is such that it tends to cause the magnetic dipole to point in the direction of the magnetic field. For the case depicted above, that would be clockwise as viewed from the vantage point of the creator of the diagram. The magnitude of the torque for such a case can be calculated as follows:

$$\tau = \mu B \sin\theta$$

$$\tau = \left(.045\ A \cdot m^2\right)\left(5.0 \times 10^{-5} T\right)\sin 23°$$

$$\tau = 8.8 \times 10^{-7} A \cdot m^2 \cdot T$$

Recalling that a tesla is a $\frac{\text{N}\cdot\text{m}}{\text{A}\cdot\text{m}^2}$ we have:

$$\tau = 8.8 \times 10^{-7}\ \text{A}\cdot\text{m}^2 \cdot \frac{\text{N}\cdot\text{m}}{\text{A}\cdot\text{m}^2}$$

$$\tau = 8.8 \times 10^{-7}\ \text{N}\cdot\text{m}$$

Thus, the torque on the magnetic dipole is 8.8×10^{-7} N·m clockwise, as viewed from the vantage point of the creator of the diagram.

Example 15-2

A particle having a magnetic dipole moment $\vec{\mu} = 0.025\ \text{A}\cdot\text{m}^2\,\hat{\imath} - 0.035\ \text{A}\cdot\text{m}^2\,\hat{\jmath} + 0.015\ \text{A}\cdot\text{m}^2\,\hat{k}$ is at a point in space where the magnetic field $\vec{B} = 2.3\,\text{mT}\,\hat{\imath} + 5.3\,\text{mT}\,\hat{\jmath} - 3.6\,\text{mT}\,\hat{k}$.
Find the torque exerted on the particle by the magnetic field.

$$\vec{\tau} = \vec{\mu} \times \vec{B}$$

$$\vec{\tau} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ 0.025\,\text{Am}^2 & -0.035\,\text{Am}^2 & 0.015\,\text{Am}^2 \\ 0.0023\frac{\text{Nm}}{\text{Am}^2} & 0.0053\frac{\text{Nm}}{\text{Am}^2} & -0.0036\frac{\text{Nm}}{\text{Am}^2} \end{vmatrix}$$

$$\vec{\tau} = \hat{\imath}\left[\left(-0.035\,\text{Am}^2\right)\left(-0.0036\frac{\text{Nm}}{\text{Am}^2}\right) - \left(0.015\,\text{Am}^2\right)\left(0.0053\frac{\text{Nm}}{\text{Am}^2}\right)\right]$$

$$+\hat{\jmath}\left[\left(0.015\,\text{Am}^2\right)\left(0.0023\frac{\text{Nm}}{\text{Am}^2}\right) - \left(0.025\,\text{Am}^2\right)\left(-0.0036\frac{\text{Nm}}{\text{Am}^2}\right)\right]$$

$$+\hat{k}\left[\left(0.025\,\text{Am}^2\right)\left(0.0053\frac{\text{Nm}}{\text{Am}^2}\right) - \left(-0.035\,\text{Am}^2\right)\left(0.0023\frac{\text{Nm}}{\text{Am}^2}\right)\right]$$

$$\vec{\tau} = 1.2\times10^{-4}\,\text{Nm}\,\hat{\imath} - 1.2\times10^{-4}\,\text{Nm}\,\hat{\jmath} + 2.1\times10^{-4}\,\text{Nm}\,\hat{k}$$

The Magnetic Force Exerted Upon a Magnetic Dipole

A uniform magnetic field exerts no force on a bar magnet that is in the magnetic field. You should probably pause here for a moment and let that sink in. A uniform magnetic field exerts no force on a bar magnet that is in that magnetic field.

You have probably had some experience with bar magnets. You know that like poles repel and unlike poles attract. And, from your study of the electric field, you have probably (correctly) hypothesized that in the field point of view, the way we see this is that one bar magnet (call it the source magnet) creates a magnetic field in the region of space around itself, and, that if there is another bar magnet in that region of space, it will be affected by the magnetic field it is in. We have already discussed the fact that the victim bar magnet will experience a torque. But you know, from your experience with bar magnets, that it will also experience a force. How can that be when I just stated that a uniform magnetic field exerts no force on a bar magnet? Yes, of course. The magnetic field of the source magnet must be non-uniform. Enough about the nature of the magnetic field of a bar magnet, I'm supposed to save that for an upcoming chapter. Suffice it to say that it is non-uniform and to focus our attention on the effect of a non-uniform field on a bar magnet that finds itself in that magnetic field.

First of all, a non-uniform magnetic field will exert a torque on a magnetic dipole (a bar magnet) just as before ($\vec{\tau} = \vec{\mu} \times \vec{\mathbf{B}}$). But, a non-uniform magnetic field (one for which the magnitude, and/or direction, depends on position) also exerts a *force* on a magnetic dipole. The force is given by:

$$\vec{\mathbf{F}}_B = \nabla(\vec{\mu} \cdot \vec{\mathbf{B}}) \tag{15-2}$$

where

- $\vec{\mathbf{F}}_B$ is the force exerted by the magnetic field $\vec{\mathbf{B}}$ on a particle having a magnetic dipole moment $\vec{\mu}$,
- $\vec{\mu}$ is the magnetic dipole of the "victim", and,
- $\vec{\mathbf{B}}$ is the magnetic field at the position in space where the victim finds itself. To evaluate the force, one must know $\vec{\mathbf{B}}$ as a function of x, y, and z (whereas $\vec{\mu}$ is a constant).

Note that after you take the gradient of $\vec{\mu} \cdot \vec{\mathbf{B}}$, you have to evaluate the result at the values of x, y, and z corresponding to the location of the victim.

Just to make sure that you know how to use this equation, please note that if $\vec{\mu}$ and $\vec{\mathbf{B}}$ are expressed in $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, $\hat{\mathbf{k}}$ notation, so that they appear as $\vec{\mu} = \mu_x\hat{\mathbf{i}} + \mu_y\hat{\mathbf{j}} + \mu_z\hat{\mathbf{k}}$ and $\vec{\mathbf{B}} = B_x\hat{\mathbf{i}} + B_y\hat{\mathbf{j}} + B_z\hat{\mathbf{k}}$ respectively, then:

$$\vec{\mu} \cdot \vec{\mathbf{B}} = (\mu_x\hat{\mathbf{i}} + \mu_y\hat{\mathbf{j}} + \mu_z\hat{\mathbf{k}}) \cdot (B_x\hat{\mathbf{i}} + B_y\hat{\mathbf{j}} + B_z\hat{\mathbf{k}})$$

$$\vec{\mu} \cdot \vec{\mathbf{B}} = \mu_x B_x + \mu_y B_y + \mu_z B_z$$

And the gradient of $\vec{\mu}\cdot\vec{\mathbf{B}}$ (which by equation 15-2 is the force we seek) is given by

$$\nabla(\vec{\mu}\cdot\vec{\mathbf{B}}) = \frac{\partial(\vec{\mu}\cdot\vec{\mathbf{B}})}{\partial x}\hat{\imath} + \frac{\partial(\vec{\mu}\cdot\vec{\mathbf{B}})}{\partial y}\hat{\jmath} + \frac{\partial(\vec{\mu}\cdot\vec{\mathbf{B}})}{\partial z}\hat{\mathbf{k}}$$

where derivatives in this equation can (using $\vec{\mu}\cdot\vec{\mathbf{B}} = \mu_x B_x + \mu_y B_y + \mu_z B_z$ from just above) can be expressed as:

$$\frac{\partial(\vec{\mu}\cdot\vec{\mathbf{B}})}{\partial x} = \mu_x\frac{\partial B_x}{\partial x} + \mu_y\frac{\partial B_y}{\partial x} + \mu_z\frac{\partial B_z}{\partial x},$$

$$\frac{\partial(\vec{\mu}\cdot\vec{\mathbf{B}})}{\partial y} = \mu_x\frac{\partial B_x}{\partial y} + \mu_y\frac{\partial B_y}{\partial y} + \mu_z\frac{\partial B_z}{\partial y}, \text{ and}$$

$$\frac{\partial(\vec{\mu}\cdot\vec{\mathbf{B}})}{\partial z} = \mu_x\frac{\partial B_x}{\partial z} + \mu_y\frac{\partial B_y}{\partial z} + \mu_z\frac{\partial B_z}{\partial z};$$

where we have taken advantage of the fact that the components of the magnetic dipole moment of the victim are not functions of position. Also note that the derivatives are all partial derivatives. Partial derivatives are the easy kind in the sense that, when, for instance, you take the derivative with respect to x, you are to treat y and z as if they were constants. Finally, it is important to realize that, after you take the derivatives, you have to plug the values of x, y, and z corresponding to the location of the magnetic dipole (the victim), into the given expression for the force.

Example 15-3

There exists, in a region of space, a magnetic field, given in terms of Cartesian unit vectors by:

$$\vec{\mathbf{B}} = -5.82\times10^{-6}\,\text{T·m}\frac{y}{x^2+y^2}\hat{\imath} + 5.82\times10^{-6}\,\text{T·m}\frac{x}{x^2+y^2}\hat{\jmath}$$

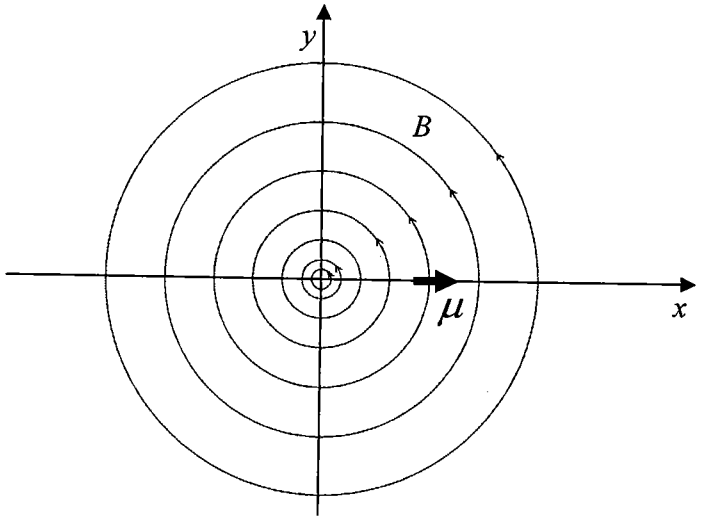
A particle is in the region of space where the magnetic field exists. The particle has a magnetic dipole moment given by:

$$\vec{\mu} = .5\,14\,\text{A·m}^2\,\hat{\imath}$$

The particle is at (0.110 m, 0, 0).

Find the force exerted on the particle by the magnetic field.

Solution: First, we sketch the configuration:



Substituting the given $\vec{\mu}$ and $\vec{B}$, into our expression for the force yields:

$$\vec{F}_B = \nabla(\vec{\mu}\cdot\vec{B})$$

$$\vec{F}_B = \nabla[(.514\,\text{A}\cdot\text{m}^2\hat{\imath})\cdot(-5.82\times10^{-6}\,\text{T}\cdot\text{m}\frac{y}{x^2+y^2}\hat{\imath} + 5.82\times10^{-6}\,\text{T}\cdot\text{m}\frac{x}{x^2+y^2}\hat{\jmath})]$$

$$\vec{F}_B = \nabla\left(-2.99\times10^{-6}\,\text{A}\cdot\text{m}^2\cdot\text{T}\cdot\text{m}\frac{y}{x^2+y^2}\right)$$

$$\vec{F}_B = -2.99\times10^{-6}\,\text{N}\cdot\text{m}^2\,\nabla[y(x^2+y^2)^{-1}]$$

$$\vec{F}_B = -2.99\times10^{-6}\,\text{N}\cdot\text{m}^2\left\{\frac{\partial}{\partial x}[y(x^2+y^2)^{-1}]\hat{\imath}+\frac{\partial}{\partial y}[y(x^2+y^2)^{-1}]\hat{\jmath}+\frac{\partial}{\partial z}[y(x^2+y^2)^{-1}]\hat{k}\right\}$$

$$\vec{F}_B = -2.99\times10^{-6}\,\text{N}\cdot\text{m}^2\left\{[y(-1)(x^2+y^2)^{-2}2x\hat{\imath}+[(x^2+y^2)^{-1}+y(-1)(x^2+y^2)^{-2}2y]\hat{\jmath}+0\hat{k}\right\}$$

$$\vec{F}_B = -2.99\times10^{-6}\,\text{N}\cdot\text{m}^2\left\{-\frac{2xy}{(x^2+y^2)^2}\hat{\imath}+\left[\frac{1}{x^2+y^2}-\frac{2y^2}{(x^2+y^2)^2}\right]\hat{\jmath}\right\}$$

Recalling that we have to evaluate this expression at the location of the victim, a location that was given as (0.110 m, 0, 0), we find that:

$$\vec{F}_B = -2.99\times10^{-6}\,\text{N}\cdot\text{m}^2\left\{-\frac{2(0.110\,\text{m})\,0}{[(0.110\,\text{m})^2+0^2]^2}\hat{\imath}+\left[\frac{1}{(0.110\,\text{m})^2+0^2}-\frac{2(0)^2}{[(0.110\,\text{m})^2+0^2]^2}\right]\hat{\jmath}\right\}$$

$$\vec{F}_B = -2.47\times10^{-4}\,\text{N}\,\hat{\jmath}$$

Characteristics of the Earth's Magnetic Field

We live in a magnetic field produced by the earth. Both its magnitude and its direction are different at different locations on the surface of the earth. Furthermore, at any given location, the earth's magnetic field varies from year to year in both magnitude and direction. Still, on the geographical scale of a college campus, and, on a time scale measured in days, the earth's magnetic field is approximately uniform and constant.

To align your index finger with the magnetic field of the earth on the Saint Anselm College campus, first point in the horizontal direction 15.4° West of North[2]. Then tilt your arm downward so that you are pointing in a direction that is 68.9° below the horizontal. (Yes! Can you believe it? It's mostly downward!) You are now pointing your finger in the direction of the earth's magnetic field. The magnitude of the magnetic field, on the Saint Anselm College campus, is 5.37×10^{-5} T. In other words:

The Earth's Magnetic Field on the Saint Anselm College Campus in 2006

Characteristic	Value	Rate of Change
Declination	−15.4°	+0.074°/year
Inclination (Dip Angle)	68.8°	−0.096°/year
Magnitude	5.36×10^{-5} T	-0.012×10^{-5} T/year
Horizontal Component	1.93×10^{-5} T	$+0.004 \times 10^{-5}$ T/year
Vertical Component	5.00×10^{-5} T	-0.014×10^{-5} T/year

A compass needle is a tiny bar magnet that is constrained to rotate about a vertical axis. The earth's magnetic field exerts a torque on the compass needle that tends to make the compass needle point in the direction of the horizontal component of the earth's magnetic field, a direction we call "magnetic north". Recall that when we talk about which way a bar magnet (such as a compass needle) is pointing, we imagine there to be an arrowhead at its north pole.

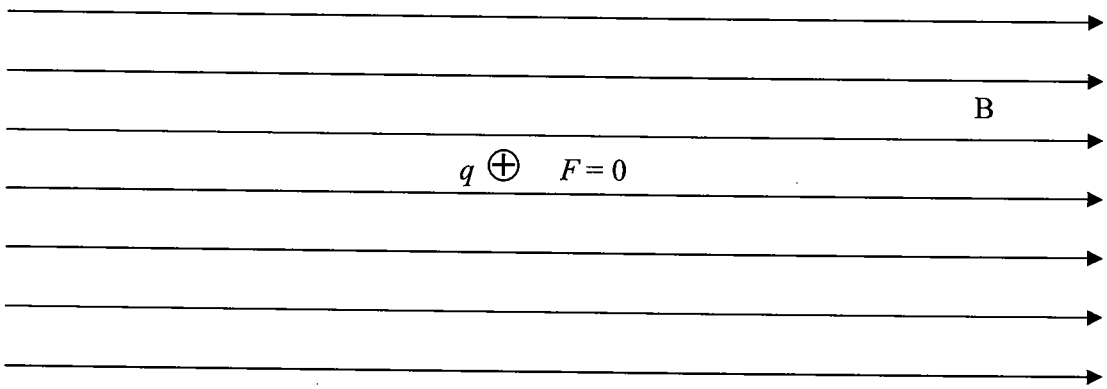
[2] The values of the earth's magnetic field presented here were obtained from the (United States) National Geophysical Data Center (NGDC) geomagnetism web site at http://www.ngdc.noaa.gov/seg/geomag/geomag.shtml. I used the magnetic field values calculator at the site to obtain the presented values. I used latitude 42° 59' 7", longitude 71° 30' 20" (the location of my office in the Goulet Science Center, obtained from a topographic map) and date February 20, 2006 as input values. Check out the web site. It provides some interesting insight into the earth's magnetic field.

16 Magnetic Field: More Effects

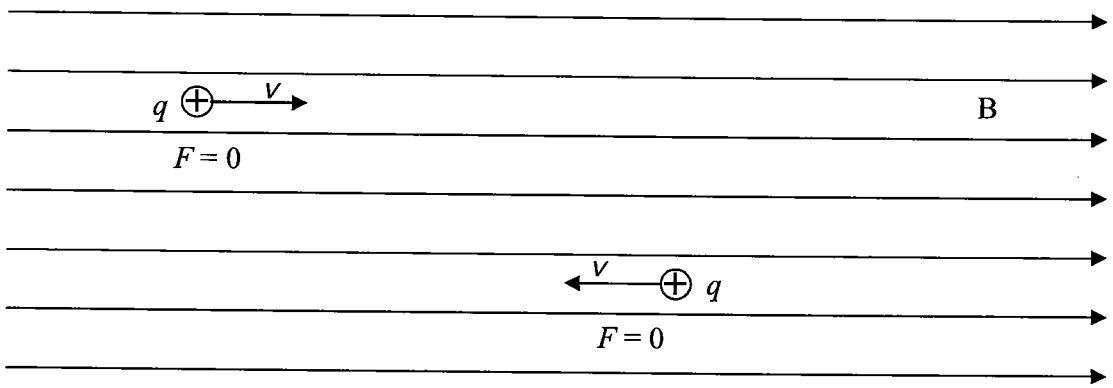
The electric field and the magnetic field are not the same thing. An electric dipole with positive charge on one end and negative charge on the other is not the same thing as a magnetic dipole having a north and a south pole. More specifically: An object can have positive charge but it can't have "northness".

On the other hand, electricity and magnetism are not unrelated. In fact, under certain circumstances, a magnetic field will exert a force on a charged particle that has no magnetic dipole moment. Here we consider the effect of a magnetic field on such a charged particle.

FACT: A magnetic field exerts no force on a charged particle that is at rest in the magnetic field.



FACT: A magnetic field exerts no force on a charged particle that is moving along the line along which the magnetic field, at the location of the particle, lies.

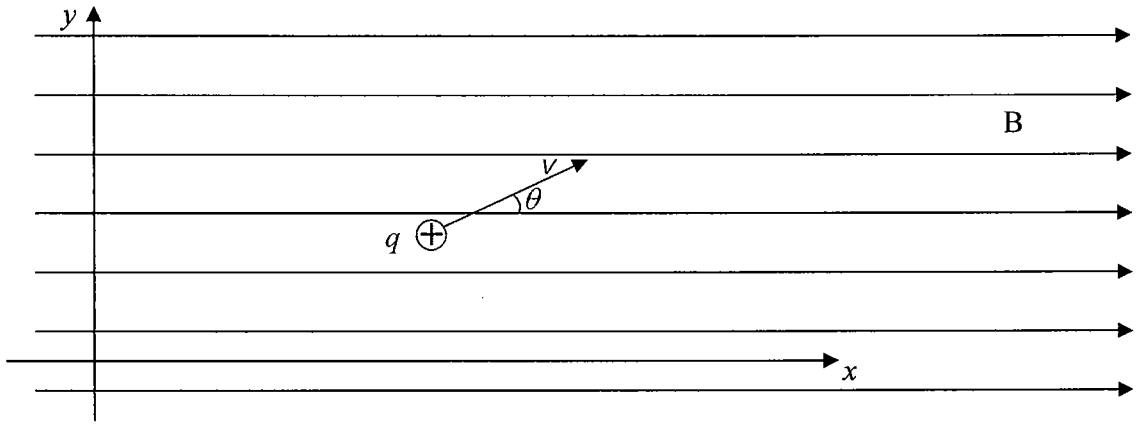


FACT: A magnetic field *does* exert a force on a charged particle that is in the magnetic field, and, is moving, as long as the velocity of the particle is not along the line, along which, the magnetic field is directed. The force in such a case is given by:

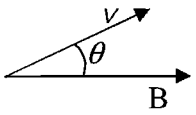
$$\vec{\mathbf{F}} = q\,\vec{\mathbf{v}} \times \vec{\mathbf{B}} \tag{16-1}$$

Note that the cross product yields a vector that is perpendicular to each of the multiplicands. Thus the force exerted on a moving charged particle by the magnetic field within which it finds itself, is always perpendicular to both its own velocity, and the magnetic field vector at the particle's location.

Consider a positively-charged particle moving with velocity v at angle θ in the x-y plane of a Cartesian coordinate system in which there is a uniform magnetic field in the +x direction.



To get the magnitude of the cross product $\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ that appears in $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ we are supposed to establish the angle that $\vec{\mathbf{v}}$ and $\vec{\mathbf{B}}$ make with each other when they are placed tail to tail. Then the magnitude $\left|\vec{\mathbf{v}} \times \vec{\mathbf{B}}\right|$ is just the absolute value of the product of the magnitudes of the vectors times the sine of the angle in between them. Let's put the two vectors tail to tail and establish that angle. Note that the magnetic field as a whole is an infinite set of vectors in the +x direction. So, of course, the magnetic field vector $\vec{\mathbf{B}}$, at the location of the particle, is in the +x direction.



Clearly the angle between the two vectors is just the angle θ that was specified in the problem. Hence,

$$\left|\vec{\mathbf{v}} \times \vec{\mathbf{B}}\right| = \left|vB\sin\theta\right| ,$$

so, starting with our given expression for $\vec{\mathbf{F}}$, we have:

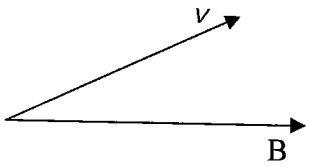
$$\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}\times\vec{\mathbf{B}}$$

$$\left|\vec{\mathbf{F}}\right| = \left|q\,\vec{\mathbf{v}}\times\vec{\mathbf{B}}\right|$$

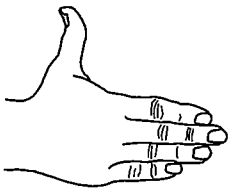
$$\left|\vec{\mathbf{F}}\right| = \left|q\,v\,B\sin\theta\right|$$

Okay, now let's talk about the direction of $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}\times\vec{\mathbf{B}}$. We get the direction of $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$ and then we think. The charge q is a scalar. If q is positive, then, when we multiply the vector $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$ by q (to get $\vec{\mathbf{F}}$), we get a vector in the same direction as that of $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$. So, whatever we get (using the right-hand rule for the cross product) for the direction of $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$ *is* the direction of $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}\times\vec{\mathbf{B}}$. *But*, if q is *negative*, then, when we multiply the vector $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$ by q (to get $\vec{\mathbf{F}}$), we get a vector in opposite direction to that of $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$. So, once we get the direction of $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$ by means of the right-hand rule for the cross product of two vectors, we have to realize that (because the charge is *negative*) the direction of $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}\times\vec{\mathbf{B}}$ is *opposite* the direction that we found for $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$.

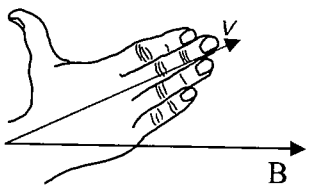
Let's do it. To get the direction of the cross product vector $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$ (which appears in $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}\times\vec{\mathbf{B}}$), draw the vectors $\vec{\mathbf{v}}$ and $\vec{\mathbf{B}}$ tail to tail.



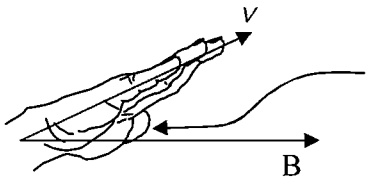
Extend the fingers of your *right* hand so that they are pointing directly away from your right elbow. Extend your thumb so that it is at right angles to your fingers.



Now, keeping your fingers aligned with your forearm, align your fingers with the first vector appearing in the cross product $\vec{\mathbf{v}}\times\vec{\mathbf{B}}$, namely $\vec{\mathbf{v}}$.

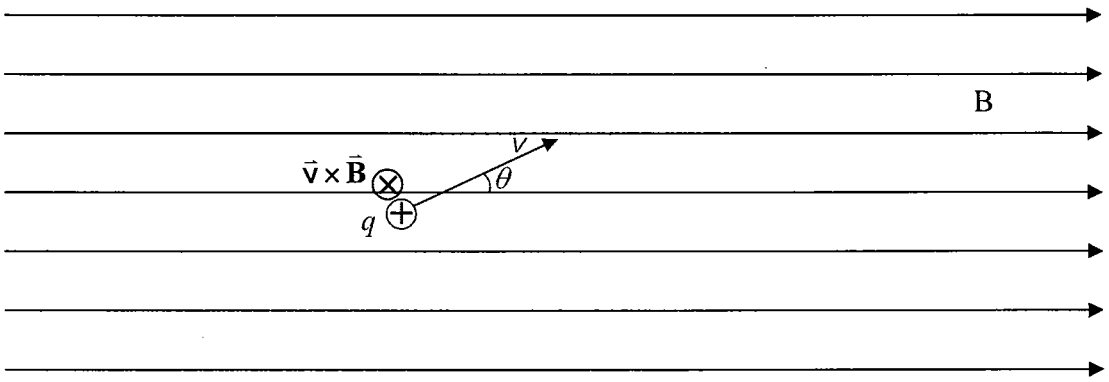


Now rotate your hand, as necessary, about an imaginary axis extending along your forearm and along your middle finger, until your hand is oriented such that, if you were to close your fingers, they would point in the direction of the second vector.

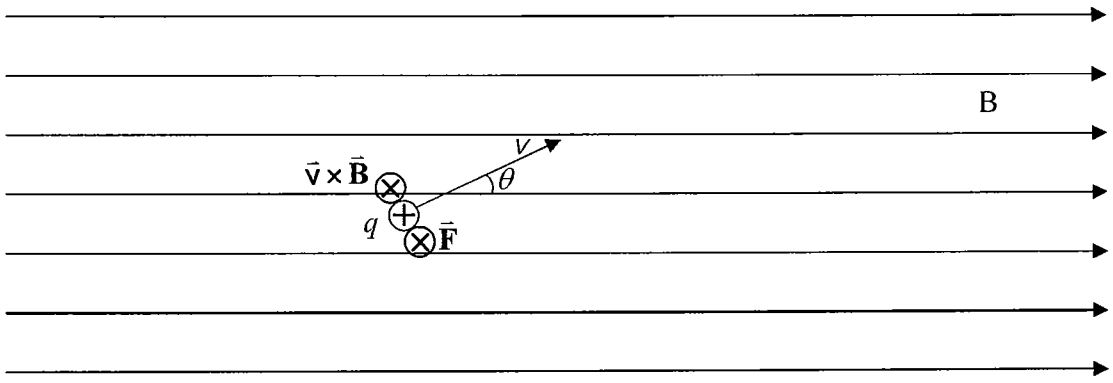


This thumb is pointing straight into the page, directly away from you!

The direction in which your right thumb is now pointing is the direction of $\vec{v}\times\vec{B}$. We depict a vector in that direction by means of an x with a circle around it. That symbol is supposed to represent the tail feathers of an arrow that is pointing away from you.

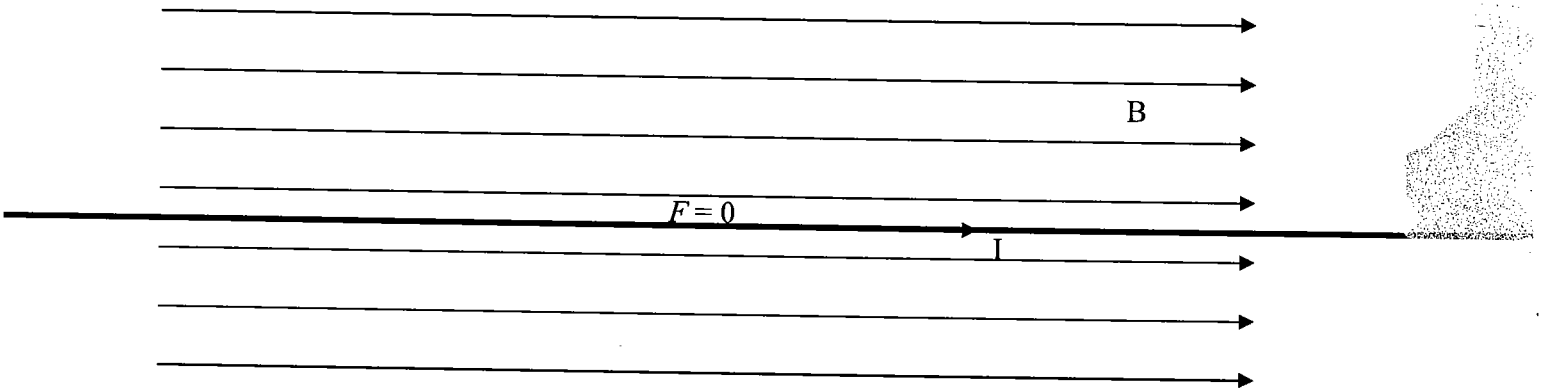


Let's not forget about that q in the expression $\vec{F} = q\vec{v}\times\vec{B}$. In the case at hand, the charged particle under consideration is positive. In other words q is positive. So, $\vec{F} = q\vec{v}\times\vec{B}$ is in the same direction as $\vec{v}\times\vec{B}$.

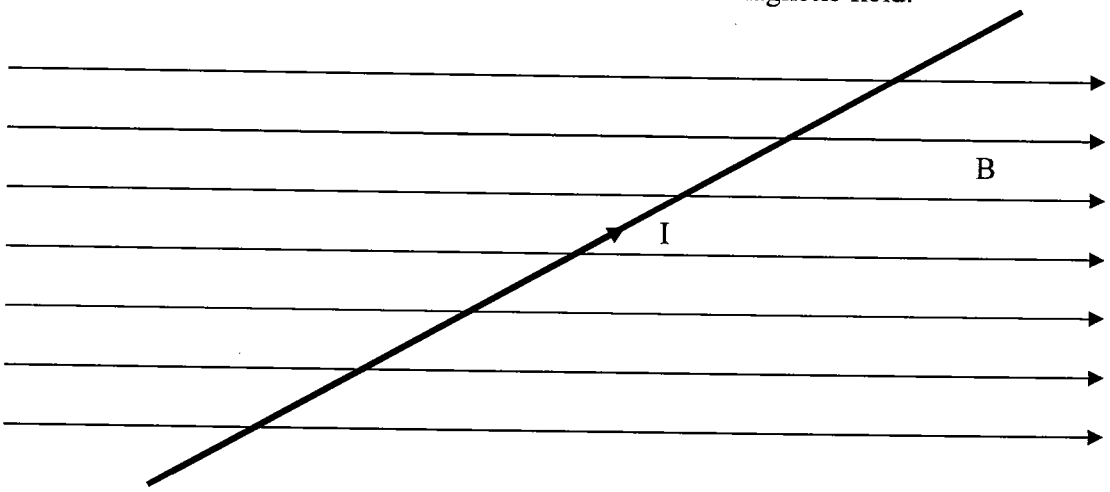


A magnetic field will also interact with a current-carrying conductor. We focus our attention on the case of a straight current-carrying wire segment in a magnetic field:

FACT: Given a straight, current carrying conductor in a magnetic field, the magnetic field exerts no force on the wire segment if the wire segment lies along the line along which the magnetic field is directed. (Note: The circuit used to cause the current in the wire must exist, but, is not shown in the following diagram.)



FACT: A magnetic field exerts a force on a current-carrying wire segment that is in the magnetic field, as long as the wire is not collinear with the magnetic field.

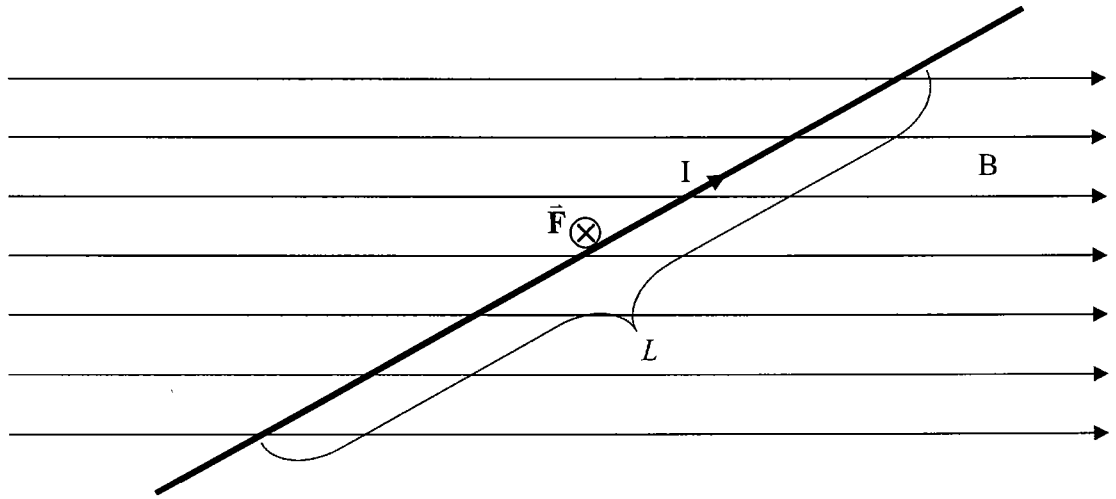


The force exerted on a straight current-carrying wire segment, by the (uniform) magnetic field in which the wire is located, is given by

$$\vec{\mathbf{F}} = I\,\vec{\mathbf{L}} \times \vec{\mathbf{B}} \qquad (16\text{-}2)$$

where:

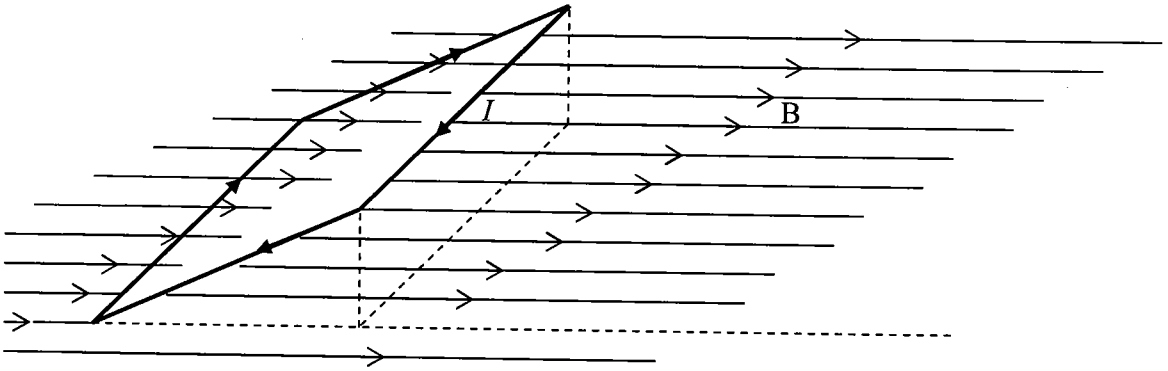
- $\vec{\mathbf{F}}$ is the force exerted on the wire-segment-with-current by the magnetic field the wire is in,
- I is the current in the wire,
- $\vec{\mathbf{L}}$ is a vector whose magnitude is the length of that segment of the wire which is actually *in* the magnetic field, and, whose direction is the direction of the current (which depends both on how the wire segment is oriented and how it is connected in the (not-shown) circuit.)
- $\vec{\mathbf{B}}$ is the magnetic field vector. The magnetic field must be uniform along the entire length of the wire for this formula to apply, so, $\vec{\mathbf{B}}$ is the magnetic field vector at each and every point along the length of the wire.



Note that, in the preceding diagram, $\vec{\mathbf{F}}$ is directed into the page as determined from $\vec{\mathbf{F}} = I\,\vec{\mathbf{L}} \times \vec{\mathbf{B}}$ by means of the right-hand rule for the cross product of two vectors.

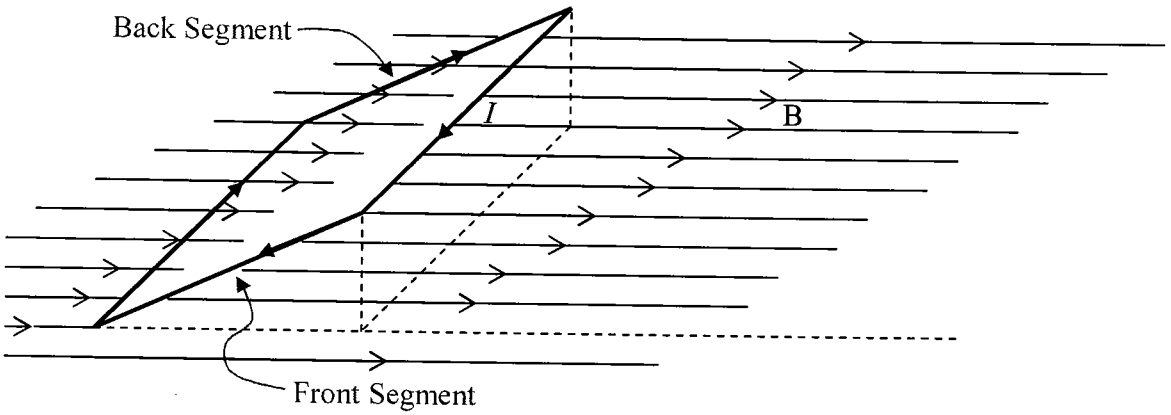
Effect of a Uniform Magnetic Field on a Current Loop

Consider a rectangular loop of wire. Suppose the loop to be in a uniform magnetic field as depicted in the following diagram:

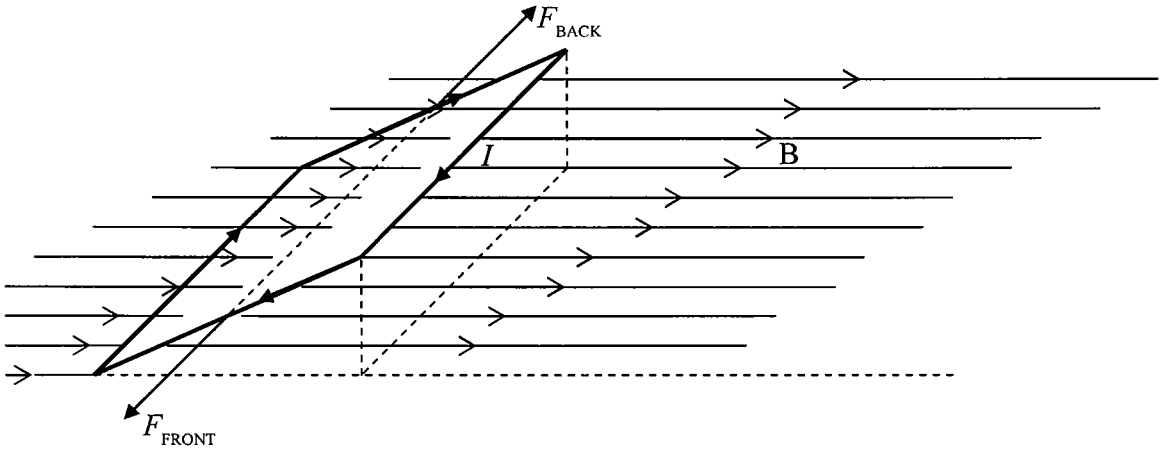


Note that, to keep things simple, we are not showing the circuitry that causes the current in the loop and we are not showing the cause of the magnetic field. Also, the magnetic field exists throughout the region of space in which the loop finds itself. We have not shown the full extent of either the magnetic field lines depicted, or, the magnetic field itself.

Each segment of the loop has a force exerted on it by the magnetic field the loop is in. Let's consider the front and back segments first:

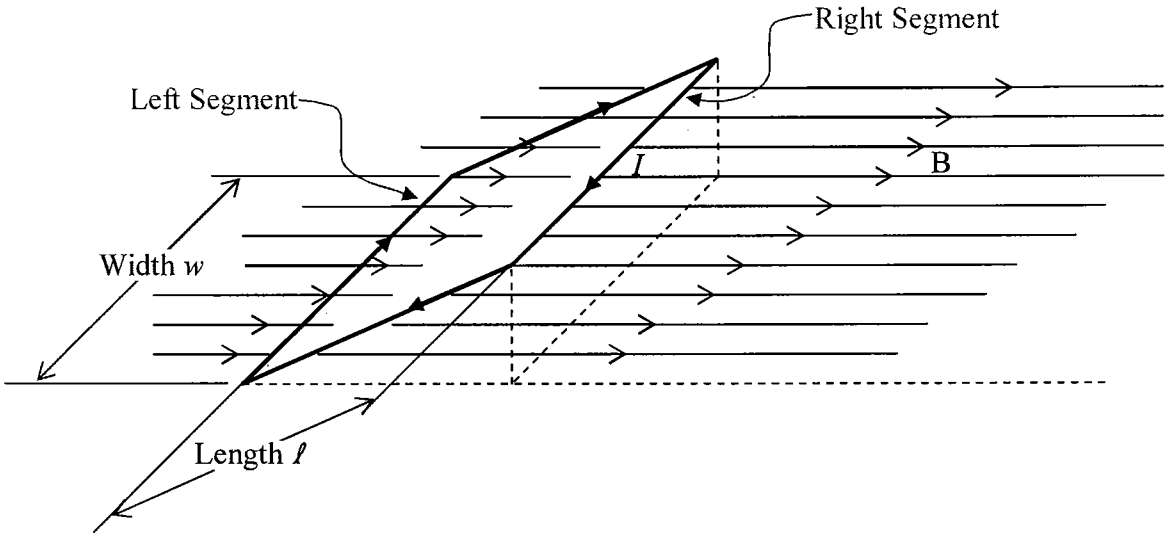


Because both segments have the same length, both segments make the same angle with the same magnetic field, and both segments have the same current; the force $\vec{\mathbf{F}} = I\vec{\mathbf{L}} \times \vec{\mathbf{B}}$ will be of the same magnitude in each. (If you write the magnitude as $F = ILB\sin\theta$, you know the magnitudes are the same as long as you know that for any angle θ, $\sin(\theta) = \sin(180° - \theta)$.) Using the right-hand rule for the cross product to get the direction, we find that each force is directed perpendicular to the segment upon which it acts, and, away from the center of the rectangle:

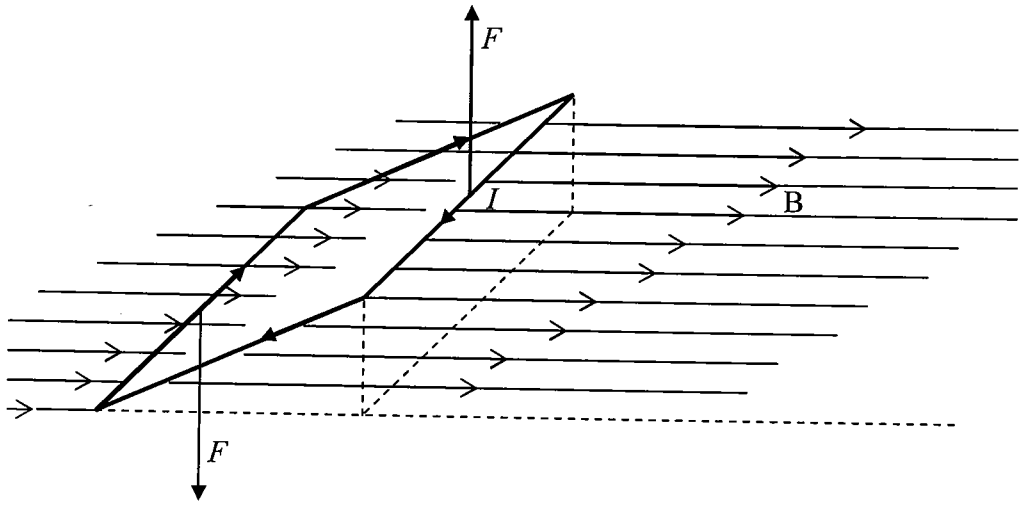


The two forces, F_{FRONT} and F_{BACK} are equal in magnitude, collinear, and opposite in direction. About the only effect they could have would be to stretch the loop. Assuming the material of the loop is rigid enough not to stretch, the net effect of the two forces is no effect at all. So, we can forget about them and focus our attention on the left and right segments in the diagram.

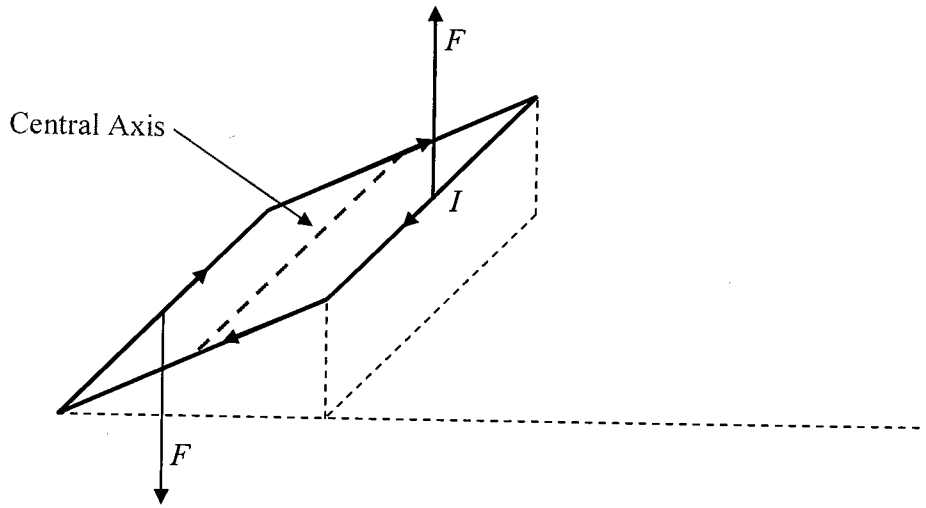
Both the left segment and the right segment are at right angles to the magnetic field. They are also of the same length and carry the same current. For each, the magnitude of $\vec{\mathbf{F}} = I\vec{\mathbf{L}} \times \vec{\mathbf{B}}$ is just IwB where w is the width of the loop and hence the length of both the left segment and the right segment.



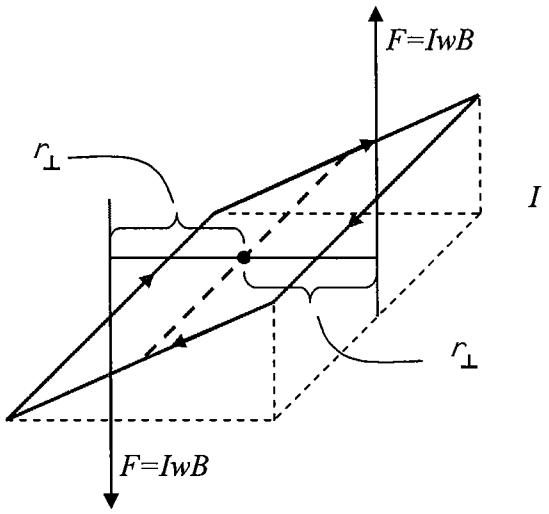
Using the right-hand rule for the cross product of two vectors, applied to the expression $\vec{\mathbf{F}} = I\vec{\mathbf{L}} \times \vec{\mathbf{B}}$ for the force exerted on a wire segment by a magnetic field, we find that the force $F = IwB$ on the right segment is upward and the force $F = IwB$ on the left segment is downward.



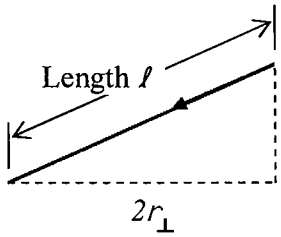
The two forces are equal (both have magnitude $F = IwB$) and opposite in direction, but, they are *not* collinear. As such, they *will* exert a net *torque* on the loop. We can calculate the torque about the central axis:



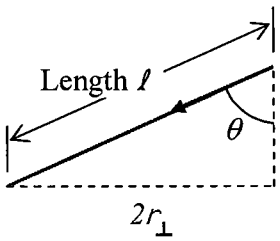
by extending the lines of action of the forces and identifying the moment arms:



The torque provided by each force is $r_\perp F$. Both torques are counterclockwise as viewed in the diagram. Since they are both in the same direction, the magnitude of the sum of the torques is just the sum of the magnitudes of the two torques, meaning that the magnitude of the total torque is just $\tau = 2\, r_\perp F$. We can get an expression for $2r_\perp$ by recognizing, in the diagram, that $2\, r_\perp$ is just the distance across the bottom of the triangle in the front of the diagram:



and defining the angle θ, in the diagram, to be the angle between the plane of the loop and the vertical.



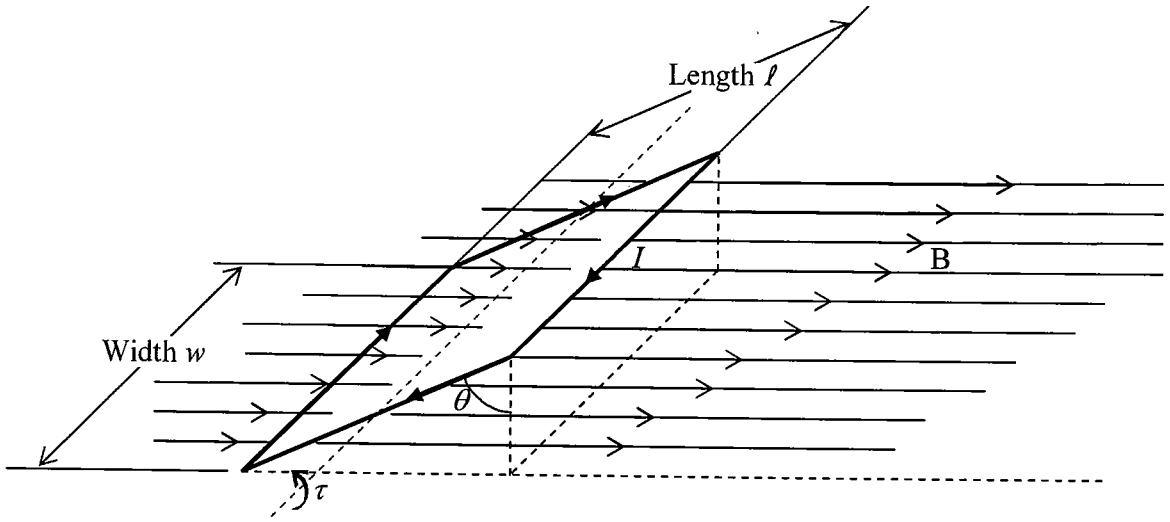
From the diagram, it is clear that $2r_\perp = \ell \sin\theta$.

Thus the magnetic field exerts a torque of magnitude

$$\tau = r_{\perp} F$$

$$\tau = \left[\ell(\sin\theta)\right](IwB)$$

on the current loop.



The expression for the torque can be written more concisely by first reordering the multiplicands so that the expression appears as

$$\tau = I\ell wB\sin\theta$$

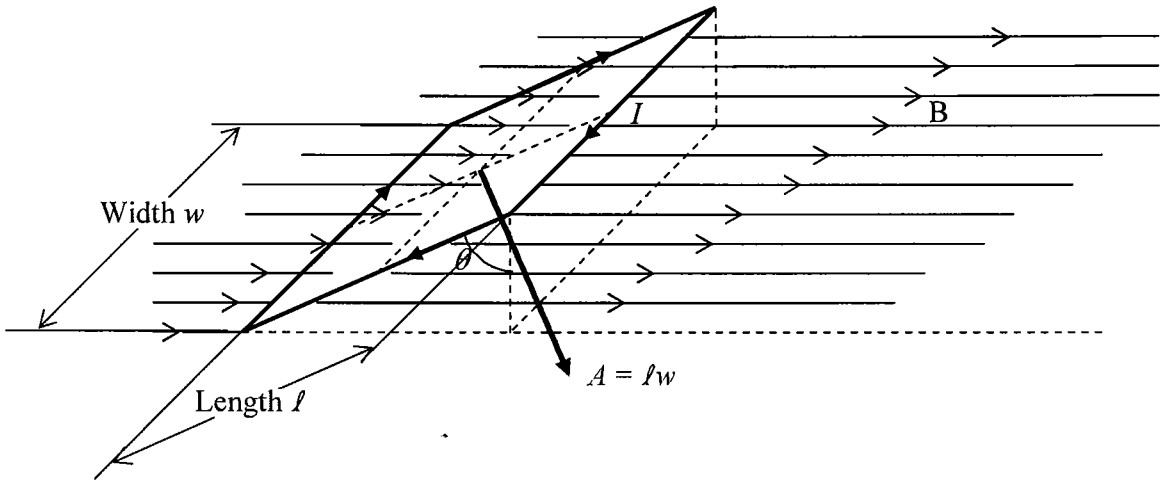
and then recognizing that the product ℓw is just the area A of the loop. Replacing ℓw with A yields:

$$\tau = IAB\sin\theta$$

Torque is something that has direction, and, you might recognize that $\sin\theta$ appearing in the preceding expression as something that can result from a cross product. Indeed, if we define an area vector to have a magnitude equal to the area of the loop,

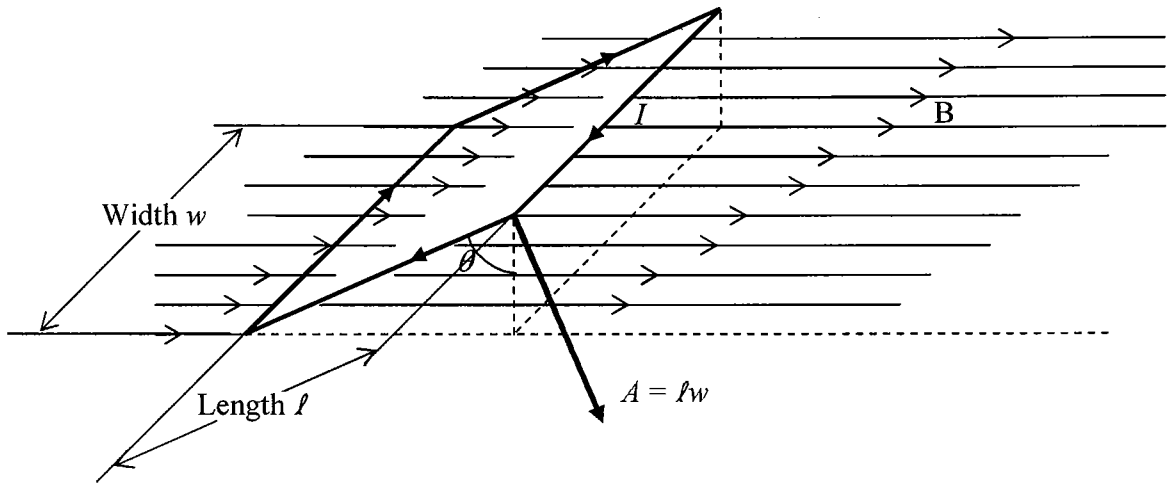
$$\left|\vec{\mathbf{A}}\right| = \ell w$$

and, a direction perpendicular to the plane of the loop,

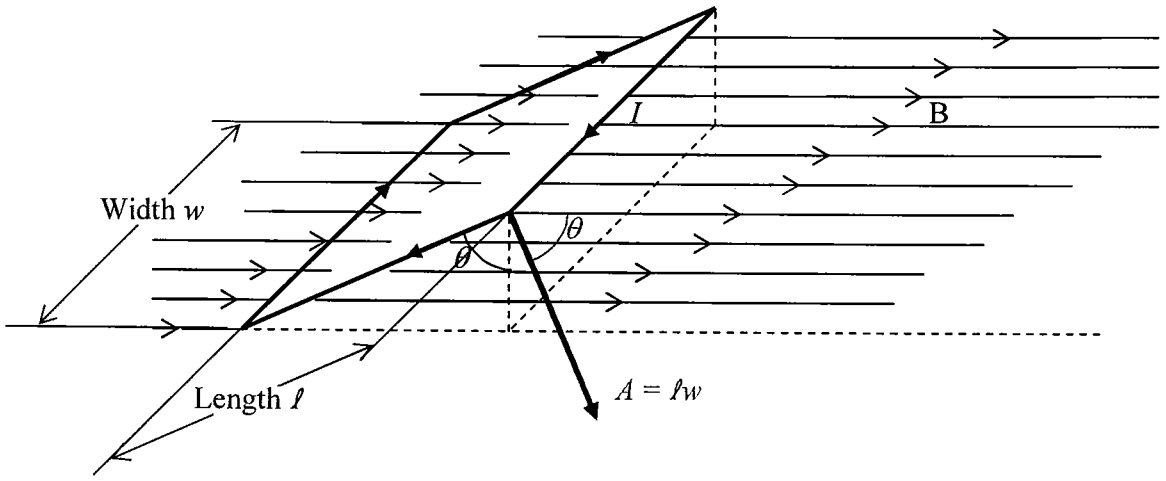


we can write the torque as a cross product. First note that the area vector as I have defined it in words to this point, could point in the exact opposite direction to the one depicted in the diagram. If, however, we additionally stipulate that the area vector is directed in accord with the right-hand rule for something curly something straight, with the loop current being the something curly and the area vector the something straight (and we do so stipulate), then the direction of the area vector is uniquely determined to be the direction depicted in the diagram.

Now, if we slide that area vector over to the right front corner of the loop,



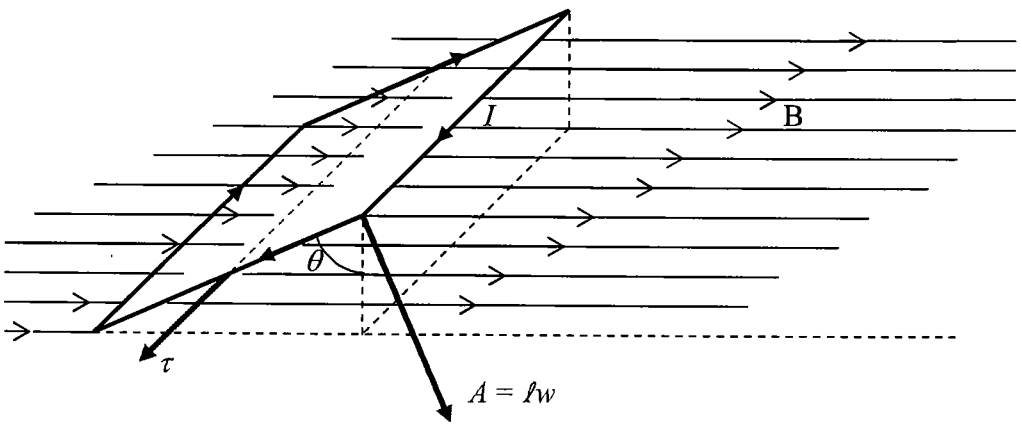
it becomes more evident (you may have already noticed it) that the angle between the area vector $\vec{\mathbf{A}}$ and the magnetic field vector $\vec{\mathbf{B}}$, is the same θ defined earlier and depicted in the diagram just above.



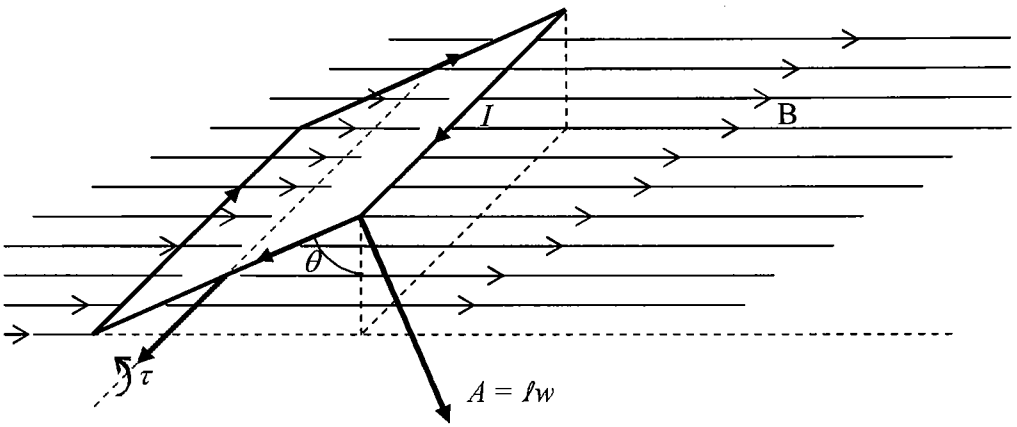
This allows us to write our expression for the torque $\tau = IAB\sin\theta$ counterclockwise as viewed in the diagram, as:

$$\vec{\tau} = I\vec{\mathbf{A}}\times\vec{\mathbf{B}}$$

Check it out. The magnitude of the cross product $\left|\vec{\mathbf{A}}\times\vec{\mathbf{B}}\right|$ is just $AB\sin\theta$, meaning that our new expression yields the same magnitude $\tau = IAB\sin\theta$ for the torque as we had before. Furthermore, the right-hand rule for the cross product of two vectors yields the torque direction depicted in the following diagram.



Recalling that the sense of rotation associated with an axial vector is determined by the right-hand rule for something curly, something straight; we point the thumb of our cupped right hand in the direction of the torque vector and note that our fingers curl around counterclockwise, as viewed in the diagram.



Okay, we're almost there. So far, we have the fact that if you put a loop of wire carrying a current I in it, in a uniform magnetic field $\vec{\mathbf{B}}$, with the loop oriented such that the area vector $\vec{\mathbf{A}}$ of the current loop makes an angle θ with the magnetic field vector, then, the magnetic field exerts a *torque*

$$\vec{\boldsymbol{\tau}} = I\vec{\mathbf{A}} \times \vec{\mathbf{B}}$$

on the loop.

This is identical to what happens to a magnetic dipole when you put it in a uniform magnetic field. It experiences a torque $\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{\mu}} \times \vec{\mathbf{B}}$. In fact, if we identify the product $I\vec{\mathbf{A}}$ as the magnetic dipole moment of the *current loop*, then the expressions for the torque are completely identical:

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{\mu}} \times \vec{\mathbf{B}} \qquad (16\text{-}3)$$

where:

- $\vec{\boldsymbol{\tau}}$ is the torque exerted on the victim. The victim can be either a particle that has an inherent magnetic dipole moment, or, a current loop.

- $\vec{\boldsymbol{\mu}}$ is the magnetic dipole moment of the victim. If the victim is a particle, $\vec{\boldsymbol{\mu}}$ is simply the magnitude and direction of the inherent magnetic dipole moment of the particle. If the victim is a current loop, then $\vec{\boldsymbol{\mu}} = I\vec{\mathbf{A}}$ where I is the current in the loop and $\vec{\mathbf{A}}$ is the area vector of the loop, a vector whose magnitude is the area of the loop and whose direction is the direction in which your right thumb points when you curl the fingers of your right hand around the loop in the direction of the current. (See the discussion below for the case in which the victim is actually a coil of wire rather than a single loop.)

- $\vec{\mathbf{B}}$ is the magnetic field vector at the location of the victim.

A single loop of wire can be thought of as a coil of wire that is wrapped around once. If the wire is wrapped around N times, rather than once, then the coil is said to have N turns or N windings. Each winding makes a contribution of $I\vec{\mathbf{A}}$ to the magnetic dipole moment of the current loop. The contribution from all the loops is in one and the same direction. So, the magnetic moment of a current-carrying *coil* of wire is:

$$\vec{\boldsymbol{\mu}} = NI\vec{\mathbf{A}} \tag{16-4}$$

where:

$\vec{\boldsymbol{\mu}}$ is the magnetic moment of the coil of wire.

N is the number of times the wire was wrapped around to form the coil. N is called the number of windings. N is also known as the number of turns.

I is the current in the coil. The coil consists of one long wire wrapped around many times, so, there is only one current in the wire. We call that one current the current in the coil.

$\vec{\mathbf{A}}$ is the area vector of the loop or coil. Its magnitude is the area of the plane shape whose perimeter is the loop or coil. Its direction is the direction your extended right thumb would point if you curled the fingers of your right hand around the loop in the direction of the current.

Some Generalizations Regarding the Effect of a Uniform Magnetic Field on a Current Loop:

We investigated the effect of a *uniform* magnetic field on a current loop. A magnetic field will exert a torque on a current loop whether or not the magnetic field is uniform. Since a current loop has some spatial extent (it is not a point particle), using a single value-plus-direction for $\vec{\mathbf{B}}$ in $\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{\mu}} \times \vec{\mathbf{B}}$ will yield an approximation to the torque. It is a good approximation as long as the magnetic field is close to being uniform in the region of space occupied by the coil.

We investigated the case of a rectangular loop. The result for the torque exerted on the current-carrying loop or coil is valid for any plane loop or coil, whether it is circular, oval, or rectangular[1].

[1] We have not proved this to be the case. We simply state it here, without proof.

17 Magnetic Field: Causes

This chapter is about magnetism but let's think back to our introduction to charge for a moment. We talked about the electric field before saying much about what caused it. We said the electric field exerts a force on a particle that has charge. Later we found out that charged particles play not only the role of "victim" to the electric field but, that charged particles *cause* electric fields to exist.

Now we have been talking about the magnetic field. We have said that the magnetic field exerts a torque on a particle that has magnetic dipole moment. You might guess that a particle that has magnetic dipole moment would cause a magnetic field. You'd be right! A particle that has the physical property known as *magnetic dipole moment* causes a magnetic field to exist in the region of space around it. A magnetic field can be caused to exist by a particle having magnetic dipole moment or a distribution of particles having magnetic dipole moment.

The magnetic field at point P, an empty point in space in the vicinity of a particle that has a magnetic dipole moment, due to that particle-with-magnetic-dipole-moment, is given by

$$\vec{\mathbf{B}} = \frac{\mu_o}{4\pi} \frac{3(\vec{\boldsymbol{\mu}} \cdot \hat{r})\hat{r} - \vec{\boldsymbol{\mu}}}{r^3} \qquad (17\text{-}1)$$

where

$\mu_o = 4\pi \times 10^{-7} \frac{\mathrm{T \cdot m}}{\mathrm{A}}$ is a universal constant which goes by the name of "the magnetic permeability of free space." This value is to be taken as exact. (Do not treat the "4" as a value known to only one significant digit.)

$\vec{\mathbf{B}}$ is the magnetic field vector at point P, where P is an empty point in space a distance r away from the particle-with-magnetic-dipole-moment that is causing $\vec{\mathbf{B}}$.

$\vec{\boldsymbol{\mu}}$ is the magnetic dipole moment of the particle that is causing the magnetic field.

$\hat{r}$ is a unit vector in the direction "from the particle, toward point P". Defining $\vec{r}$ to be the position vector of point P relative to the location of the particle-with-magnetic-dipole-moment, $\vec{r} = r\hat{r}$ so $\hat{r} = \frac{\vec{r}}{r}$.

r is the distance that point P is from the particle-with-magnetic-dipole-moment.

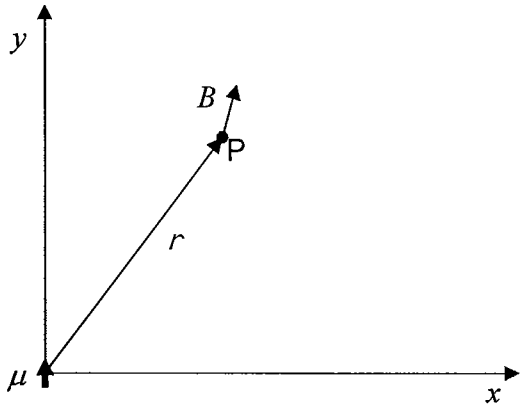
A particle-with-magnetic-dipole-moment is called a magnetic dipole. Note that the magnetic field due to a magnetic dipole dies off like $\frac{1}{r^3}$.

Example

A particle is at the origin of a Cartesian coordinate system. The magnetic dipole moment of the particle is 1.0 A·m^2 $\hat{\jmath}$. Find the magnetic field vector, due to the particle, at (3.0 cm, 4.0 cm).

Solution

I'm going to start with a diagram of the configuration.



Note that I do *not* know the direction of B in advance, so, I have drawn $\vec{\mathbf{B}}$ on the diagram in a fairly arbitrary direction. I did want to put $\vec{\mathbf{B}}$ on there to make it more evident that we are dealing with the magnetic field at point P, caused by the particle at the origin. Also, I intentionally drew $\vec{\mathbf{B}}$ in a direction other than that of $\vec{r}$, to avoid conveying the false impression that $\vec{\mathbf{B}}$ is necessarily in the direction of $\vec{r}$. (At some points, it is, but those points are the exception. In general, $\vec{\mathbf{B}}$ is *not* in the same direction as $\vec{r}$. As we shall soon see, for the case at hand, it turns out that $\vec{\mathbf{B}}$ is *not* in the same direction as $\vec{r}$.)

Given x = 0.030 m and y =0.040 m, the position vector, for point P is $\vec{r} = 0.030\text{ m}\hat{\imath} + 0.040\text{ m}\hat{\jmath}$. The magnitude of $\vec{r}$ is given by:

$$r = \sqrt{x^2 + y^2}$$
$$r = \sqrt{(.030\,\text{m})^2 + (.040\,\text{m})^2}$$
$$r = .050\text{ m}$$

The unit vector $\hat{r}$ is thus given by:

$$\hat{r} = \frac{\vec{r}}{r}$$

$$\hat{r} = \frac{0.030\ \text{m}\uparrow + 0.040\ \text{m}\hat{\jmath}}{0.050\ \text{m}}$$

$$\hat{r} = 0.60\uparrow + 0.80\,\hat{\jmath}$$

Substituting what we have into our expression for $\vec{B}$ we find:

$$\vec{B} = \frac{\mu_o}{4\pi}\frac{3(\vec{\mu}\cdot\hat{r})\hat{r} - \vec{\mu}}{r^3}$$

$$\vec{B} = \frac{4\pi\times10^{-7}\,\text{T·m/A}}{4\pi}\ \frac{3\left[(1.0\ \text{A·m}^2\hat{\jmath})\cdot(0.60\uparrow + 0.80\,\hat{\jmath})\right](0.60\uparrow + 0.80\,\hat{\jmath}) - 1.0\ \text{A·m}^2\hat{\jmath}}{(.050\ \text{m})^3}$$

$$\vec{B} = 1\times10^{-7}\,\frac{\text{T·m}}{\text{A}}\ \frac{3\left[0.80\ \text{A·m}^2\right](0.60\uparrow + 0.80\,\hat{\jmath}) - 1.0\ \text{A·m}^2\hat{\jmath}}{(.050\ \text{m})^3}$$

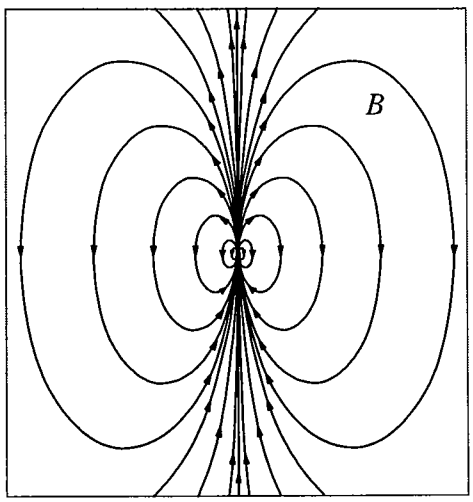
$$\vec{B} = 1\times10^{-7}\,\frac{\text{T·m}}{\text{A}}\ \frac{1.44\ \text{A·m}^2\uparrow + 1.92\ \text{A·m}^2\,\hat{\jmath} - 1.0\ \text{A·m}^2\hat{\jmath}}{(.050\ \text{m})^3}$$

$$\vec{B} = 1\times10^{-7}\,\frac{\text{T·m}}{\text{A}}\ \frac{1.44\ \text{A·m}^2\uparrow + 0.92\ \text{A·m}^2\,\hat{\jmath}}{(.050\ \text{m})^3}$$

$$\vec{B} = 1\times10^{-7}\,\frac{\text{T·m}}{\text{A}}\,(11520\frac{\text{A}}{\text{m}}\uparrow + 7360\frac{\text{A}}{\text{m}}\,\hat{\jmath})$$

$$\vec{B} = 1.152\ \text{mT}\uparrow + .736\ \text{mT}\,\hat{\jmath}$$

So ends our solution to the sample problem. Here's a magnetic field diagram of the magnetic field due to a particle that has a magnetic dipole moment.



The Magnetic Field Due to a Loop or Coil

We discovered in the last chapter that, as a victim to a magnetic field, a current loop or a current-carrying coil behaves as if it were a particle with a magnetic dipole moment

$$\vec{\mu} = NI\vec{A}$$

where:

- $\vec{\mu}$ is the magnetic moment of the coil of wire.
- N is the number of windings, a.k.a. the number of turns. ($N = 1$ in the case of a loop.)
- I is the current in the coil.
- $\vec{A}$ is the area vector of the loop or coil. Its magnitude is the area of the plane shape whose perimeter is the loop or coil. Its direction is the direction your extended right thumb would point if you curled the fingers of your right hand around the loop in the direction of the current.

You might guess that if a coil of wire responds to a magnetic field as if it were a particle with a magnetic dipole moment, then perhaps it can also behave as a source of magnetic field lines and create the same kind of magnetic field that a particle with a magnetic dipole moment produces. Indeed it does. As compared to a particle like the electron that has a magnetic dipole moment but itself has no extent in space, a loop or coil of wire does have extent in space. The magnetic field very near the loop or coil is more complicated than a dipole field, but, at points whose distance from the loop or coil are large compared to the diameter of the coil, the magnetic field of the loop or coil is the dipole magnetic field

$$\vec{B} = \frac{\mu_o}{4\pi}\frac{3(\vec{\mu}\cdot\hat{r})\hat{r} - \vec{\mu}}{r^3}$$

In the case of a loop or coil, the $\vec{\mu}$ that appears in this equation is $\vec{\mu} = NI\vec{A}$.

A Bar Magnet

An atom is made of a nucleus containing neutrons and protons; and; electrons in orbit *about* the nucleus. Each of these elementary particles has a magnetic moment. The magnetic moment[1] of the electron is $9.28\times10^{-24}\,\text{A}\cdot\text{m}^2$, the magnetic moment of the proton is $1.41\times10^{-26}\,\text{A}\cdot\text{m}^2$, and, the magnetic moment of the neutron is $9.66\times10^{-27}\,\text{A}\cdot\text{m}^2$. When these particles combine to form atoms, they each contribute to the magnetic field of the atom. In addition to these contributions to the magnetic field, the protons move in loops within the nucleus and the electrons move in loops about the nucleus. A charged particle that is moving in a loop is a current loop and such current loops contribute to the overall magnetic field of the atom. In many atoms the various contributions to the magnetic field cancel each other out in such a manner that the overall

[1] I got the magnetic moment values from the U.S. National Institute of Standards and Technology (NIST) www.nist.gov web site and rounded them to three significant figures.

magnetic field is essentially zero. In some atoms, such as iron, cobalt, and neodymium, the various contributions to the magnetic field do not cancel out. In such cases, the observed total magnetic field of the atom is a dipole magnetic field, and, the atom behaves as a magnetic dipole. Substances consisting of such atoms are referred to as ferromagnetic materials.

Consider an iron rod or bar that is not a magnet. The bar was formed from molten iron. As the iron cooled, seed crystals formed at various locations within the iron. At the start of crystallization, the iron atoms forming the seed crystal tend to align with each other, south pole to north pole. The magnetic field of the seed crystal causes neighboring iron atoms to align with the seed crystal magnetic dipole moment so that when they crystallize and become part of the growing crystal they also align south pole to north pole. The contributions of the atoms making up the crystal to the magnetic field of the crystal tend to add together constructively to form a relatively large magnetic field. There is a multitude of sites at which crystals begin to form and at each site, in the absence of an external magnetic field, the seed crystal is aligned in a random direction. As the crystals grow, they collectively form a multitude of microscopic bar magnets. When the iron bar is completely solidified it consists of a multitude of microscopic bar magnets called *domains*. Because they are aligned in random directions, their magnetic fields cancel each other out. Put the iron rod or bar in a magnetic field and the magnetic field will cause the microscopic bar magnets, the domains, in the iron to line up with each other to an extent that depends on the strength of the magnetic field. This turns the iron rod or bar into a magnet. Remove the rod or bar from the magnetic field and local forces on the domains cause them to revert back toward their original orientations. They do not achieve their original orientations and the iron remains at least weakly magnetized, an effect known as hysteresis.

Getting back to the cooling process, if we allow the molten iron to crystallize within an external magnetic field, the seed crystals, will all tend to line up with the external magnetic field, and hence, with each other. When the iron is completely solidified, you have a permanent magnet.

So a bar magnet consists of a bunch of microscopic bar magnets which themselves consist of a bunch of atoms each of which has a magnetic dipole moment because it consists of particles that each have a magnetic dipole moment and in some cases have charge and move in a loop within the atom.

The magnetic field of a bar magnet is thus the superposition (vector sum at each point in space) of a whole lot of magnetic dipole fields. As such, at distances large compared to the length of the magnet, the magnetic field of a bar magnet is a magnetic dipole field. As such, we can assign, based on measurements, a magnetic dipole vector $\vec{\mu}$ to the bar magnet as a whole, and compute its magnetic field (valid for distances large compared to the length of the magnet) as

$$\vec{\mathbf{B}} = \frac{\mu_o}{4\pi} \frac{3(\vec{\mu} \cdot \hat{r})\hat{r} - \vec{\mu}}{r^3}$$

The Dipole-Dipole Force

The magnetic field produced by one bar magnet will exert a torque on another bar magnet. Because the magnetic field due to a magnetic dipole is non-uniform (you can see in $\vec{\mathbf{B}} = \frac{\mu_o}{4\pi}\frac{3(\vec{\boldsymbol{\mu}}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \vec{\boldsymbol{\mu}}}{r^3}$ that it dies off like $\frac{1}{r^3}$), it also exerts a *force* on another bar magnet.

We are now in a position to say something quantitative about the force that one bar magnet exerts on another. Consider an object that is at the origin of a Cartesian coordinate system. Suppose that object to have a magnetic dipole moment given by $\vec{\boldsymbol{\mu}}_1 = \mu_1\hat{\mathbf{\imath}}$. Clearly we're talking about a magnet pointing (treating the magnet as an arrow with its head at the north pole of the magnet) in the +x direction. Let's find the force that *that* magnet would exert on another one at (x, 0, 0) given that the magnetic dipole moment of the second magnet is $\vec{\boldsymbol{\mu}}_2 = -\mu_2\hat{\mathbf{\imath}}$. The second magnet is pointing back toward the origin, so we are talking about two magnets whose north poles are facing each other. Knowing that like poles repel, you should be able to anticipate that the second magnet will experience a force in the +x direction. The magnetic field produced by the first magnet is given (for any point in space, as long as the distance to that point, from the origin, is large compared to the size of the magnet) by

$$\vec{\mathbf{B}} = \frac{\mu_o}{4\pi}\frac{3(\vec{\boldsymbol{\mu}}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \vec{\boldsymbol{\mu}}}{r^3}$$

$$\vec{\mathbf{B}}_1 = \frac{\mu_o}{4\pi}\frac{3(\mu_1\hat{\mathbf{\imath}}\cdot\hat{\mathbf{r}})\,\hat{\mathbf{r}} - \mu_1\hat{\mathbf{\imath}}}{r^3}$$

$$\vec{\mathbf{B}}_1 = \frac{\mu_o}{4\pi}\left(\frac{3(\mu_1\hat{\mathbf{\imath}}\cdot\vec{\mathbf{r}})\vec{\mathbf{r}}}{r^5} - \frac{\mu_1\hat{\mathbf{\imath}}}{r^3}\right)$$

The force on the second particle is given by:

$$\vec{\mathbf{F}} = \nabla(\vec{\boldsymbol{\mu}}_2\cdot\vec{\mathbf{B}}_1)$$

evaluated at the position of magnet 2, namely at (x, 0, 0). Substituting the given $\vec{\boldsymbol{\mu}}_2 = -\mu_2\hat{\mathbf{\imath}}$ in for the magnetic dipole of particle 2, and, the expression just above for $\vec{\mathbf{B}}_1$, we obtain:

$$\vec{\mathbf{F}} = \nabla\left\{-\mu_2\hat{\mathbf{\imath}}\cdot\left[\frac{\mu_o}{4\pi}\left(\frac{3(\mu_1\hat{\mathbf{\imath}}\cdot\vec{\mathbf{r}})\vec{\mathbf{r}}}{r^5} - \frac{\mu_1\hat{\mathbf{\imath}}}{r^3}\right)\right]\right\}$$

$$\vec{\mathbf{F}} = -\frac{\mu_o}{4\pi}\mu_1\mu_2\nabla\left\{\hat{\mathbf{\imath}}\cdot\left[\left(\frac{3(\hat{\mathbf{\imath}}\cdot\vec{\mathbf{r}})\vec{\mathbf{r}}}{r^5} - \frac{\hat{\mathbf{\imath}}}{r^3}\right)\right]\right\}$$

Now, if you substitute $\vec{r} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$ and $r = \sqrt{x^2 + y^2 + z^2}$, take the gradient, and then (*after* taking the gradient) evaluate the result at (*x*, 0, 0), you find that

$$\vec{\mathbf{F}} = \frac{3\mu_o}{2\pi}\frac{\mu_1\mu_2}{x^4}\hat{\mathbf{i}}$$

So, when like poles are facing each other, two magnets repel each other with a force that dies off like $\frac{1}{r^4}$ where r is the distance between the magnets (measure it center to center), the x in the case that we investigated.

The Magnetic Field Due to a Long Straight Current-Carrying Wire

A current-carrying conductor causes a magnetic field. You are already aware of this because we have already discussed the fact that a current in a loop or coil behaves as a magnetic dipole, and, you know that a magnetic dipole creates a magnetic field in the region of space around it. As it turns out, a wire with a current in it doesn't have to be wrapped around in the shape of a loop or coil to produce a magnetic field. In fact, experimentally, we find that a straight wire segment creates a magnetic field in the region of space around it. The magnitude of the magnetic field due to a long straight wire, valid for any point whose distance from the wire is small compared to the length of the wire and whose distance from either end of the straight wire segment in question is large compared to the distance from the wire, is given by

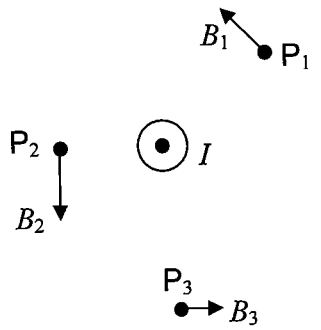
$$B = \frac{\mu_o}{2\pi}\frac{I}{r} \qquad (17\text{-}2)$$

where

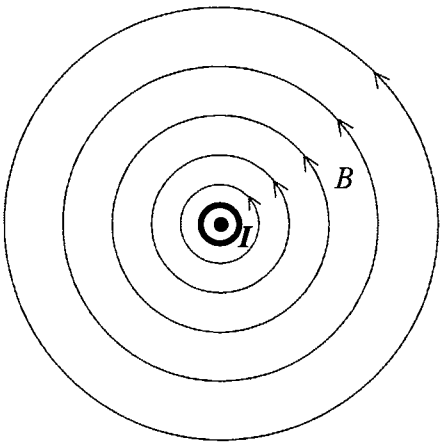
- μ_0 is a constant referred to as the magnetic permeability of free space,
- I is the current in the wire segment, and,
- r is the distance that the point in question is from the long straight wire segment. The equation gives the magnitude of the magnetic field at any specified point P. The symbol r represents the distance that point P is from the wire.

The direction of the magnetic field due to a long, straight, current-carrying wire, at some empty point in space, call it point P, is always perpendicular to both the wire and the imaginary line segment that extends from point P, straight to (and thus perpendicular to), the current-carrying wire.

Consider the case of a long straight wire carrying current straight at you. The magnetic field at a few points is depicted in the diagram below (where the empty points in space in question are labeled P_1, P_2, and P_3.)



While the magnetic field vector at any point in space is, of course, directed along a straight line, the overall pattern of the magnetic field lines in the vicinity of a long straight wire segment, in a plane perpendicular to the wire segment, forms circles around the wire. The magnetic field lines are directed tangent to the circles, and, the direction is given by the right hand rule for something curly something straight. The magnetic field line pattern is the something curly and the current in the straight wire is the something straight. Point your right thumb in the direction of the current and the fingers of your cupped right hand curl around in the direction of the magnetic field lines.

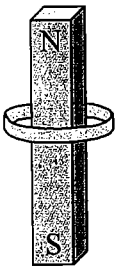


18 Faraday's Law, Lenz's Law

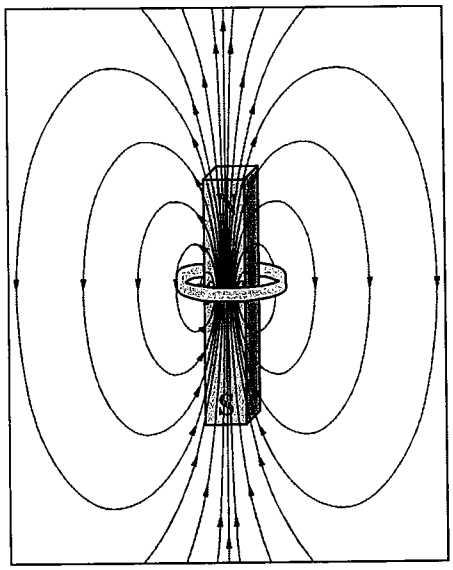
Do you remember Archimedes's Principle? We were able to say something simple, specific, and useful about a complicated phenomenon. The gross idea was that a submerged object being pressed upon on every surface element in contact with fluid, by the fluid, experiences a net upward force because the pressure in a fluid is greater at greater depth. The infinite sum, over all the surface area elements of the object in contact with the fluid, of the force of magnitude pressure-times-the-area, and direction normal to and into the area element, resulted in an upward force that we called the buoyant force. The thing is, we were able to prove that the buoyant force is equal in magnitude to the weight of that amount of fluid that would be where the object is if the object wasn't there. Thus we can arrive at a value for the buoyant force without having to even think about the vector integration of pressure-related force that causes it.

We are about to encounter another complicated phenomenon which can be characterized in a fruitful fashion by a relatively simple rule. I'm going to convey the idea to you by means of a few specific processes, and then sum it up by stating the simple rule.

Consider a gold[1] ring and a bar magnet in the hands of a person. The person is holding the ring so that it encircles the bar magnet. She is holding the magnet, north end up.



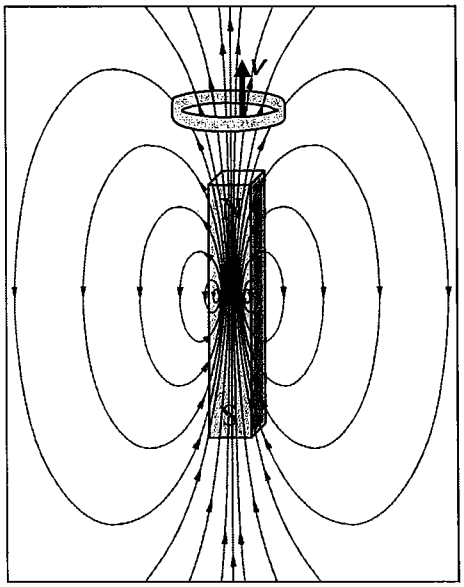
There is a magnetic field, due to the bar magnet, within the bar magnet, and in the region of space around it.



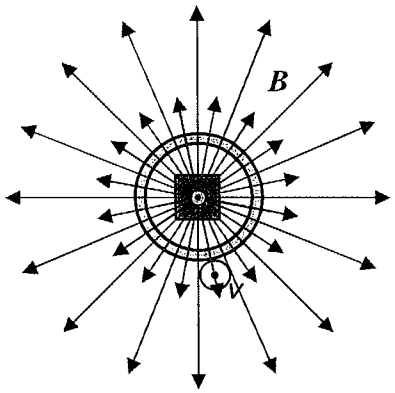
[1] Any conductive material will do here. I chose gold arbitrarily.

It is important to note that the magnetic field lines are most densely packed inside the bar magnet.

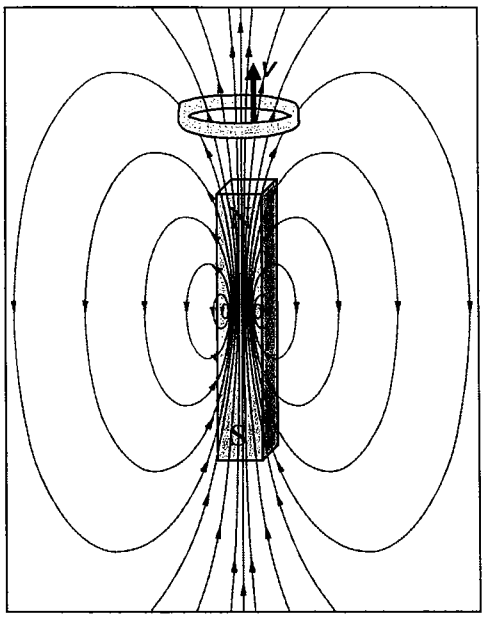
Now suppose that the person, holding the magnet at rest in one hand, moves the loop upward. I want to focus on what is going on while she is moving it upward. As she moves the loop upward, she is moving it roughly along the direction of the magnetic file lines, but, and this is actually the important part, that loop will also be "cutting through" some magnetic field lines. Consider an instant in time when the loop is above the magnet, and moving upward:



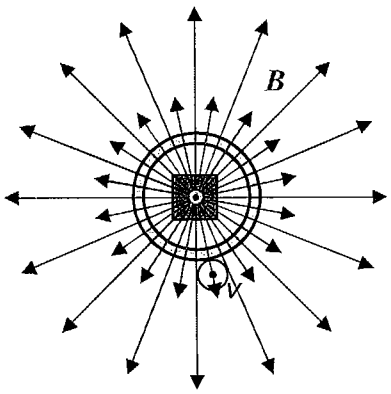
From above, the scene looks like:



where it is important to realize that none of those magnetic field lines begin on the magnet or end at the tip of the arrow depicted, rather, they extend out of the magnet toward us, flower out and over, back down away from us, and then they loop around to enter the south pole of the magnet from which they extend back up through the magnet toward us. In fact, no magnetic field line ever begins or ends anywhere. They all form closed loops. This is a manifestation of the fact that there is no such thing as magnetic charge. (There are no magnetic monopoles.)



View From Above



Here's where we're going with this: The motion of the ring relative to the magnet is going to cause a current in the ring. Here's how: The ring is neutral, but, it is chock full of charged particles that are free to move around within the gold. [I'm going to discuss it in our positive charge carrier model but you can verify that you get the same result if the charge carriers are negative (recalling that our current is in the direction opposite that in which negative charge carriers are moving.)] Pick any short segment of the ring and get the direction of the force exerted on the charge carriers of that segment using $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ and the right-hand rule for the cross product of two vectors. In the view from above, all we can see is the horizontal component of the magnetic field vectors in the vicinity of the moving ring but that's just dandy; the vertical component, being parallel to the ring's velocity (and hence parallel to the velocity of the charge in the ring), makes no contribution to $\vec{\mathbf{v}} \times \vec{\mathbf{B}}$. Now, pick your segment of the ring. Make your fingers point away from your elbow, and, in the direction of the first vector (the velocity vector) in $\vec{\mathbf{v}} \times \vec{\mathbf{B}}$, namely, "out of the page". Now, keeping your fingers pointing both away from your elbow, and, out of the page, rotate your forearm as necessary so that your palm is facing in the direction of $\vec{\mathbf{B}}$ (at the location of the segment you are working on), meaning that if you were to close your fingers, they would point in the direction of $\vec{\mathbf{B}}$. Your extended thumb is now pointing in the direction of the force exerted on the positive charge carriers in the ring segment you chose. No matter what ring segment you pick, the force is always in that direction which tends to push the positive charge carriers counterclockwise around the ring! The result is a counterclockwise (as viewed from above) current in the ring.

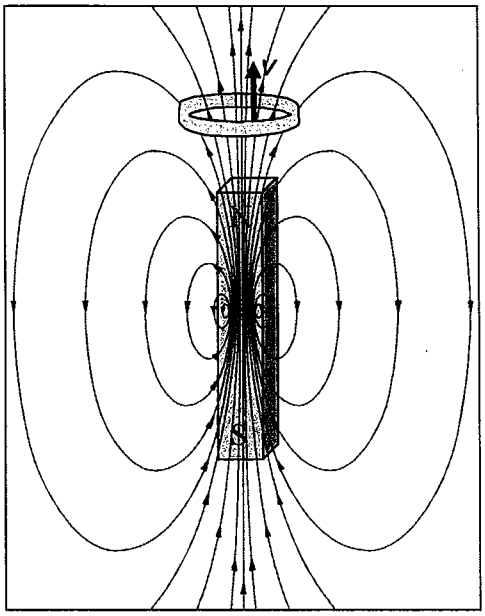
Suppose that, starting with the ring encircling the magnet, the person who was holding the ring and the magnet moved the magnet downward rather than moving the ring upward. She holds the ring stationary, and moves the magnet. I said earlier that a charged particle at rest in a magnetic field has no force exerted on it by the magnetic field. But we were talking about *stationary* magnetic fields at the time. Now we are talking about the magnetic field of a magnet that is moving. Since the magnet responsible for it is moving, the magnetic field itself must be moving. Will that result in a force on the charges in the ring (and hence a current in the ring)? This brings us to a consideration of relative motion. To us, the two processes (person moves ring upward

while holding magnet at rest, vs. person moves magnet downward while holding ring at rest) are different. But that is just because we are so used to viewing things from the earth's reference frame. Have you ever been riding along a highway and had the sense that you were at rest and the lampposts on the side of the road were moving past you at high speed. That is a valid viewpoint. Relative to your reference frame of the car, the lampposts are indeed moving and the car is a valid reference frame. Suppose we view the *magnet moving downward through a ring* situation from a platform that is moving downward at the same speed as the magnet. In that reference frame, the magnet is at rest. If for instance, as we, while seated on the platform, see the magnet at eye level, it remains at eye level. But the ring, as viewed from the platform reference frame is moving upward. So in the platform reference frame, we have, in the new processes (which in the room reference frame is a magnet moving downward through and away from a ring) the same situation that we had in the room frame for the original process (which in the room reference frame is a ring, originally encircling a stationary magnet, moving upward). Thus in the platform reference frame, we must have the same result for the new process that we had for the original process in the room frame, namely, a counterclockwise (as viewed from above) current in the ring. The current in the ring doesn't depend on what reference frame we view the ring from. Hence, we can conclude that the magnet moving downward through the stationary ring at speed v results in the same current as we have when the ring moves upward at the same speed v relative to the stationary magnet.

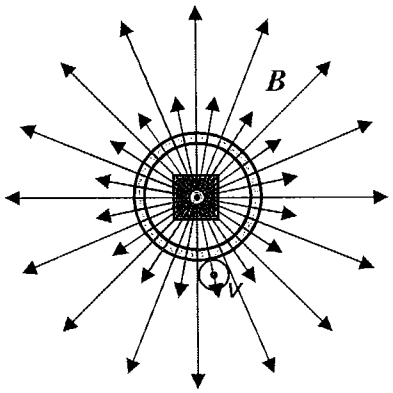
When the person holding the magnet and the ring moved the ring upward, there was a current in the ring. Now we have established that if, instead of moving the ring, she moves the magnet downward at the same speed, she will get the same current in the ring. Based on what caused that current, the $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ force on the charged particles in the ring, you can surmise that the current will depend on things like the velocity of the ring relative to the magnet, the strength of the magnetic field, and the relative orientation of the velocity vector and the magnetic field. It has probably occurred to you that the current also depends on the resistance of the ring.

Michael Faraday came up with a very fruitful way of looking at the phenomenon we are discussing and I will convey his idea to you by means of the example we have been working with.

Looking at the diagrams of that ring moving relative to the magnet again,



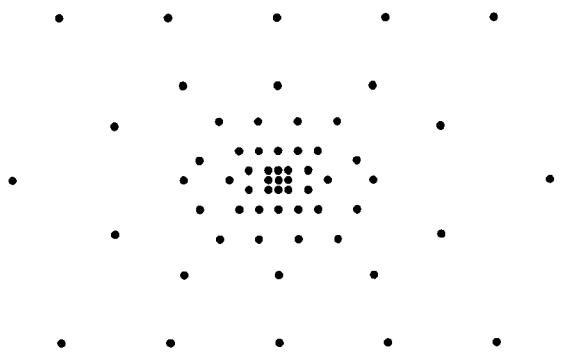
View From Above



we can describe what's happening by saying that the ring is "cutting through" magnetic field lines (or, equivalently, by saying that the magnetic field lines are "cutting through" the ring). What Faraday recognized was that, in conceptual terms, by the ring cutting through magnetic field lines (or vice versa depending on your point of view), what was happening was, that the number of magnetic field lines encircled by the loop was changing. In the diagrams above, each time the ring "cuts through" one more field line, the number of field lines encircled by the loop decreases by one. The rate at which the ring "cuts through" magnetic field lines (or the magnetic field lines cut through the ring) is determined by the same things that determine the force on the charged particles making up the ring (relative speed between ring and magnetic field, strength of magnetic field, relative orientation of velocity of ring and magnetic field) such that, the greater the rate at which the ring "cuts through" magnetic field lines (or the greater the rate at which magnetic field lines cut through the ring), the greater the force on the charged particles and hence the greater the current. Faraday expressed this in a manner that is easier to analyze. He said that the current is determined by the rate at which the number of magnetic field lines encircled by the loop is changing. In fact, Faraday was able to write this statement in equation form. Before I show you that, I have to be a lot more specific about what I mean by "the number of magnetic field lines."

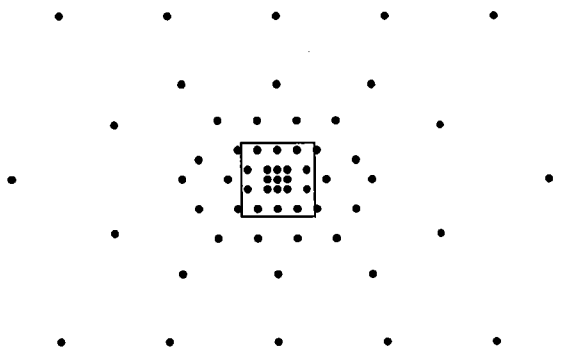
I'm going to call the statement I have just attributed to Faraday, the conceptual form of Faraday's Law. In other words, *Faradays Law*, in conceptual form is: *A changing number of magnetic field lines through a closed loop or coil causes a current in that loop or coil, and, the faster the number is changing, the greater the current.*

Our field line concept is essentially a diagrammatic scheme used to convey some information about the direction and the relative strength of a field. We have used it both for the electric field and the magnetic field. What I say here about the number of field lines can be applied to both, but, since we are presently concerned with the magnetic field, I will talk about it in terms of the magnetic field. Conceptually, the number of field lines encircled by a loop is going to depend on how closely packed the field lines are, how big the loop is, and to what degree the loop is oriented "face-on" to the field lines. (Clearly, if the loop is oriented edge-on to the field lines, it will encircle none of them.) Now, diagrammatically, how closely packed the field lines are is representative of how strong the magnetic field is. The more closely-packed the field lines, the greater the value of B. Imagine that someone has created a beautiful, three-dimensional, magnetic field diagram. Now if you view the field lines end-on, e.g. such that the magnetic field lines are directed right at you, and depict a cross section of "what you see" in a two-dimensional diagram, you would get something like this.



This is a graphical representation of the magnitude of that component of the magnetic field which is directed straight at you.

Suppose the scale of the diagram to be given by $(1\mu T\cdot m^2)n$ where n is the magnetic field line density, the number-of-magnetic-field-lines-per-area, directed through the plane represented by the page, straight at you. Let's use a square, one centimeter on a side, to sample the field at a position near the center,



I count 19 field lines that are clearly in the square centimeter and four that are touching it, I'm going to count two of those four for an estimated 21 field lines in one square centimeter. Thus, in that region,

$$n = \frac{21\text{ lines}}{(1\times10^{-2}\text{ m})^2}$$

$$n = 2100\frac{\text{lines}}{\text{m}^2}$$

Using the given scale factor,

$$B = (1.0\mu T\cdot m^2)n$$

$$B = (1.0\mu T\cdot m^2)2100\frac{\text{lines}}{\text{m}^2}$$

$$B = 2.1\text{ mT}$$

Let's make it more clear what the number of lines represents by replacing n with $\frac{Number\ of\ Lines}{A}$ and solving the expression $B = (1.0\mu T)n$ for the number of lines.

$$B = (1.0\mu T\cdot m^2)\frac{Number\ of\ Lines}{A}$$

$$Number\ of\ Lines = \frac{BA}{1.0\ \mu T\cdot m^2}$$

So the number of lines through a loop encircling a plane region of area A is proportional to BA, with the constant of proportionality being the reciprocal of our scale factor for the field diagram. The simple product BA is really only good if the magnetic field lines are "hitting" the area encircled by the loop "head on," and, if the magnetic field is single-valued over the whole area. We can take care of the "which way the loop is facing" issue by replacing BA with $\vec{\mathbf{B}} \cdot \vec{\mathbf{A}}$ where $\vec{\mathbf{A}}$, the area vector, is a vector whose magnitude is the area of the plane region encircled by the loop and whose direction is perpendicular to the plane of the loop. There are actually two directions that are perpendicular to the loop. One is the opposite of the other. In practice, one picks one of the two directions arbitrarily, but, picking a direction for the area vector establishes a positive direction for the current around the loop. The positive direction for the current is the direction around the loop that makes the current direction and the area vector direction, together, conform to the right-hand rule for something curly something straight. We take care of the possible variation of the magnetic field over the region enclosed by the loop, by cutting that plane region up into an infinite number of infinitesimal area elements dA, calculating $\vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$ for each area element, and adding up all the results. The final result is the integral $\int \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$. You won't be held responsible for using the calculus algorithms for analyzing such an integral, but, you are responsible for knowing what $\int \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$ means. It is the infinite sum you get when you subdivide the area enclosed by the loop up into an infinite number of infinitesimal area elements, and, for each area element, dot the magnetic field vector at the location of that area element into the area vector of that area element, and add up all the resulting dot products. You also need to know that, in the *special case* of a magnetic field that is *constant in both magnitude and direction* over the entire area enclosed by the loop, $\int \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$ is just $\vec{\mathbf{B}} \cdot \vec{\mathbf{A}}$.

Using a generic "*constant*" for the reciprocal of the field diagram scale factor yields

$$\textit{Number of Lines} = (\textit{constant}) \int \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$$

for the number of field lines encircled by the loop or coil. The quantity $\int \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$ is called the *magnetic flux* through the plane region enclosed by the loop. Note that the flux is directly proportional to the number of magnetic field lines through the loop.

The magnetic flux is given the name Φ_B (the Greek letter upper case phi).

$$\Phi_B = \int \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$$

The expression yields $\mathrm{T \cdot m^2}$ as the units of magnetic flux. This combination of units is given a name, the Weber, abbreviated Wb.

$$1\ \mathrm{Wb} = \mathrm{T \cdot m^2}$$

Faraday's Law, the one that says that the current induced in a loop or coil is proportional to the rate of change in the number of magnetic field lines encircled by the loop or coil, can be written in terms of the flux as:

$$I = -\frac{N}{R}\frac{d\Phi_B}{dt}$$

where:

N is the number of windings or turns making up the closed coil of wire. $N = 1$ for a single loop.

R is the resistance of the loop or coil.

$\frac{d\Phi_B}{dt}$ is the rate of change in the flux through the loop.

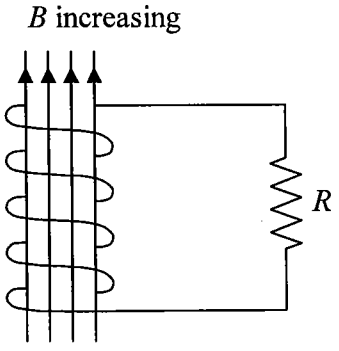
The derivative of a function with respect to *time* is often abbreviated as the function itself with a dot over it. In other words,

$$\dot{\Phi}_B = \frac{d\Phi_B}{dt}$$

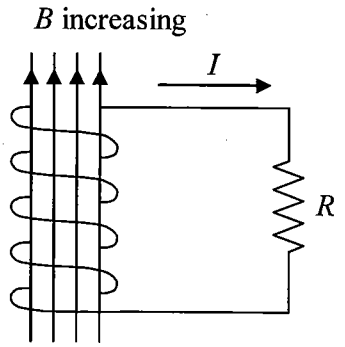
Using this notation in our expression for the current in Faradays Law of induction we have:

$$I = -\frac{N}{R}\dot{\Phi}_B$$

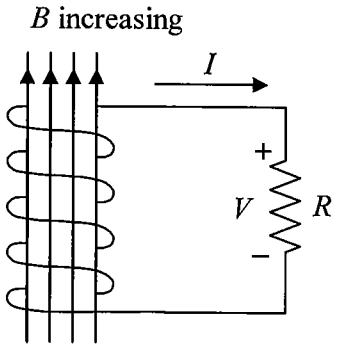
Faraday's Law is usually expressed in terms of an EMF rather than a current. I'm going to use a specific case study to develop the idea which is of general applicability. Consider a coil of *ideally-conducting* wire in series with a resistor. For closure of the loop, the resistor is to be considered part of the loop (and hence is the resistance *of* the loop), but, we have a negligible number of magnetic field lines cutting through the resistor itself. Suppose there to be an increasing magnetic flux directed upward through the coil.



By Faraday's Law of Induction, there will be a current $I = -\frac{N}{R}\dot{\Phi}_B$ induced in the coil. The charge will flow around and around the coil, out the top of the coil and down through the resistor.



But, for a resistor to have a current in it, there must be a potential difference $V = IR$ between the terminals of the resistor.



Recognizing that, in the case at hand, the I in $V = IR$ is the $I = -\frac{N}{R}\dot{\Phi}_B$ resulting from the changing magnetic flux through the coil, we have

$$V = \left(-\frac{N}{R}\dot{\Phi}_B\right)R$$

which we can write as

$$V = -N\dot{\Phi}_B$$

Where there is a voltage across a resistor, there is an electric field in the resistor. What exactly causes that electric field? The answer is, the changing flux through the coil. More specifically, it is the magnetic field lines cutting through the coil as they must be doing to cause a change in the number of field lines through the coil. The field lines through the coil causes a force on the charge carriers in the coil. In our positive charge carrier model, this causes positive charge carriers in the coil all to surge toward the top of the resistor, leaving an absence of same on the bottom of the resistor. It only takes a minuscule amount of charge to cause an appreciable electric field in the resistor. A dynamic equilibrium is reached in which the changing magnetic field force on the charged particle becomes unable to push any more charge to the top terminal of the resistor than is forced through the resistor by the electric field in the resistor. The changing magnetic field can't push more charge there because of the repulsion of the charge that is already there. The changing magnetic field force in the coil maintains the potential difference across the resistor in spite of the fact that charge carriers keep "falling" through the resistor. This should sound familiar. A seat of EMF does the same thing. It maintains a constant potential difference between two conductors (such as the terminals of the resistor in the case at hand). The coil with the changing flux through it is acting like a seat of EMF. One says that the changing flux induces an EMF in the coil, calls that Faraday's Law of Induction, and writes:

$$\mathcal{E} = -N\dot{\Phi}_B$$

where:

$\mathcal{E}$ is the EMF induced in the loop.

N is the number of windings or turns making up the coil of wire.

$\dot{\Phi}_B$ is the rate of change in the flux through the loop.

Lenz's Law

Faraday's Law of Induction has the direction of the current built into it. One arbitrarily establishes the direction of the area vector of the loop. This determines, via the right-hand rule for something curly something straight, the positive direction for the current. Then the algebraic sign of the result of $I = -\frac{N}{R}\dot{\Phi}_B$ determines whether the current is really in that direction ("+") or in the opposite direction ("–"). This is tough to keep track of. I advise using Faraday's Law in the form

$$|I| = \frac{N}{R}\left|\dot{\Phi}_B\right|$$

to get the magnitude of the current, and, using Lenz's Law to get the direction.

The current induced in the loop or coil, by the changing flux through the loop or coil, *produces* a magnetic field of its own. I call *that* magnetic field B_{PIN} for "the magnetic field produced by the induced current." At points inside the loop or coil, B_{PIN} is related to the induced current itself by the right hand rule for something curly something straight. Lenz's Law states that B_{PIN} is in that direction which tends to keep the number of magnetic field lines what it was.

Consider, for instance, a horizontal loop.

> Suppose there are magnetic field lines directed *upward* through the loop, and, that they are *increasing* in number. By Faraday's Law, the changing number of field lines through the loop will induce a current in the loop. By Ampere's Law, a current in the loop will produce a magnetic field (B_{PIN}). By Lenz's Law, B_{PIN} will be downward to cancel out some of the newly-appearing upward magnetic field lines, in a futile attempt to keep the number of magnetic field lines directed upward through the loop, what it was. By the right-hand rule for something curly something straight; to produce a downward-directed magnetic field line inside the loop, the induced current must be *clockwise*, around the loop, as viewed from above.
>
> Suppose there are magnetic field lines directed *upward* through the loop, and, that they are *decreasing* in number. By Faraday's Law, the changing number of field lines through the loop will induce a current in the loop. By Ampere's Law, a current in the loop will produce a magnetic field (B_{PIN}). By Lenz's Law, B_{PIN} will be upward to make up for the departure of upward-directed magnetic field lines, in a futile attempt to keep the number of magnetic field lines directed upward through the loop, what it was. By the right-hand rule for something curly something straight; to produce an upward-directed magnetic field line inside the loop, the induced current must be *counterclockwise*, around the loop, as viewed from above.

19 Induction, Transformers, and Generators

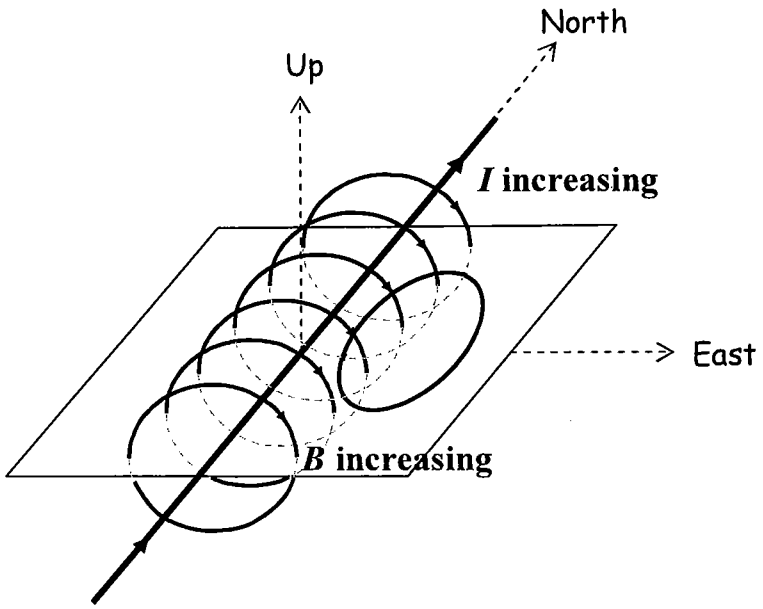
In this chapter we provide examples chosen to further familiarize you with Faraday's Law of Induction and Lenz's Law. The last example is the generator, the device used in the world's power plants to convert mechanical energy into electrical energy.

Example 19-1

A straight wire carries a current due northward. Due east of the straight wire, at the same elevation as the straight wire, is a horizontal loop of wire. The current in the straight wire is increasing. Which way is the current, induced in the loop by the changing magnetic field of the straight wire, directed around the loop?

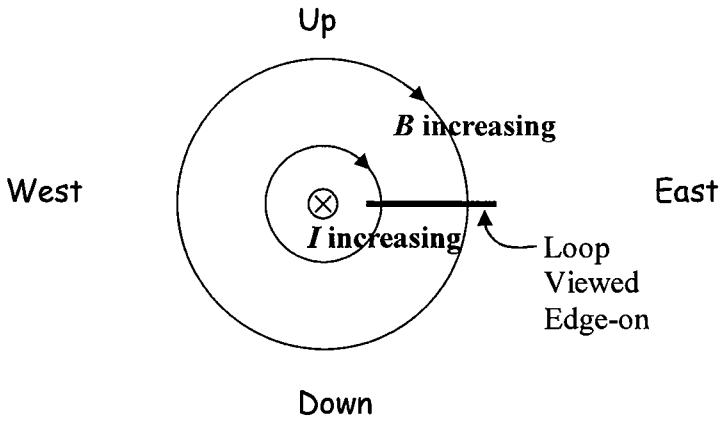
Solution

I'm going to draw the given situation from a few different viewpoints, just to help you get used to visualizing this kind of situation. As viewed from above-and-to-the-southeast, the configuration (aside from the fact that magnetic field lines are invisible) appears as:



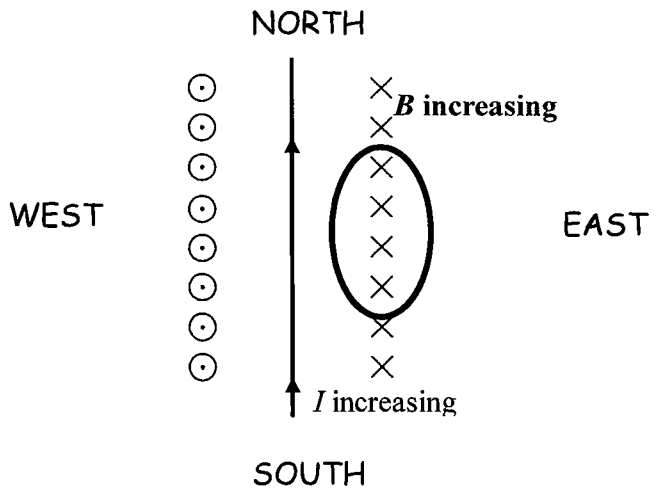
where I included a sheet of paper in the diagram to help you visualize things.

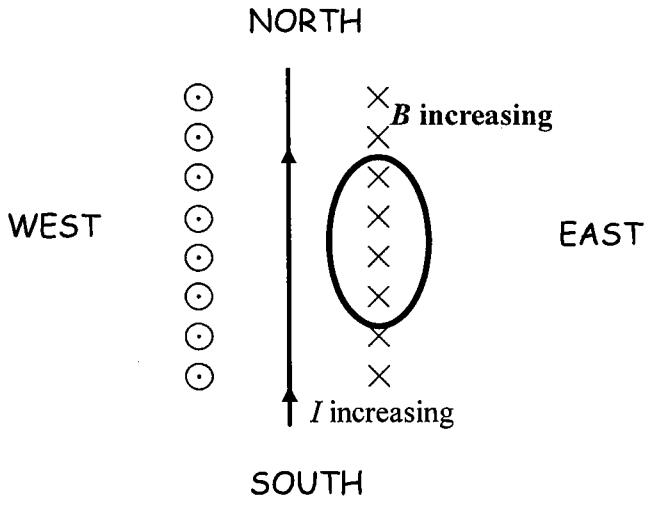
Here's a view of the same configuration from the south, looking due north:



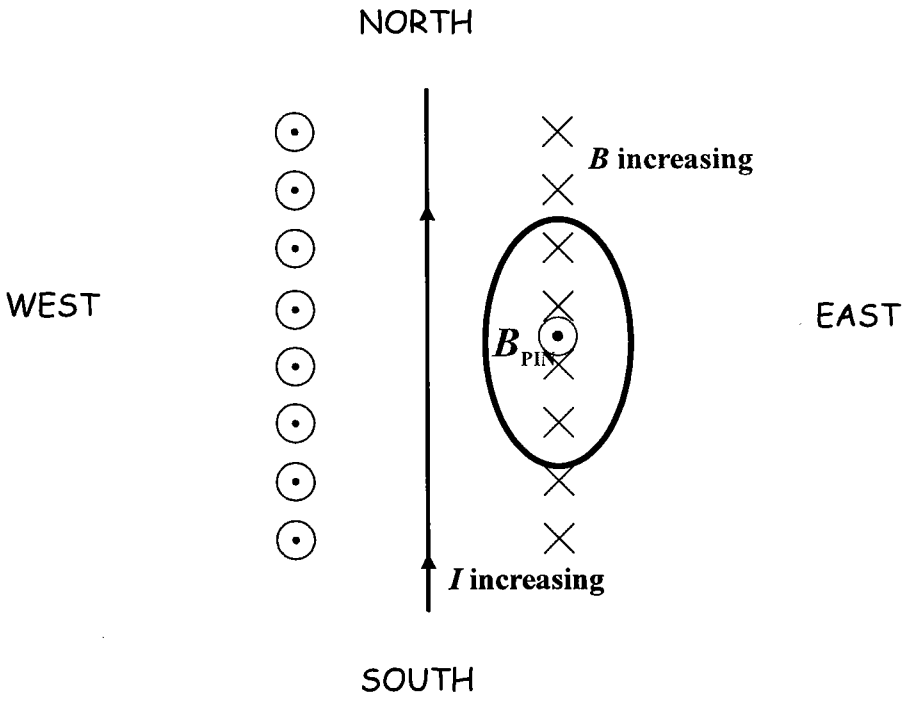
Both diagrams make it clear that we have an increasing number of downward-directed magnetic field lines through the loop. It is important to keep in mind that a field diagram is a diagrammatic manner of conveying information about an infinite set of vectors. *There is no such thing as a curved vector.* A vector is always directed along a straight line. The magnetic field vector is tangent to the magnetic field lines characterizing that vector. At the location of the loop, every magnetic field vector depicted in the diagram above is straight downward. While it is okay to say that we have an increasing number of magnetic field lines directed downward through the loop, please keep in mind that the field lines characterize *vectors*.

In presenting my solution to the example question, "What is the direction of the current induced in a horizontal loop that is due east of a straight wire carrying an increasing current due north?" I wouldn't draw either one of the diagrams above. The first one takes too long to draw and there is no good way to show the direction of the current in the loop in the second one. The view from above is the most convenient one:

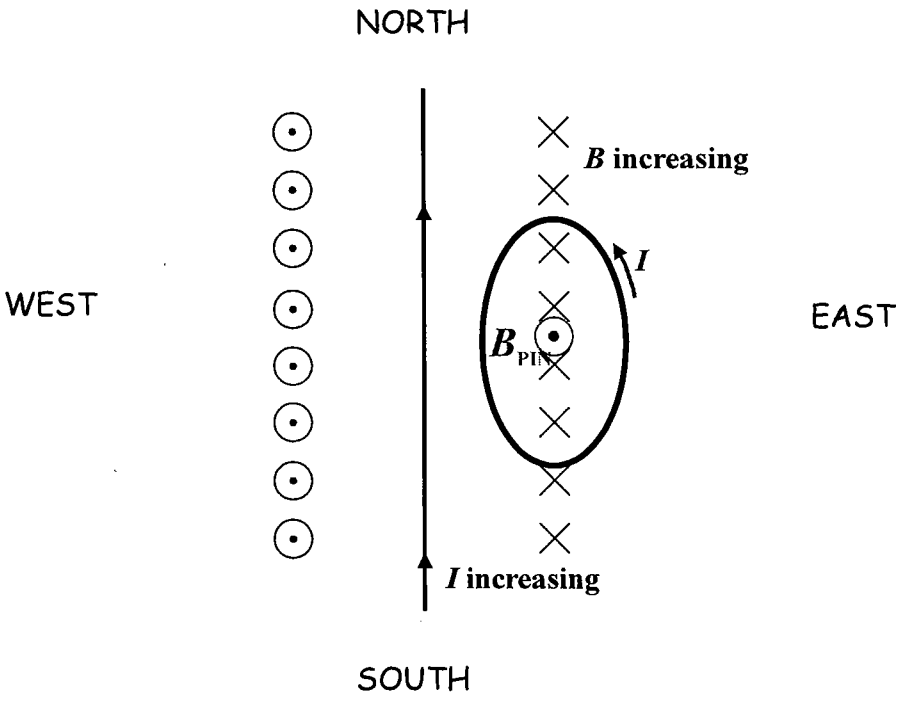




In this view (in which the downward direction is into the page) it is easy to see that what we have is an increasing number of downward-directed magnetic field lines through the loop (more specifically, through the region enclosed by the loop). In its futile attempt to keep the number of magnetic field lines directed downward through the loop the same as what it was, $\vec{\mathbf{B}}_{\text{PIN}}$ must be directed *upward* in order to cancel out the newly-appearing downward-directed magnetic field lines. [Recall the sequence: The changing number of magnetic field lines *induces* (by Faraday's Law) a current in the loop. That current *produces* (by Ampere's Law) a magnetic field ($\vec{\mathbf{B}}_{\text{PIN}}$) of its own. Lenz's Law relates the *end product* ($\vec{\mathbf{B}}_{\text{PIN}}$) to the *original change* (increasing number of downward-through-the-loop magnetic field lines).]



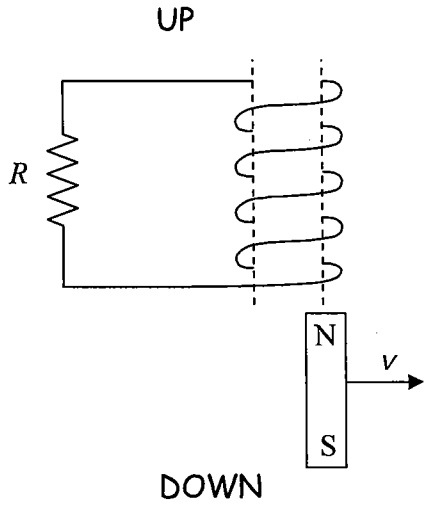
That's interesting. We know the direction of the magnetic field produced by the induced current before we even known the direction of the induced current itself. So, what must the direction of the induced current be in order to produce an upward-directed magnetic field ($\vec{\mathbf{B}}_{PIN}$)? Well, by the right-hand rule for something curly something straight, the current must be counterclockwise, as viewed from above.



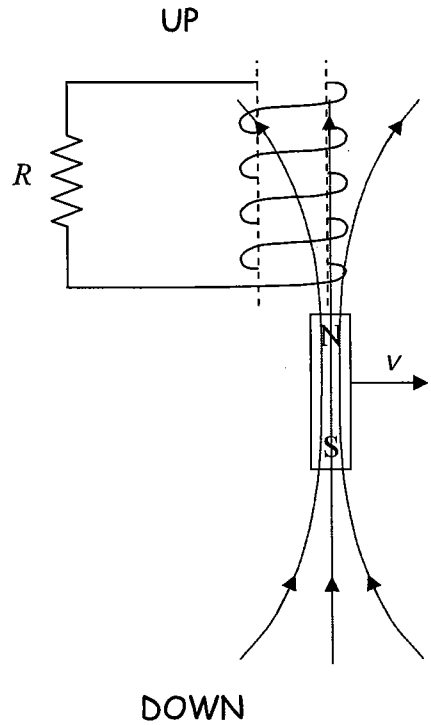
Hey. That's the answer to the question. We're done with that example. Here's another one:

Example 19-2

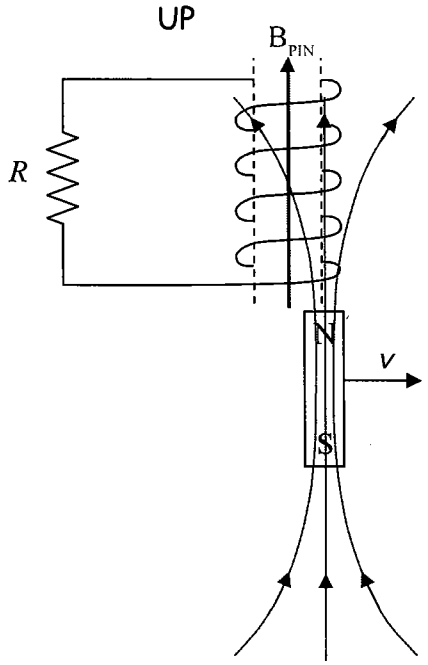
A person is moving a bar magnet, aligned north pole up, out from under a coil of wire, as depicted below. What is the direction of the current in the resistor?



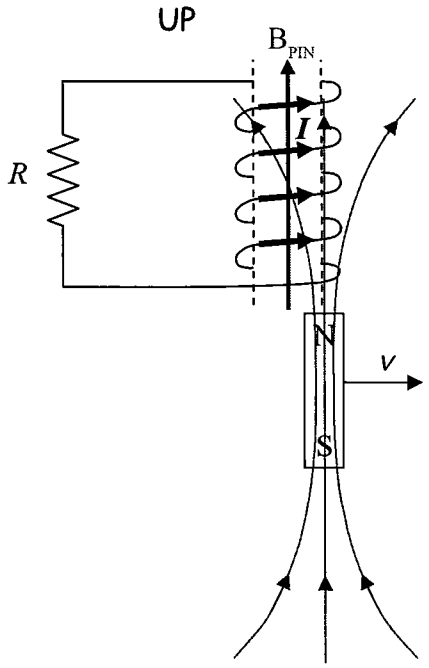
The magnetic field of the bar magnet extends upward through the coil.



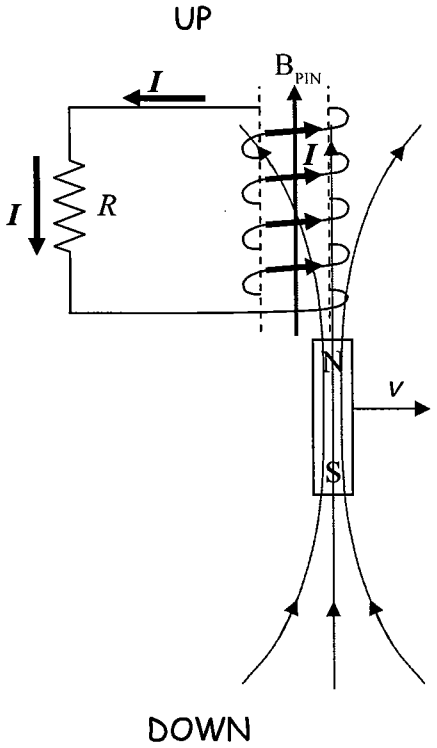
As the magnet moves out from under the coil, it takes its magnetic field with it. So, as regards the coil, what we have is a decreasing number of upward-directed magnetic field lines through the coil. By Faraday's Law, this induces a current in the coil. By Ampere's Law, the current produces a magnetic field, $\vec{\mathbf{B}}_{\text{PIN}}$. By Lenz's Law $\vec{\mathbf{B}}_{\text{PIN}}$ is upward, to make up for the departing upward-directed magnetic field lines through the coil.



So, what is the direction of the current that is causing $\vec{\mathbf{B}}_{\text{PIN}}$? The right-hand rule will tell us that. Point the thumb of your cupped right hand in the direction of $\vec{\mathbf{B}}_{\text{PIN}}$. Your fingers will then be curled around (counterclockwise as viewed from above) in the direction of the current.



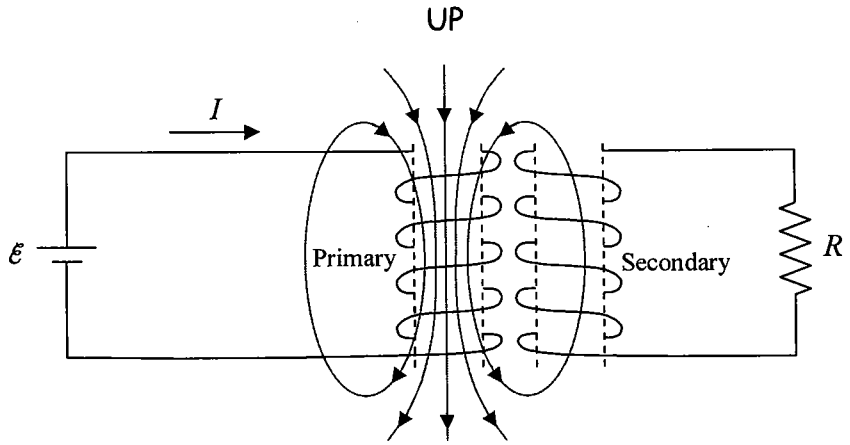
Because of the way the coil is wound, such a current will be directed out the top of the coil *downward* through the resistor.



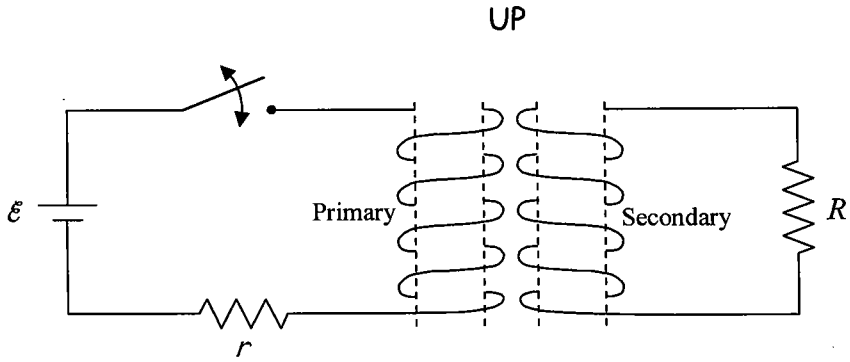
That's the answer to the question posed in the example. (What is the direction of the current in the resistor?)

Example 19-3 The Transformer

When you put two coils of wire near each other, such that when you create a magnetic field by using a seat of EMF to cause a current in one coil, *that magnetic field* extends through the region encircled by the other coil, you create a *transformer*. Let's call the coil in which you initially cause the current, the primary coil, and the other one, the secondary coil.



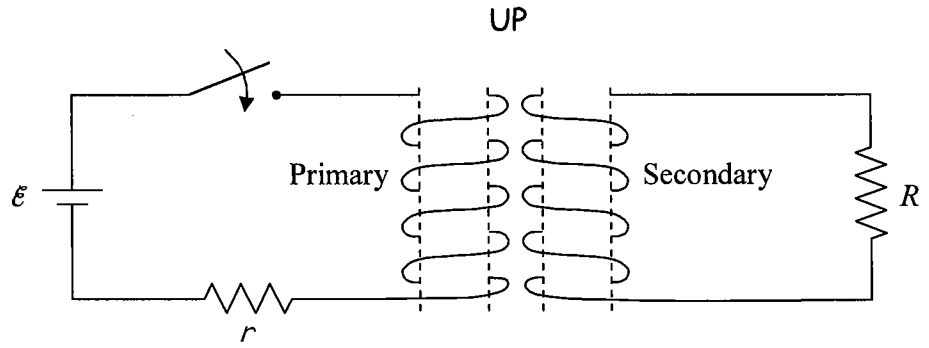
If you cause the current in the primary coil to be changing, then the magnetic field produced by that coil is changing. Thus, the flux through the secondary coil is changing, and, by Faraday's Law of Induction, a current will be induced in the secondary coil. One way to cause the current in the primary coil to be changing would be to put a switch in the primary circuit (the circuit in which the primary coil is wired) and to repeatedly open and close it.



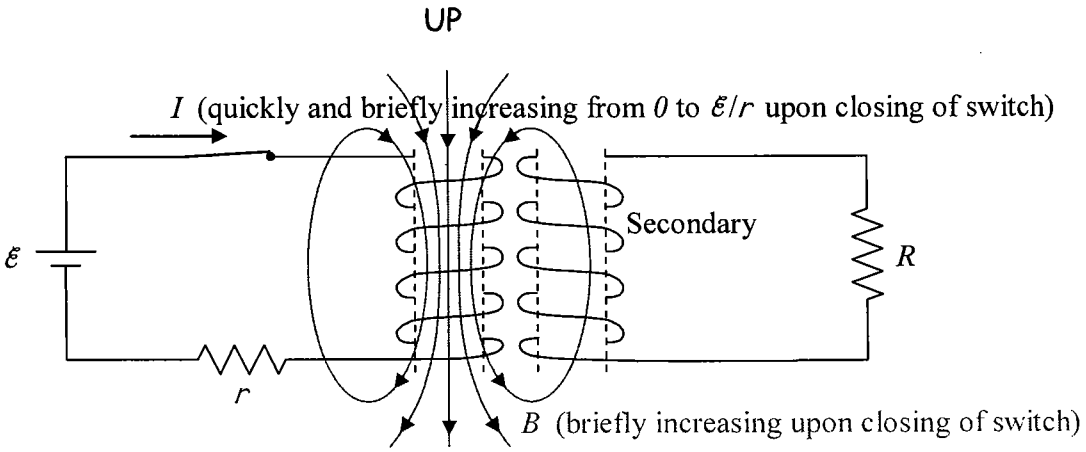
Okay, enough preamble, here's the question: What is the direction of the transient[1] current induced in the circuit above when the switch is closed?

[1] (This footnote is about the English language rather than physics.) Transient means *existing for a short time interval.*

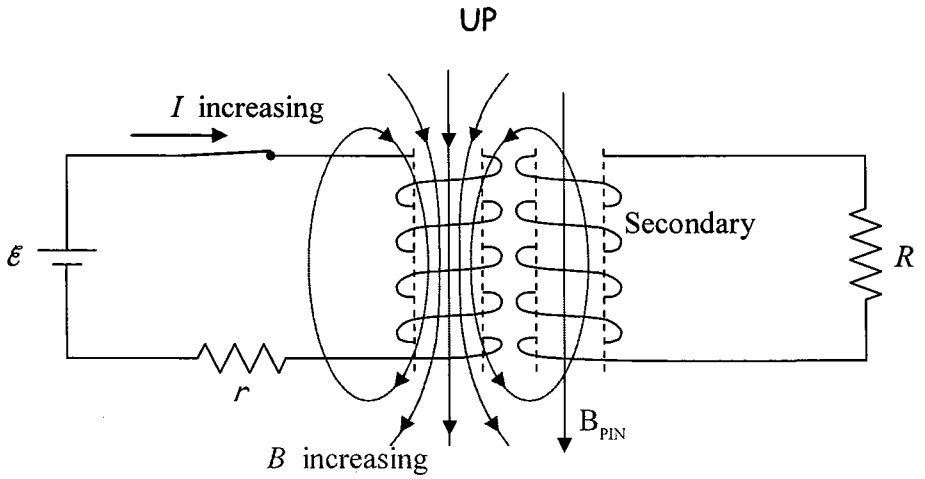
Solution to *Example 19-3*:



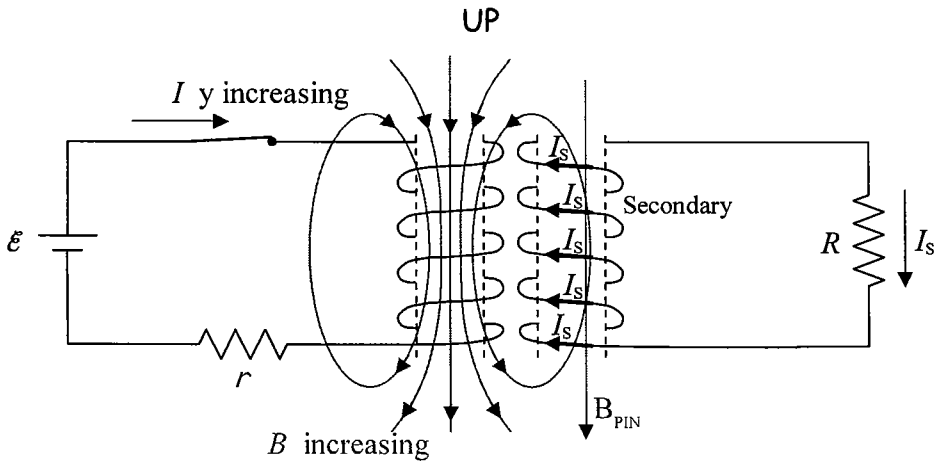
Upon closing the switch, the current in the primary circuit very quickly builds up to $\mathcal{E}/r$. While the time that it takes for the current to build up to $\mathcal{E}/r$ is very short, it is during this time interval that the current is changing. Hence, it is on this time interval that we must focus our attention in order to answer the question about the direction of the transient current in resistor R in the secondary circuit. The current in the primary causes a magnetic field. Because the current is increasing, the magnetic field vector at each point in space is increasing in magnitude.



The increasing magnetic field causes upward-directed magnetic field lines in the region encircled by the secondary coil. There were no magnetic field lines through that coil before the switch was closed, so clearly, what we have here is an increasing number of upward-directed magnetic field lines through the secondary coil. By Faraday's Law this will induce a current in the coil. By Ampere's law, the current induced in the secondary will produce a magnetic field of its own, one that I like to call $\vec{\mathbf{B}}_{\text{PIN}}$ for "The Magnetic field Produced by the Induced Current." By Lenz's Law, $\vec{\mathbf{B}}_{\text{PIN}}$ must be downward to cancel out some of the newly-appearing upward-directed magnetic field lines through the secondary. (I hope it is clear that what I call the magnetic field lines *through* the secondary, are the magnetic field lines passing through the region encircled by the secondary coil.)



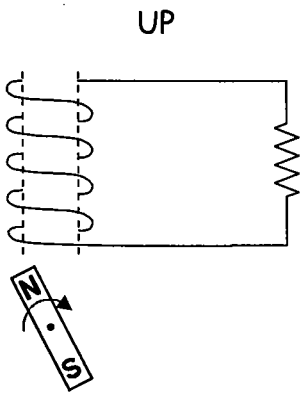
Okay. Now the question is, which way must the current be directed around the coil in order to create the downward-directed magnetic field $\vec{\mathbf{B}}_{\text{PIN}}$ that we have deduced it does create. As usual, the right-hand rule for something curly something straight reveals the answer. We point the thumb of the cupped right hand in the direction of $\vec{\mathbf{B}}_{\text{PIN}}$ and cannot fail to note that the fingers curl around in a direction that can best be described as "clockwise as viewed from above."



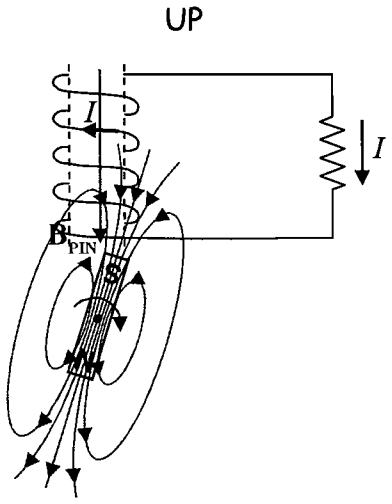
Because of the way the secondary coil is wound, such a current will be directed out of the secondary at the top of the coil and downward through resistor R. This is the answer to the question posed in the example.

An Electric Generator

Consider a magnet that is caused to rotate in the vicinity of a coil of wire as depicted below.

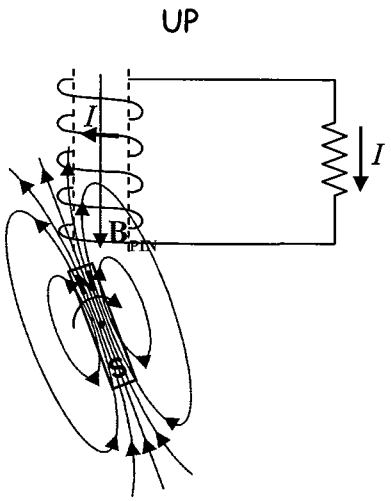


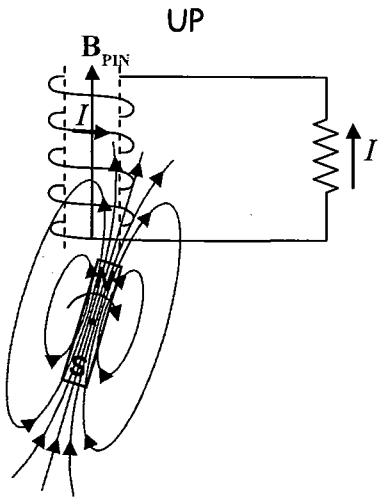
As a result of the rotating magnet, the number and direction of the magnetic field lines through the coil is continually changing. This induces a current in the coil, which, as it turns out, is also changing. Check it out in the case of magnet that is, from our viewpoint, rotating clockwise. In the orientation of the rotating magnet depicted here:

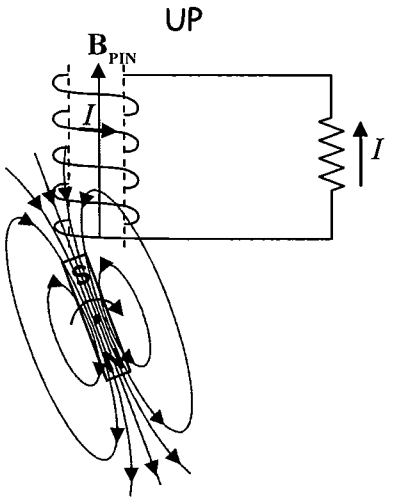


as the magnet rotates, the number of its magnetic field lines extending *downward* through the coil is *decreasing*. In accord with Faraday's Law, this induces a current in the coil which, in accord with Ampere's Law, produces a magnetic field of its own. By Lenz's Law, the field ($\vec{\mathbf{B}}_{PIN}$) produced by the induced current must be downward to make up for the loss of downward-directed magnetic field lines through the coil. To produce $\vec{\mathbf{B}}_{PIN}$ downward, the induced current must be clockwise, as viewed from above. Based on the way the wire is wrapped and the coil is connected in the circuit, a current that is clockwise as viewed from above, in the coil, is directed out of the coil at the top of the coil and downward through the resistor.

In the following diagrams we show the magnet in each of several successive orientations. Keep in mind that someone or something is spinning the magnet by mechanical means. You can assume for instance that a person is turning the magnet with her hand. As the magnet turns the number of magnetic field lines is changing in a specific manner for each of the orientations depicted. You the reader are asked to apply Lenz's Law and the Right Hand Rule for Something Curley, Something Straight to verify that the current (caused by the spinning magnet) through the resistor is in the direction depicted:





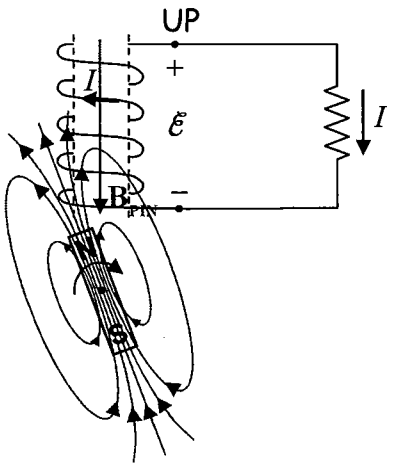


As the magnet continues to rotate clockwise, the next orientation it achieves is our starting point and the process repeats itself over and over again.

Recapping and extrapolating, the current through the resistor in the series of diagrams above, is:

downward, downward, *upward, upward*, downward, downward, *upward, upward*, ...

For half of each rotation, the current is downward, and for the other half of each rotation, the current is upward. In quantifying this behavior, one focuses on the EMF induced in the coil:



The EMF across the coil varies sinusoidally with time as:

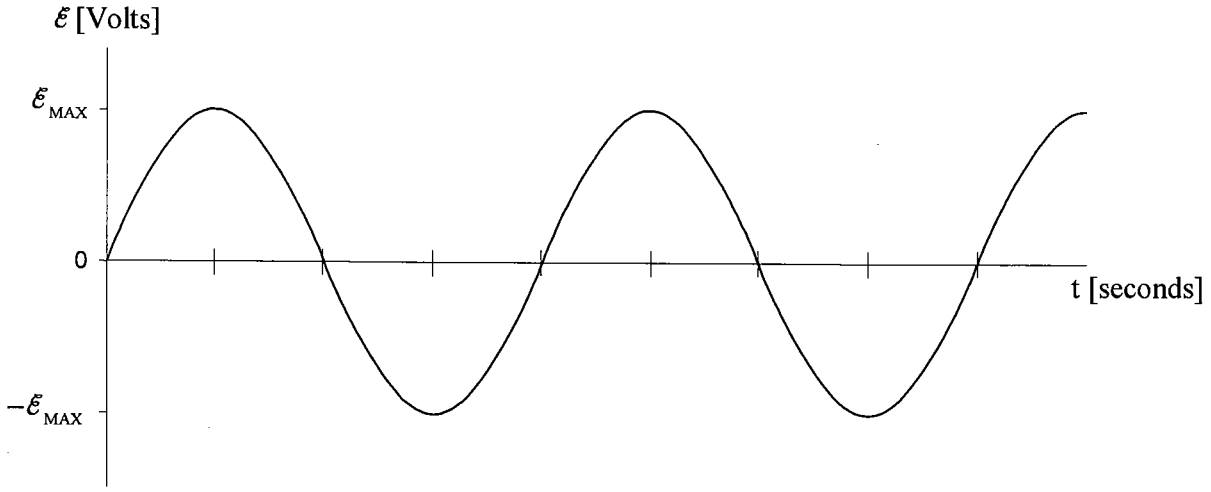
$$\mathcal{E} = \mathcal{E}_{MAX} \sin(2\pi f t) \tag{19-1}$$

where:

$\mathcal{E}$ which stands for EMF, is the time-varying electric potential difference between the terminals of a coil in close proximity to a magnet that is rotating relative to the coil as depicted in the diagrams above. This potential difference is caused to exist, and to vary the way it does, by the changing magnetic flux through the coil.

$\mathcal{E}_{MAX}$ is the maximum value of the EMF of the coil.

f is the frequency of oscillations of the EMF across the coil. It is exactly equal to the rotation rate of the magnet expressed in rotations per second, a unit that is equivalent to hertz.

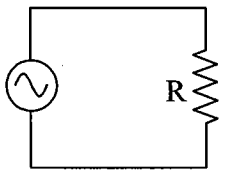


The device that we have been discussing (coil-plus-rotating magnet) is called a generator, or more specifically, an *electric generator*. A generator is a seat of EMF that causes there to be a potential difference between its terminals that varies sinusoidally with time. The schematic representation of such a time-varying seat of EMF is:

It takes work to spin the magnet. The magnetic field caused by the current induced in the coil exerts a torque on the magnet that always tends to slow it down. So, to keep the magnet spinning, one must continually exert a torque on the magnet in the direction in which it is spinning. The generator is the main component of any electrical power plant. *It converts mechanical energy to electrical energy.* The kind of power plant you are dealing with is determined by what your power company uses to spin the magnet. If moving water is used to

spin the magnet, we call the power plant a hydroelectric plant. If a steam turbine is used to spin the magnet, then the power plant is designated by its method of heating and vaporizing water. For instance, if one heats and vaporizes the water by means of burning coal, one calls the power plant a coal-fired power plant. If one heats and vaporizes the water by means of a nuclear reactor, one calls the power plant a nuclear power plant.

Consider a "device which causes a potential difference between its terminals that varies sinusoidally with time" in a simple circuit:

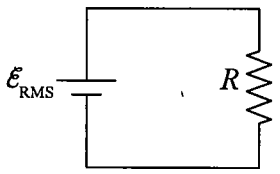


The time-varying seat of EMF causes a potential difference across the resistor, in this simple circuit, equal, at any instant in time, to the voltage across the time-varying seat of EMF. As a result, there is a current in the resistor. The current is given by $I = \frac{V}{R}$, our defining equation for resistance, solved for the current I. Because the algebraic sign of the potential difference across the resistor is continually alternating, the direction of the current in the resistor is continually alternating. Such a current is called an *alternating current* (AC). It has become traditional to use the abbreviation AC to the extent that we do so in a redundant fashion, often referring to an alternating current as an AC current. (When we need to distinguish it from AC, we call the "one-way" kind of current that, say, a battery causes in a circuit, *direct current*, abbreviated DC.)

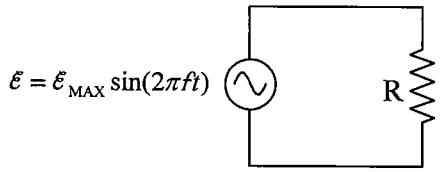
A device that causes current in a resistor, whether that current is alternating or not, is delivering energy to the resistor at a rate that we call power. The power delivered to a resistor can be expressed as $P=IV$ where I is the current through the resistor and V is the voltage across the resistor. Using the defining equation of resistance, $V=IR$, the power can be expressed as $P=I^2R$. A "device which causes a potential difference between its terminals that varies sinusoidally with time", what I have been referring to as a "time-varying seat of EMF" is typically referred to as an *AC power source*. An AC power source is typically referred to in terms of the frequency of oscillations, and, the voltage that a DC power source, an ordinary seat of EMF, would have to maintain across its terminals to cause the same average power in any resistor that might be connected across the terminals of the AC power source. The voltage in question is typically referred to as $\mathcal{E}_{\text{RMS}}$ or V_{RMS} where the reasoning behind the name of the subscript will become evident shortly.

Since the power delivered by an ordinary seat of EMF is a constant, its average power is the value it always has.

Here's the fictitious circuit



that would cause the same resistor power as the AC power source in question. The average power (which is just the *power* in the case of a DC circuit) is given by $P_{AVG}=I\mathcal{E}_{RMS}$, which, by means of our defining equation of resistance solved for I, $I=V/R$, (where the voltage across the resistor is, by inspection, $\mathcal{E}_{RMS}$) can be written $P_{AVG}=\frac{\mathcal{E}_{RMS}^2}{R}$. So far, this is old stuff, with an unexplained name for the EMF voltage. Now let's consider the AC circuit:



The power is $P=\frac{\mathcal{E}^2}{R}=\frac{[\mathcal{E}_{MAX}\sin(2\pi ft)]^2}{R}=\frac{\mathcal{E}_{MAX}^2[\sin(2\pi ft)]^2}{R}$. The average value of the square of the sine function is $\frac{1}{2}$. So the average power is $P_{AVG}=\frac{1}{2}\frac{\mathcal{E}_{MAX}^2}{R}$. Combining this with our expression $P_{AVG}=\frac{\mathcal{E}_{RMS}^2}{R}$ from above yields:

$$\frac{\mathcal{E}_{RMS}^2}{R}=\frac{1}{2}\frac{\mathcal{E}_{MAX}^2}{R}$$

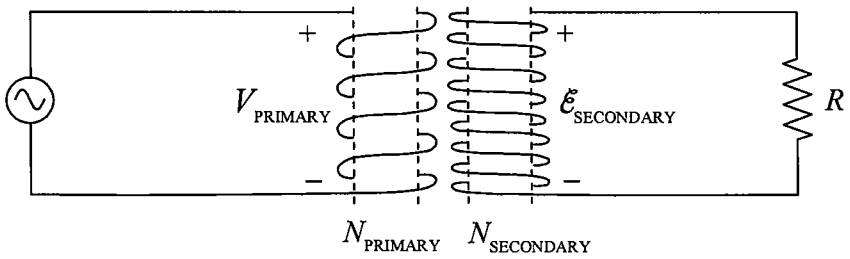
$$\mathcal{E}_{RMS}=\sqrt{\frac{1}{2}}\,\mathcal{E}_{MAX} \qquad (19\text{-}2)$$

Now we are in a position to explain why we called the equivalent EMF, $\mathcal{E}_{RMS}$. In our expression $P_{AVG}=\frac{1}{2}\frac{\mathcal{E}_{MAX}^2}{R}$, we can consider $\frac{\mathcal{E}_{MAX}^2}{2}$ to be the average value of the square of our time-varying EMF $\mathcal{E}=\mathcal{E}_{MAX}\sin(2\pi ft)$. Another name for "average" is "mean" so we can consider $\frac{\mathcal{E}_{MAX}^2}{2}$ to be the *mean* value of $\mathcal{E}^2$. On the right side of our expression for our equivalent EMF, $\mathcal{E}_{RMS}=\frac{1}{\sqrt{2}}\mathcal{E}_{MAX}$, we have the square root of $\frac{\mathcal{E}_{MAX}^2}{2}$, that is, we have the *square root* of the *mean* of the *square* of the EMF $\mathcal{E}$. And indeed the subscript "RMS" stands for "root mean squared." RMS values are convenient for circuits consisting of resistors and AC power sources in that, one can analyze such circuits using RMS values the same way one analyzes DC circuits.

More on the Transformer

When the primary coil of a transformer is driven by an AC power source, it creates a magnetic field which varies sinusoidally in such a manner as to cause a sinusoidal EMF, of the same frequency as the source, to be induced in the secondary coil. The RMS value of the EMF induced in the secondary coil is directly proportional to the RMS value of the sinusoidal potential difference imposed across the primary. The constant of proportionality is the ratio of the number of turns in the secondary to the number of turns in the primary.

$$\mathcal{E}_{\text{SECONDARY}} = \frac{N_{\text{SECONDARY}}}{N_{\text{PRIMARY}}} V_{\text{PRIMARY}} \tag{19-3}$$



When the number of windings in the secondary coil is *greater than* the number of windings in the primary coil, the transformer is said to be a *step-up transformer* and the secondary voltage is greater than the primary voltage. When the number of windings in the secondary coil is *less than* the number of windings in the primary coil, the transformer is said to be a *step-down transformer* and the secondary voltage is less than the primary voltage.

The Electrical Power in Your House

When you plug your toaster into a wall outlet, you bring the prongs of the plug into contact with two conductors between which there is a time-varying potential difference characterized as 115 volts 60 Hz AC. The 60 Hz is the frequency of oscillations of the potential difference resulting from a magnet completing 60 rotations per second, back at the power plant. A step-up transformer is used near the power plant to step the power plant output up to a high voltage. Transmission lines at a very high potential, with respect to each other, provide a conducting path to a transformer near your home where the voltage is stepped down. Power lines at a much lower potential provide the conducting path to the wires in your home. 115 volts is the RMS value of the potential difference between the two conductors in each pair of slots in your wall outlets. Since $\mathcal{E}_{\text{RMS}} = \frac{1}{\sqrt{2}}\mathcal{E}_{\text{MAX}}$, we have $\mathcal{E}_{\text{MAX}} = \sqrt{2}\,\mathcal{E}_{\text{RMS}}$, so $\mathcal{E}_{\text{MAX}} = \sqrt{2}\,(115 \text{ volts})$, or $\mathcal{E}_{\text{MAX}} = 163 \text{ volts}$. Thus,

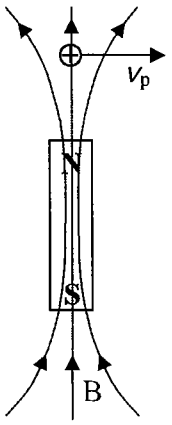
$$\mathcal{E} = (163 \text{ volts})\sin[2\pi(60\text{Hz})t]$$

which can be written as,

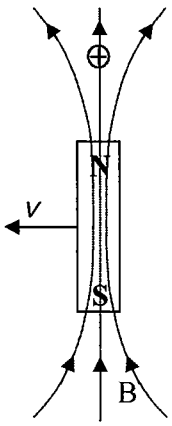
$$\mathcal{E} = (163 \text{ volts})\sin[(377\,\frac{\text{rad}}{\text{s}})t]$$

20 Faraday's Law and Maxwell's Extension to Ampere's Law

Consider the case of a charged particle that is moving in the vicinity of a moving bar magnet as depicted in the following diagram:



When we view the situation from the reference frame of the magnet, what we see (as depicted just above) is a charged particle moving in a stationary magnetic field. We have already studied the fact that a magnetic field exerts a force $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}_p \times \vec{\mathbf{B}}$ on a charged particle moving in that magnetic field. Now let's look at the same phenomenon from the point of view of the charged particle:

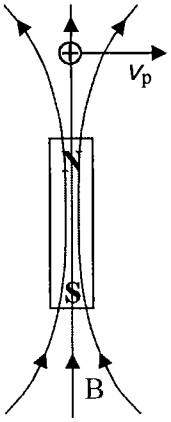


(where $\vec{\mathbf{v}} = -\vec{\mathbf{v}}_p$).

Surely we aren't going to change the force exerted on the charged particle by the magnetic field of the magnet just by looking at the situation from a different reference frame. In fact we've already addressed this issue. What I said was that it is the relative motion between the magnet and the charged particle that matters. Whether the charged particle is moving through magnetic field lines, or the magnetic field lines, due to their motion, are moving sideways through the particle, the particle experiences a force. *Now here's the new viewpoint on this situation*: What we say is, that the moving magnetic field doesn't really exert a force on the stationary charged

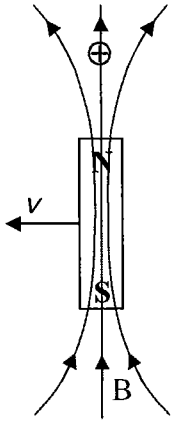
particle, but rather, that by moving sideways through the point at which the particle is located, the magnetic field creates an electric field at that location, and it is the electric field that exerts the force on the charged particle. In this viewpoint, we have, at the location of the stationary charged particle, an electric field that is exerting a force on the particle, and a magnetic field that is exerting no force on the particle. At this stage it might seem that it would be necessary to designate the magnetic field as some special kind of magnetic field that doesn't exert a force on a charged particle despite the relative velocity between the charged particle and the magnetic field. Instead, what we actually do is to characterize the magnetic field as being at rest relative to the charged particle.

So, as viewed from the reference frame in which the magnet is at rest:



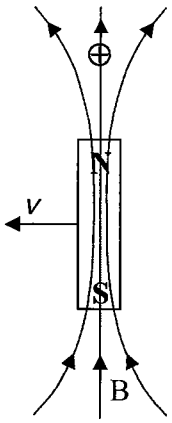
the particle experiences a force $\vec{\mathbf{F}}$ directed out of the page in the diagram above due to its motion through the magnetic field.

And, as viewed from the reference frame in which the charged particle is at rest:



the particle finds itself in a stationary magnetic field but experiences the same force $\vec{\mathbf{F}}$ because it also finds itself in an electric field directed out of the page.

So we have two models for explaining the force on the stationary charged particle in the case depicted by:



In model 1 we simply say that in terms of the Lorentz Force $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}\times\vec{\mathbf{B}}$, what matters is the relative velocity between the particle and the magnetic field and to calculate the force we identify the velocity $\vec{\mathbf{v}}_p$ of the particle relative to the magnetic field as being rightward at magnitude $v_p = v$ in the diagram above so $\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}_p\times\vec{\mathbf{B}}$ (where q is the charge of the particle). In model 2 we say that the apparent motion of the magnetic field "causes" there to be an electric field and a stationary magnet field so the particle experiences a force $\vec{\mathbf{F}} = q\,\vec{\mathbf{E}}$.

Of course we are using two different models to characterize the same force. In order for both models to give the same result we must have:

$$\vec{\mathbf{E}} = \vec{\mathbf{v}}_p\times\vec{\mathbf{B}} \qquad (20\text{-}1)$$

where:

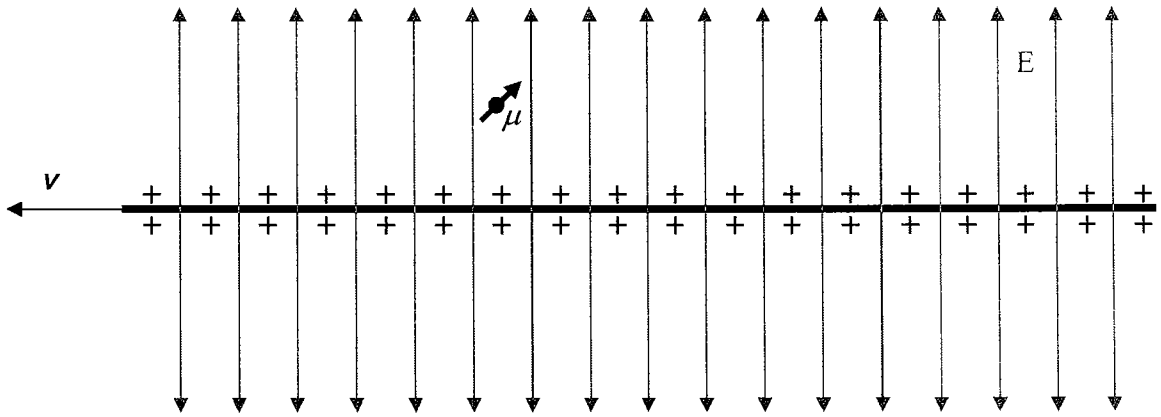
$\vec{\mathbf{E}}$ is the electric field at an empty point in space due to the motion of that point relative to a magnetic field vector that exists at that point in space,

$\vec{\mathbf{v}}_p$ is the velocity of the empty point in space relative to the magnetic field vector, and

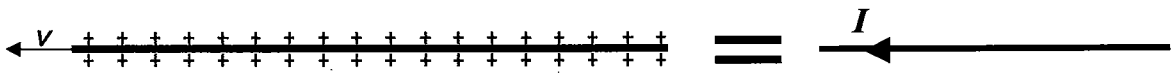
$\vec{\mathbf{B}}$ is the magnetic field vector.

Physicists have found model 2 to be more fruitful, especially when attempting to explain magnetic waves. The idea that a magnetic field in apparent sideways motion through a point in space "causes" there to be an electric field at that point in space, is referred to as Faraday's Law of Induction. Our mnemonic for Faraday's Law of Induction is: "A changing magnetic field causes an electric field."

The acceleration experienced by a charged particle in the vicinity of a magnet, when the charged particle is moving relative to the magnet represents an experimental result that we have characterized in terms of the model described in the preceding part of this chapter. The model is useful in that it can be used to predict the outcome of, and provide explanations regarding, related physical processes. Another experimental result is that a particle that has a magnetic dipole moment and is moving in an electric field with a velocity that is neither parallel nor antiparallel to the electric field, does (except for two special magnetic dipole moment directions) experience angular acceleration. We interpret this to mean that the particle experiences a torque. Recalling that a particle with a magnetic dipole moment that is at rest in an electric field experiences no torque, but one that is at rest in a magnetic field does indeed experience a torque (as long as the magnetic dipole moment and the magnetic field it is in are not parallel or antiparallel to each other), you might think that we can model the fact that a particle with a magnetic dipole moment experiences a torque when it is moving relative to an electric field, by defining a magnetic field "caused" by the apparent motion of the electric field relative to the particle. You would be right. To build such a model, we consider a charged particle that is moving in an electric field produced by a long line of charge that is uniformly distributed along the line. We start by depicting the situation in the reference frame in which the particle is at rest and the line of charge is moving:



Note that we have two different ways of accounting for the magnetic field due to the moving line of charge, at the location of the particle with a magnetic dipole moment. The moving line of charge is a current so we can think of the magnetic field as being caused by the current.



The other option is to view the magnetic field as being caused by the electric field lines moving sideways through the particle. There is, however, only one magnetic field, so, the two different ways of accounting for it must yield the same result. We are going to arrive at an expression for

the magnetic field due to the motion of an electric field by forcing the two different ways of accounting for the magnetic field to be consistent with each other. First, we'll simply use Ampere's law to determine the magnetic field at the location of the particle.

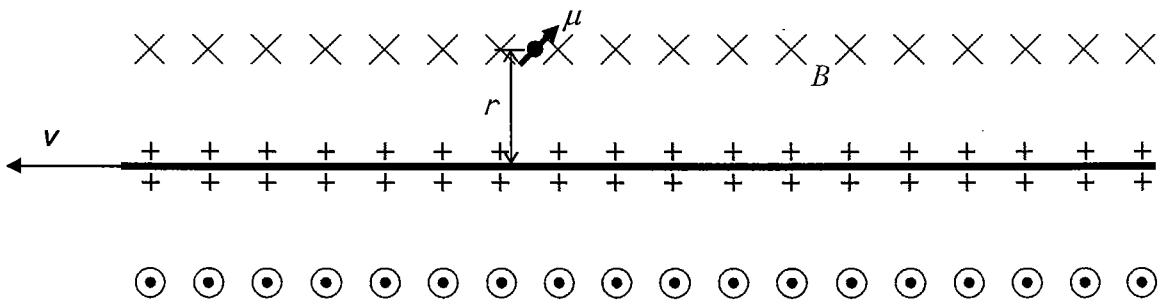
Let's define the linear charge density (the charge per length) of the line of charge to be λ and the distance that the particle is from the line of charge to be r. Suppose that in an amount of time dt the line of charge moves a distance dx. Then the amount of charge passing a fixed point on the line along which the charge is moving, in time dt, would be $\lambda\, dx$. Dividing the latter by dt yields $\lambda\, dx/dt$ which can be expressed as λv and is just the rate at which charge is flowing past the fixed point, that is, it is the current I. In other words, the moving line of charge is a current $I = \lambda v$. Back in chapter 17 we gave the experimental result for the magnetic field due to a long straight wire carrying current I in the form of an equation that we called "Ampere's Law." It was equation 17-2; it read:

$$B = \frac{\mu_o}{2\pi}\frac{I}{r}$$

and it applies here. substituting $I = \lambda v$ into this expression for B yields

$$B = \frac{\mu_o}{2\pi}\frac{\lambda v}{r} \qquad (20\text{-}2)$$

By the right hand rule for something curly something straight we know that the magnetic field is directed into the page at the location of the particle that has a magnetic dipole moment, as depicted in the following diagram:



Now let's work on obtaining an expression for the same magnetic field from the viewpoint that it is *the electric field moving sideways* through the location of the particle that causes the magnetic field. First we need an expression for the electric field due to the line of charge, at the location of the particle, that is, at a distance r from the line of charge. The way to get that is to consider the line of charge as consisting of an infinite number of bits of charged material, each of which is a segment of infinitesimal length dx of the line of charge. Since the line of charge has a linear charge density λ, this means that each of the infinitesimal segments dx has charge λdx. To get the electric field at the location of the particle that has a magnetic dipole moment, all we have to do is to add up all the contributions to the electric field at the location of the particle, due to all the infinitesimal segments of charged material making up the line of charge. Each contribution is given by Coulomb's Law for the Electric Field. The difficulty is that there are an infinite

number of contributions. You will be doing such calculations when you study chapter 30 of this textbook. At this stage, we simply provide the result for the electric field due to an infinitely long line of charge having a constant value of linear charge density λ:

$$E = \frac{\lambda}{2\pi r \epsilon_o}$$

Multiplying both sides by ϵ_0 yields

$$\epsilon_o E = \frac{\lambda}{2\pi r}$$

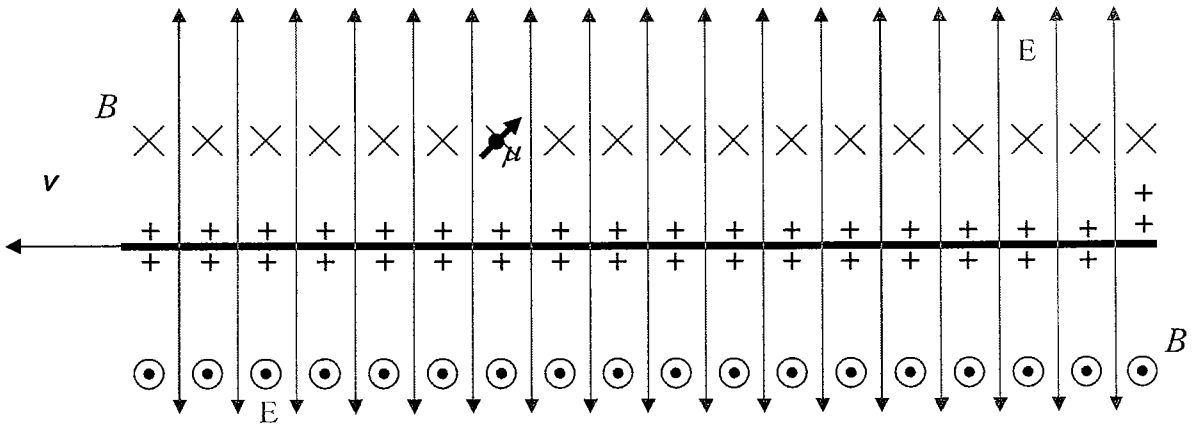
The expression on the right side of this equation appears in equation 20-2, $B = \frac{\mu_o}{2\pi}\frac{\lambda v}{r}$.

Substituting $\epsilon_o E$ for $\frac{\lambda}{2\pi r}$ where the latter appears in equation 20-2 yields:

$$B = \mu_o \epsilon_o E v$$

This represents the magnitude of the magnetic field that is experienced by a particle when it is moving with speed $v_p = v$ relative to an electric field $\vec{\mathbf{E}}$ when the velocity is perpendicular to $\vec{\mathbf{E}}$. Experimentally we find that a particle with a magnetic dipole moment experiences no torque (and hence no magnetic field) if its velocity is parallel or antiparallel to the electric field $\vec{\mathbf{E}}$. As such, we can make our result more general (not only good for the case when the velocity is perpendicular to the electric field) if we write, $E_\perp$ in place of E.

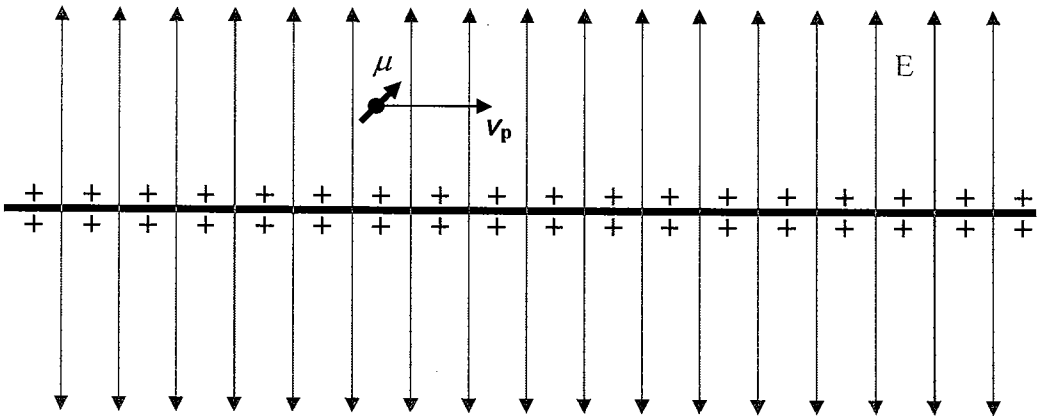
$$B = \mu_o \epsilon_o E_\perp v$$



Starting with the preceding equation, we can bundle both the magnitude and the direction (as determined from Ampere's Law and the right hand rule when we treat the moving line of charge as a current, and as depicted in the diagram above) of the magnetic field into one equation by writing:

$$\vec{\mathbf{B}} = \mu_o \epsilon_o \vec{\mathbf{v}} \times \vec{\mathbf{E}}$$

We can express $\vec{\mathbf{B}}$ in terms of the velocity $\vec{\mathbf{v}}_{\mathrm{P}}$ of the particle relative to the line of charge



(instead of the velocity $\vec{\mathbf{v}}$ of the line of charge relative to the particle) just by recognizing that

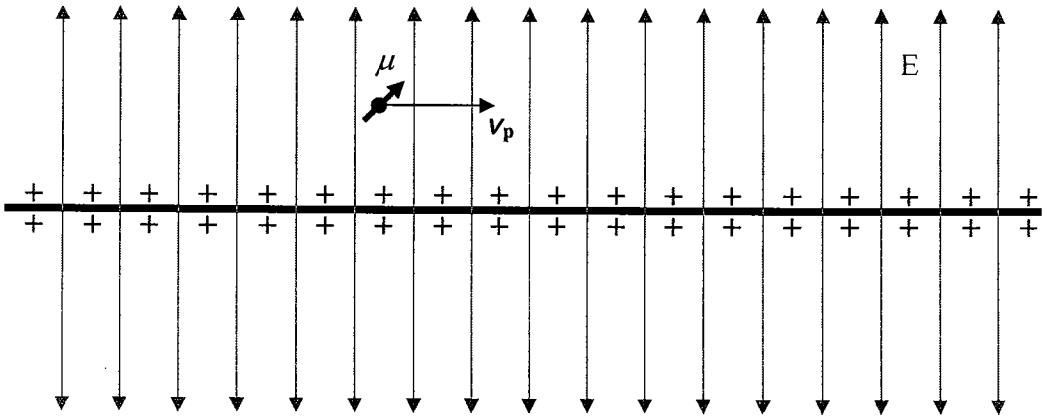
$$\vec{\mathbf{v}}_{\mathrm{P}} = -\vec{\mathbf{v}}$$

Substituting this expression ($\vec{\mathbf{v}}_{P} = -\vec{\mathbf{v}}$) into our expression for the magnetic field ($\vec{\mathbf{B}} = \mu_o \epsilon_o \vec{\mathbf{v}} \times \vec{\mathbf{E}}$) yields:

$$\vec{\mathbf{B}} = -\mu_o \epsilon_o \vec{\mathbf{v}}_{\mathrm{P}} \times \vec{\mathbf{E}}$$

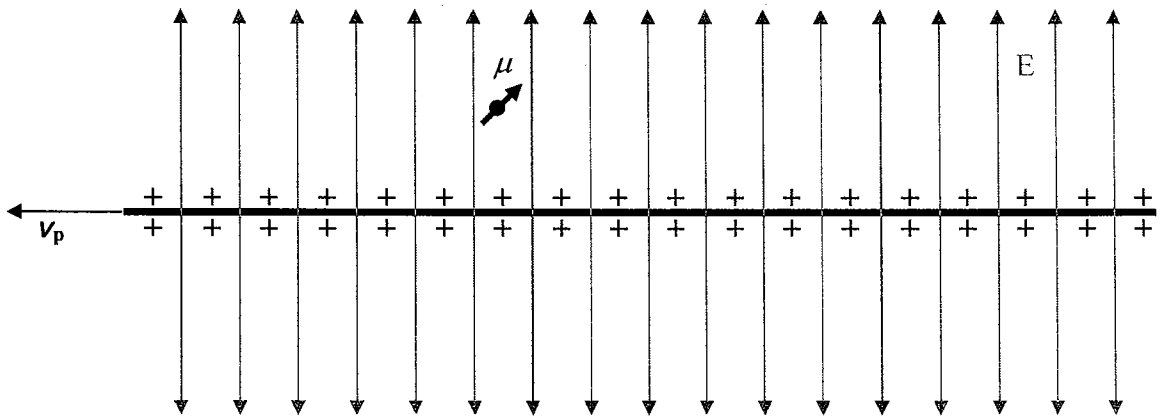
In this model, where we account for the torque experienced by a particle that has a magnetic dipole moment when that particle is moving in an electric field, by defining a magnetic field $\vec{\mathbf{B}} = -\mu_o \epsilon_o \vec{\mathbf{v}}_{\mathrm{P}} \times \vec{\mathbf{E}}$ which depends both on the velocity of the particle relative to the electric field and the electric field itself, the electric field itself is considered to exert no torque on the charged particle. At this stage it might seem that it would be necessary to designate the electric field as some special kind of electric field that doesn't exert a torque on a charged particle despite the relative velocity between the charged particle and the electric field. Instead, what we actually do is to characterize the electric field as being at rest relative to the charged particle.

So, as viewed from the reference frame in which the line of charge is at rest:



the particle that has a magnetic dipole moment experiences a torque due to its motion through the electric field.

And, as viewed from the reference frame in which the particle is at rest:

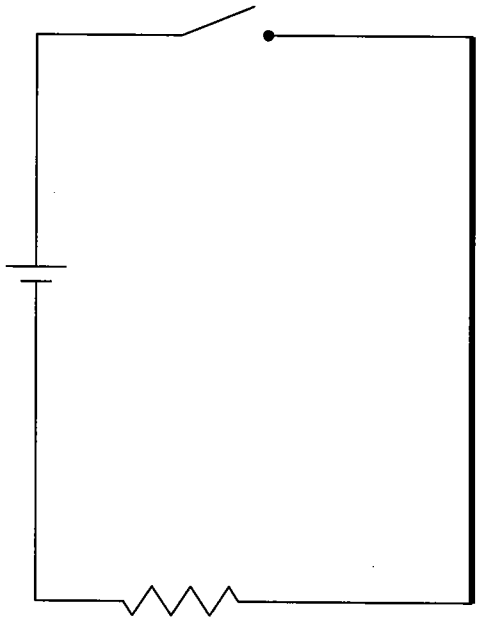


the particle that has a magnetic dipole moment finds itself in a stationary electric field but experiences the same torque because it also finds itself in a magnetic field directed, in the diagram above, into the page. One way of saying what is going on here is to say that, loosely speaking: A changing electric field "causes" a magnetic field. The phenomenon of a changing electric field "causing" a magnetic field is referred to as *Maxwell's Extension to Ampere's Law.*

So far, in this chapter we have addressed two major points: A magnetic field moving sideways through a point in space causes there to be an electric field at that point in space, and, an electric field moving sideways through a point in space causes there to be a magnetic field at that point in space. In the remainder of this chapter we find that putting these two facts together yields something interesting.

Expressing what we have found in terms of the point of view in which *point* P *is fixed and the field is moving through point* P with speed $\vec{v} = -\vec{v}_p$, we have: a magnetic field vector $\vec{B}$ moving with velocity $\vec{v}$ transversely through a point in space will "cause" an electric field $\vec{E} = -\vec{v} \times \vec{B}$ at that point in space; and; an electric field vector moving with velocity $\vec{v}$ transversely through a point in space will "cause" a magnetic field $\vec{B} = \mu_o \epsilon_o \vec{v} \times \vec{E}$ at that point in space. The word "cause" is in quotes because there is never any time delay. A more precise way of putting it would be to say that whenever we have a magnetic field vector moving transversely through a point in space, there exists, simultaneously, an electric field $\vec{E} = -\vec{v} \times \vec{B}$ at that point in space, and whenever we have an electric field vector moving transversely through a point in space there exists, simultaneously, a magnetic field $\vec{B} = \mu_o \epsilon_o \vec{v} \times \vec{E}$ at that point in space.

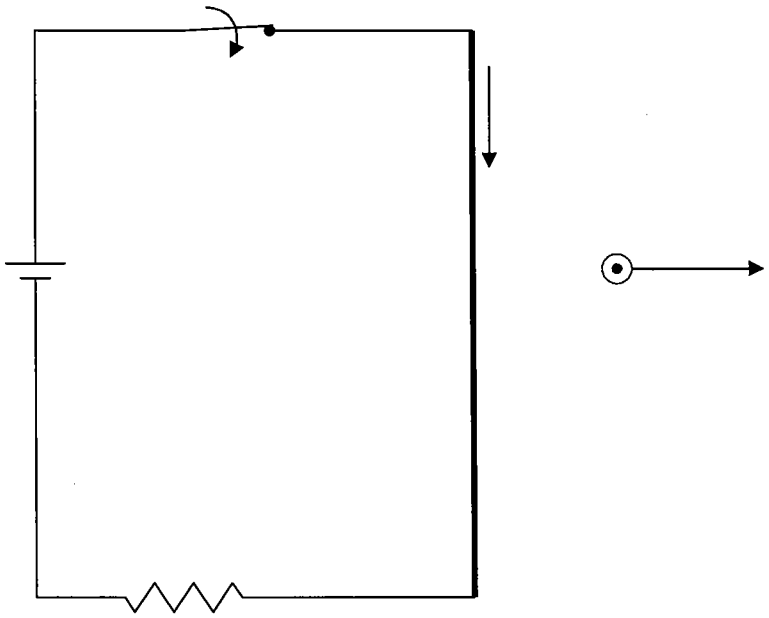
Consider the following circuit. Assume that we are looking down on the circuit from above, meaning that into the page is downward, and out of the page is upward.



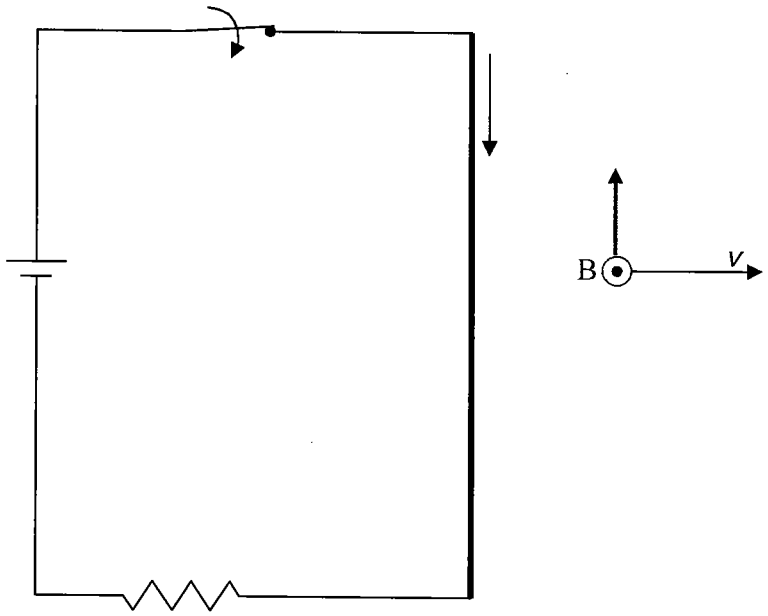
I want you to focus your attention on the rightmost wire of that circuit. As soon as someone closes that switch we are going to get a current through that wire and that current is going to produce a magnetic field. By means of the right-hand rule for something curly something straight, with the current being the something straight, and our knowledge that straight currents cause magnetic fields that make loops around the current, we can deduce that there will be an upward-directed (pointing out of the page) magnetic field at points to the right of the wire. In steady state, we understand that the upward-directed magnetic field vectors will be everywhere to the right of the wire with the magnitude of the magnetic field vector being smaller the greater the distance the point in question is from the wire. Now the question is, how long does it take for the magnetic field to become established at some point a specified distance to the right of the wire? Does the magnetic field appear instantly at every point to the right of the wire or does it take time? James Clerk Maxwell decided to explore the possibility that *it takes time*, in other words, that the magnetic field develops in the vicinity of the wire and moves outward with a finite velocity.

Here I want to talk about the leading edge of the magnetic field, the expanding boundary within which the magnetic field already exists, and outside of which, the magnetic field does not yet exist. With each passing infinitesimal time interval another infinitesimal layer is added to the region within which the magnetic field exists. While this is more a case of magnetic field vectors growing sideways through space, the effect of the motion of the leading edge through space is the same, at the growing boundary, as magnetic field vectors moving through space. As such, I am going to refer to this magnetic field growth as motion of the magnetic field through space.

To keep the drawing uncluttered I'm going to show just one of the infinite number of magnetic field vectors moving rightward at some unknown velocity (and it is *this* velocity that I am curious about) as the magnetic field due to the wire becomes established in the universe.



Again, what I'm saying is that, as the magnetic field builds up, what we have, are rightward-moving upward (pointing out of the page, toward you) magnetic field lines due to the current that just began. Well, as a magnetic field vector moves through whatever location it is moving through, it "causes" an electric field $\vec{\mathbf{E}} = -\vec{\mathbf{v}} \times \vec{\mathbf{B}}$.



At any point P through which the magnetic field vector passes, an electric field exists consistent with $\vec{\mathbf{E}} = -\vec{\mathbf{v}} \times \vec{\mathbf{B}}$. What this amounts to is that we have both a magnetic field and an electric field moving rightward through space. But we said that an electric field moving transversely through space "causes" a magnetic field. More specifically we said that it is always accompanied by a magnetic field given by $\vec{\mathbf{B}} = \mu_0 \epsilon_0 \vec{\mathbf{v}} \times \vec{\mathbf{E}}$. Now we've argued around in a circle. The current "causes" the magnetic field and its movement through space "causes" an electric field whose movement through space "causes" the magnetic field. Again, the word "causes" here should really be interpreted as "exists simultaneously with." Still, we have two explanations for the existence of one and the same magnetic field and the two explanations *must* be consistent with each other. For that to be the case, if we take our expression for the magnetic field "caused" by the motion of the electric field,

$$\vec{\mathbf{B}} = \mu_0 \epsilon_0 \vec{\mathbf{v}} \times \vec{\mathbf{E}}$$

and substitute into it, our expression $\vec{\mathbf{E}} = -\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ for the electric field "caused" by the motion of the magnetic field, we must obtain the same $\vec{\mathbf{B}}$ that, in this circular argument, is "causing" itself. Let's try it. Substituting $\vec{\mathbf{E}} = -\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ into $\vec{\mathbf{B}} = \mu_0 \epsilon_0 \vec{\mathbf{v}} \times \vec{\mathbf{E}}$, we obtain:

$$\vec{\mathbf{B}} = -\mu_0 \epsilon_0 \vec{\mathbf{v}} \times (\vec{\mathbf{v}} \times \vec{\mathbf{B}})$$

All right. Noting that $\vec{\mathbf{v}}$ is perpendicular to both $\vec{\mathbf{B}}$ and $\vec{\mathbf{v}} \times \vec{\mathbf{B}}$, meaning that the magnitude of the cross product, in each case, is just the product of the magnitudes of the multiplicand vectors, we obtain:

$$\vec{\mathbf{B}} = \mu_0 \epsilon_0 v^2 \vec{\mathbf{B}}$$

which I copy here for your convenience:

$$\vec{\mathbf{B}} = \mu_0 \epsilon_0 v^2 \vec{\mathbf{B}}$$

Again, it is one and the same $\vec{\mathbf{B}}$ on both sides, so, the only way this equation can be true is if $\mu_0 \epsilon_0 v^2$ is exactly equal to 1. Let's see where that leads us:

$$\mu_0 \epsilon_0 v^2 = 1$$

$$v^2 = \frac{1}{\mu_0 \epsilon_0}$$

$$v = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$$

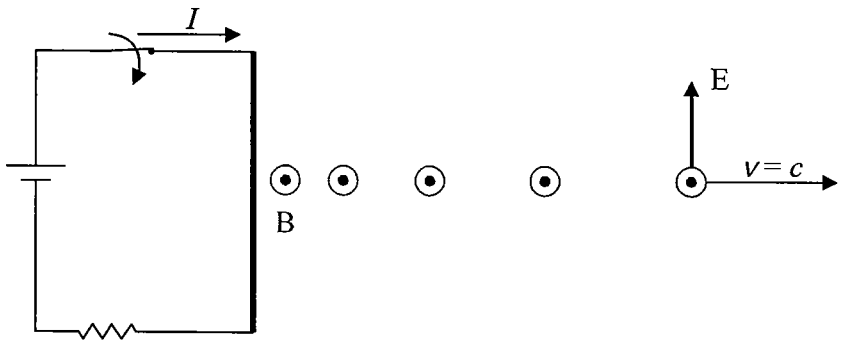
$$v = \frac{1}{\sqrt{\left(4\pi \times 10^{-7} \frac{\mathrm{T \cdot m}}{\mathrm{A}}\right) 8.85 \times 10^{-12} \frac{\mathrm{C}^2}{\mathrm{N \cdot m}^2}}}$$

$$v = 3.00 \times 10^8 \frac{\mathrm{m}}{\mathrm{s}}$$

Wow! That's the speed of light! When James Clerk Maxwell found out that electric and magnetic fields propagate through space at the (already known) speed of light he realized that light is electromagnetic waves.

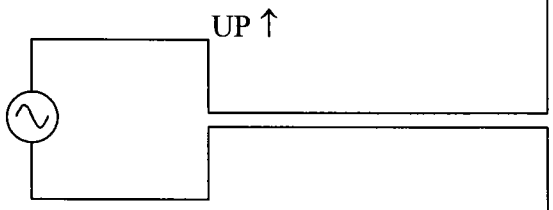
21 The Nature of Electromagnetic Waves

When we left off talking about the following circuit:



we had recently closed the switch and the wire was creating a magnetic field which was expanding outward. The boundary between that part of the universe in which the magnetic field is already established and that part of the universe in which the magnetic field has not yet been established is moving outward at the speed of light, $c = 3.00 \times 10^8$ m/s. Between that boundary and the wire we have a region in which there exists a steady unmoving magnetic field. Note that it was the act of creating the current that caused the magnetic field "edge" that is moving at the speed of light. In changing from a no-current situation to one in which there was current in the wire, charged particles in the wire went from no net velocity in the along-the-wire direction to a net velocity along the wire, meaning, that the charged particles were accelerated. In other words, accelerated charged particles cause light. We can also cause light by means of the angular acceleration of particles having a magnetic dipole moment, but, the *short* answer to the question about what causes light, is, *accelerated charged particles.*

Here's a simple circuit that one might use to intentionally cause light:



The vertical arrangement of wires on the right is referred to as a dipole antenna. As the AC power source alternately causes charge to surge upward in both parts of the antenna, and then downward, the dipole antenna creates electric and magnetic fields that oscillate sinusoidally in both time and space. The fields propagate through space away from the antenna at the speed of light.

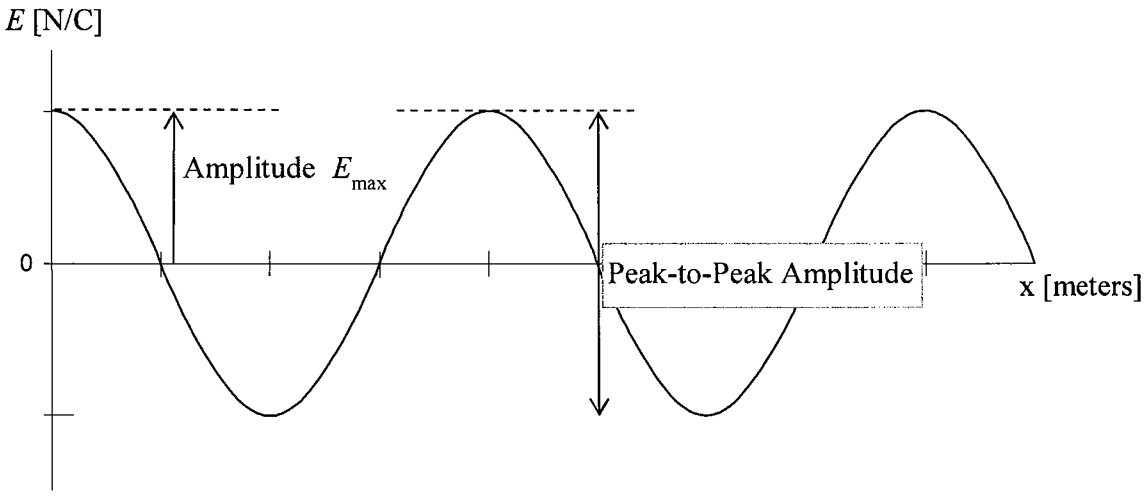
The charged particles oscillating up and down in the antenna causes waves of electric and magnetic fields known as light. The frequency of the waves is the same as the frequency of oscillations of the particles which is determined by the frequency of the power source. The speed of the waves is the speed of light $c = 3.00 \times 10^8$ m/s, because the waves are light. For any kind of wave, the frequency, wavelength, and wave speed are related by:

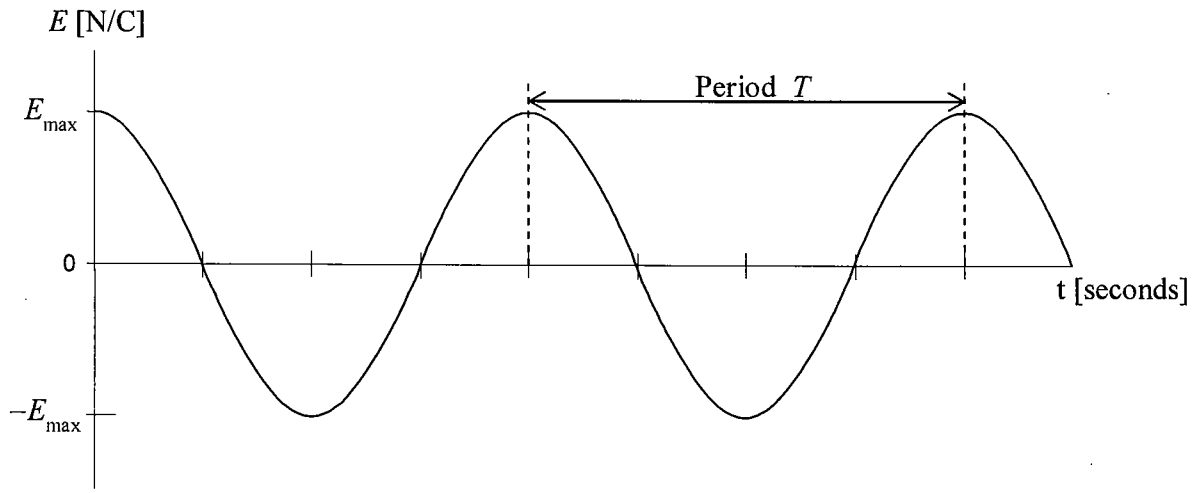
$$v = \lambda f$$

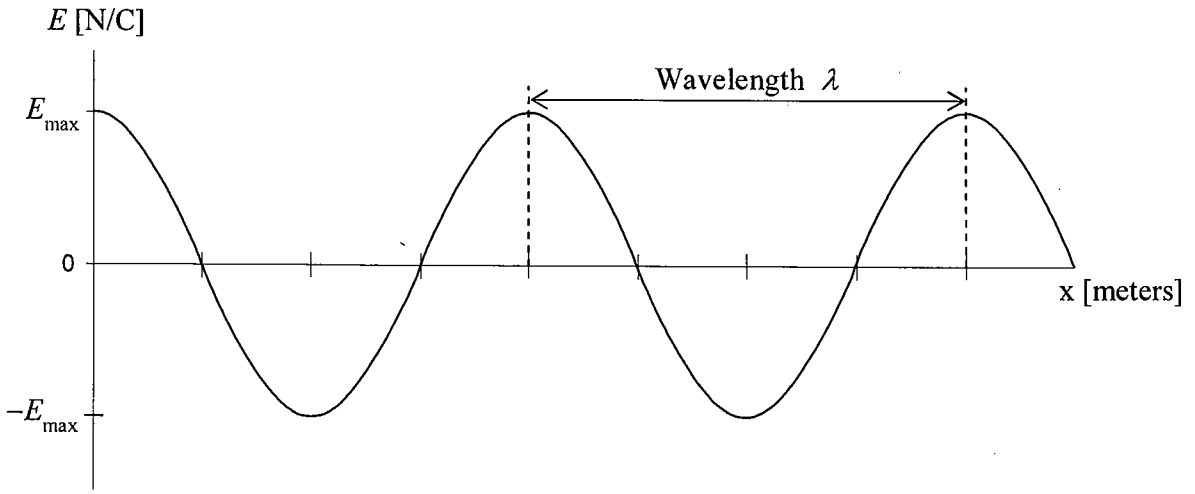
which, in the case of light reads:

$$c = \lambda f$$

Here's a quick pictorial review of some properties of waves. In the case of light, we have electric and magnetic fields oscillating in synchronization with each other. It is customary to characterize the waves in terms of the electric field. I'll do that here, but, one should keep in mind that the magnetic field oscillates and moves in the same manner that the electric field does, but, at right angles to the electric field.







The intensity of a wave is proportional to the square of its amplitude, so, in the case of light:

$$I \propto (E_{MAX})^2$$

The frequency of light is determined by the frequency of oscillations of the charged particles constituting the source of the light. How we categorize light depends on the frequency of the light. In order of increasing frequency, we refer to light as: radio waves, microwaves, infrared

radiation, visible light, ultraviolet light, X rays, and gamma rays. They are all the same thing—electric and magnetic fields that are oscillating in time and space. I am using the word light in a generic sense. It refers to waves of any one of these various frequencies of oscillations of electric and magnetic fields. In this context, if I want to talk about light whose frequency falls in the range to which our eyes are sensitive, I refer to it as visible light. Another name for light is electromagnetic radiation. The entire set of the different frequencies of light is referred to as the electromagnetic spectrum. The following table indicates the way in which humans categorize the various frequencies of light in the electromagnetic spectrum. While I do give definite values, boundaries separating one frequency from the next are not well defined and hence, should be treated as approximate values.

Kind of Light	***Frequency***	***Wavelength***
Radio Waves	< 300 MHz	> 1 m
Microwaves	300 - 750 000 MHz	. 4 mm - 1 m
Infrared	750 GHz - 430 THz	700 nm- .4 mm
Visible	430 - 750 THz	400 - 700 nm
Ultraviolet	750 - 6 000 THz	5 - 400 nm
X rays	6 000 - 50 000 000 THz	.006 - 5 nm
Gamma Rays	> 50 000 000 THz	< .006 nm

Note that the visible regime is but a tiny slice of the overall electromagnetic spectrum. Within it, red light is the long-wavelength, low-frequency visible light, and, blue/violet light is the short-wavelength high-frequency visible light. AM radio stations broadcast in the kHz range and FM stations broadcast in the MHz range. For instance, setting your AM dial to 100 makes your radio sensitive to radio waves of frequency 100 kHz and wavelength 3000 m. Setting your FM dial to 100 makes your radio sensitive to radio waves of frequency 100 MHz and wavelength 3 m.

We call the superposition of the changing electric and magnetic field vectors, with other changing electric and magnetic field vectors, interference. Many of the phenomena involving light are understood in terms of interference.

When light interacts with matter, its electric field exerts forces on the charged particles that make up matter. The direction of the force exerted on a charged particle is the same direction as the electric field if the particle is positive, and in the opposite direction if it is negative. The magnetic field exerts a torque on the magnetic-dipole-possessing particles that make up matter. Because so many of the observable effects associated with the interaction of visible light have to do with the force exerted on charged particles by the electric field, it has become customary to talk about the interaction of light with matter in terms of the interaction of the electric field with matter. I will follow that custom. Please keep in mind that the magnetic field, always at right angles to the electric field in light, is also present. As a result of the force exerted on the charged particles by the electric field of which the light consists, the charged particles accelerate, and, as a result, produce their own electric and magnetic fields. Because there is no time delay between the exertion of the force and the resulting acceleration, the newly produced electric and magnetic field vectors superpose with the very electric and magnetic field vectors causing the acceleration. Because the mass of an electron is approximately 1/2000 of the mass of a proton, the acceleration experienced by an electron is 2000 times greater than that experienced by a proton

subject to the same force. Hence, the interaction of light with matter, can often be explained in terms of the interaction of light with the *electrons* in matter.

How the electrons in matter interact with light, is largely determined by the degree to which the electrons are bound in the matter. As a rather bizarre example of how a large number of complicated interactions can combine to form a simple total effect, the mix of attractive and repulsive $1/r^2$ Coulomb forces exerted on the electron in a solid material by the protons and electrons on all sides of it, results in a net force on the electron that is well modeled by the force that would be exerted on a particle "tied" to its equilibrium position by a spring. Hence the electron acts like a "mass on a spring." As such, it can undergo simple harmonic motion like a mass on the end of a spring. The way light interacts with the electrons can thus be said to depend on the frequency of the light and the force constant of the effective spring. If we limit our discussion to visible light, the degree to which the electrons are bound (the spring constant) determines how the light interacts with the matter. In the case of what we would consider opaque light-absorbing matter such as flat black paint, the electron accelerations result in destructive interference of the incoming light with the light produced by the electrons. Light doesn't go through, nor is much reflected off the material. In the case of shiny metal surfaces, the electrons that the light interacts with are virtually free. The light emitted by these electrons as a result of the acceleration caused by the light, interferes constructively with the light in a very specific backward direction and destructively in forward directions. Hence, the light does not get through the metal, but, it is reflected in a mirror-like manner referred to as *specular reflection.* In the case of a transparent medium such as glass, the light given off by the electrons interferes with the incoming light in such a manner as to cause constructive interference in specific forward and backward directions. But, the constructive interference in the forward direction is such that the pattern of electric and magnetic waves formed by all the interference taken as a whole, moves more slowly through the glass than light moves through vacuum. We say that the speed of light in a transparent medium is less than the speed of light in vacuum. The ratio of the speed of light in vacuum to the speed of light in a transparent medium is called the index of refraction, n, of that transparent medium.

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

where:

n is the index of refraction of a transparent medium,
$c = 3.00 \times 10^8$ m/s is the speed of light in vacuum, and,
v is the speed of light in the transparent medium.

Because the speed of light never exceeds the-speed-of-light-in-vacuum, the index of refraction is always greater than or equal to 1 (equal when the medium is, or behaves as, vacuum). Some values for the index of refraction of light for a few transparent media are:

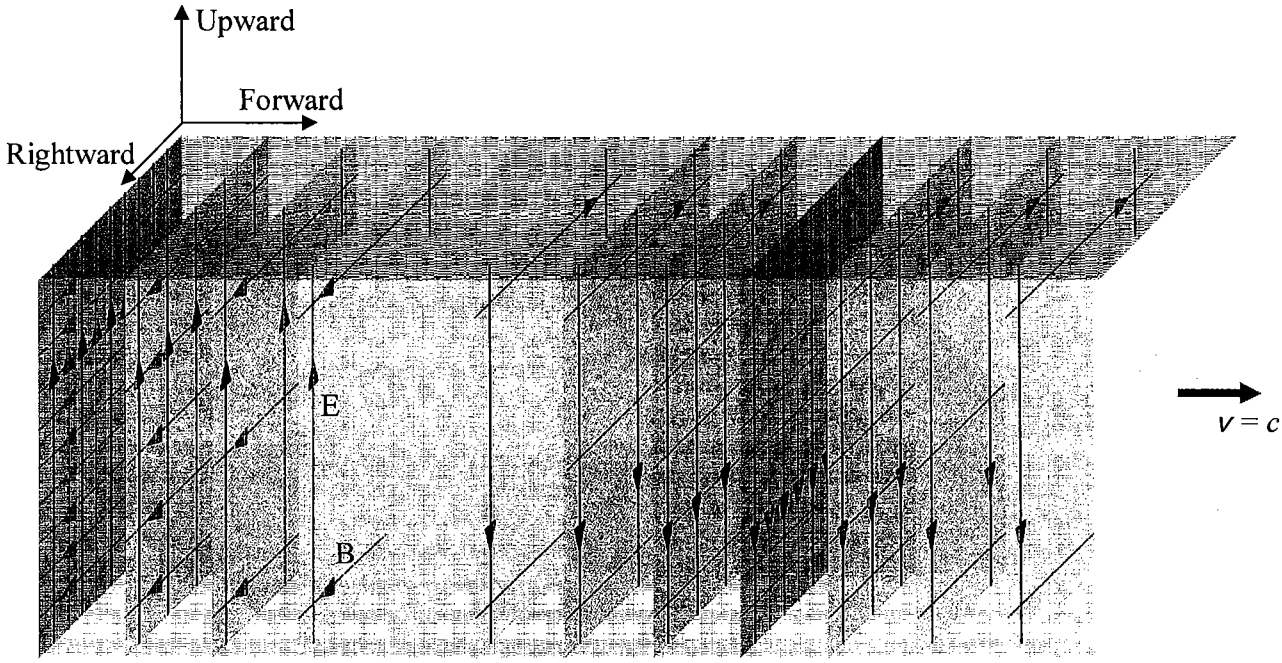
Medium	***Index of Refraction***
Vacuum	1
Air	1.00
Water	1.33
Glass (Depends on the kind of glass. Here is one typical value.)	1.5

The kind of interaction of light with matter with which we are most familiar is called diffuse reflection. It is the light that is diffusely reflecting off a person that enters your eyes when you are looking at that person. The electron motion produces light that interferes destructively with the incoming light in the forward direction (the direction in which the incoming light is traveling), so, essentially none gets through but, for a particular frequency range, for all backward directions, very little destructive interference occurs. When the object is illuminated by a mix of all visible frequencies (white light), the frequency of the reflected light depends on the force constant of the effective spring that is binding the electrons to the material of which they are a part. The frequency reflected (in all backward directions) corresponds to what we call the color of the object.

22 Huygens's Principle and 2-Slit Interference

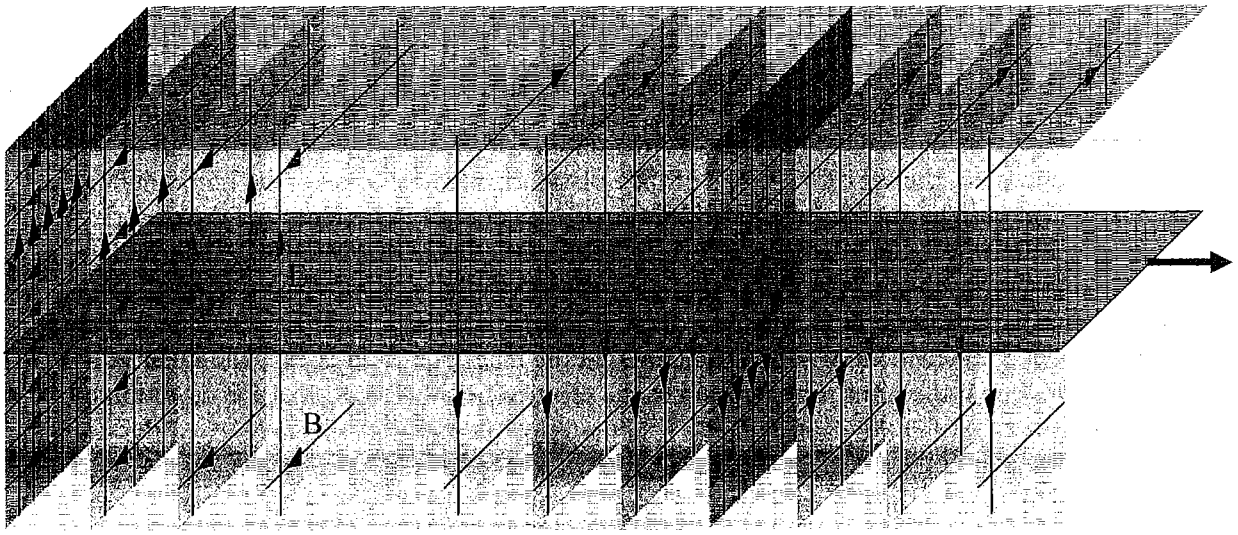
Consider a professor standing in front of the room holding one end of a piece of rope that extends, except for sag, horizontally away from her in what we'll call the forward direction. She asks, "What causes sinusoidal waves?" You say, "Something oscillating." "Correct," she replies. Then she starts moving her hand up and down and, right before your eyes, waves appear in the rope. For purposes of discussion, we will consider the waves only before any of them reach the other end, so we are dealing with traveling waves, not standing waves. "What, specifically, is causing these waves," she asks, while pointing, with her other hand, at the waves in the rope. You answer that it is her hand oscillating up and down that is causing the waves and again you are right. Now suppose you focus your attention on a point in the rope, call it point P, somewhat forward of her hand. Like all points in the rope where the wave is, that point is simply oscillating up and down. At points forward of point P, the rope is behaving just as if the professor were holding the rope at point P and moving her hand up and down the same way that point P is actually moving up and down. Someone studying only those parts of the rope forward of point P would have no way of knowing that the professor is actually holding onto the rope at a point further back and that point P is simply undergoing its part of the wave motion caused by the professors hand at the end of the rope. For points forward of point P, things are the same as if point P were the source of the waves. For predicting wave behavior forward of point P, we can treat point P, an oscillating bit of the rope, as if it were the source of the waves. This idea that you can treat one point in a wave medium as if it were the source of the waves forward of it, is called Huygens's Principle. Here, we have discussed it in terms of a one dimensional medium, the rope. When we go to more than one dimension, we can do the same kind of thing, but we have more than one point in the wave medium contributing to the wave behavior at forward points. In the case of light, that which is oscillating are the electric field and the magnetic field. For smooth regular light waves traveling in a forward direction, if we know enough about the electric and magnetic fields at all points on some imaginary surface through which all the light is passing, we can determine what the light waves will be like forward of that surface by treating all points on the surface as if they were point sources of electromagnetic waves. For any point forward of the surface, we just have to (vectorially) add up all the contributions to the electric and magnetic fields at that one point, from all the "point sources" on the imaginary surface. In this chapter and the next, we use this Huygens's Principle idea in a few simple cases (e.g. when, except for two points on the kind of surface just mentioned, all the light is blocked at the surface so you only have two "point sources" contributing to the electric and magnetic fields at points forward of the surface—you could create such a configuration by putting aluminum foil on the surface and poking two tiny holes in the aluminum foil) to arrive at some fairly general predictions (in equation form) regarding the behavior of light. The first has to do with a phenomenon called two-slit interference. In applying the Huygens's principle idea to this case we use a surface that coincides with a wave front. In diagrams, we have some fairly abstract ways of representing wave fronts which we share with you by means of a series of diagrams of wave fronts, proceeding from less abstract to more abstract.

Here's one way of depicting a portion of a beam of light traveling forward through space, at an instant in time:

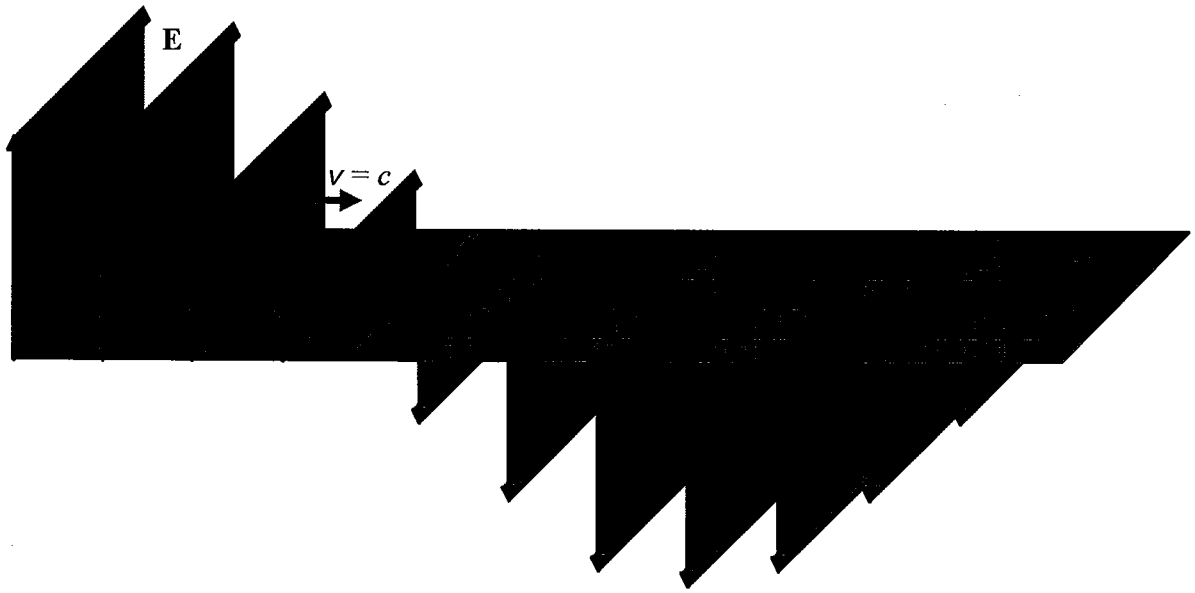


Each "sheet" in the diagram characterizes the electric (vertical arrows) and magnetic (horizontal arrows) fields at the instant in time depicted. On each sheet, we use the field diagram convention with which you are already familiar—the stronger the field, the more densely packed the field lines in the diagram.

Consider the one thin horizontal slice of the beam depicted in this diagram:



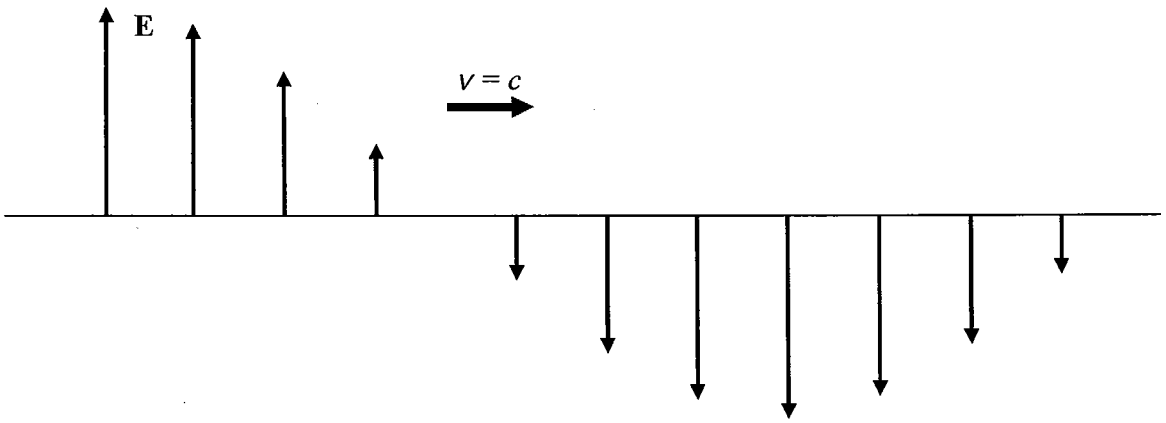
Here we depict the electric field on that one thin horizontal region:



This is a simpler diagram with less information which is, perhaps, easier to interpret, but, also, perhaps, easier to misinterpret. It may for instance, be more readily apparent to you that what we are depicting (as we were in the other diagram) is just $\frac{3}{4}$ of a wavelength of the wave. On the other hand, the diagram is more abstract—the length of the electric field vectors does not represent extent through space, but rather, the magnitude of the electric field at the tail of the arrow. As mentioned, the set of locations of the tails of the arrows, namely the horizontal plane, is the only place the diagram is giving information. It is generally assumed that the electric field has the same pattern for some distance above and below the plane on which it is specified in the diagram. Note the absence of the magnetic field. It is up to the reader to know that; as part of the light, there is a magnetic field wherever there is an electric field, and that, the greater the electric field, the greater the magnetic field, and that the magnetic field is perpendicular both to the direction in which the light is going, and to the electric field. (Recall that the direction of the magnetic field is such that the vector $\vec{\mathbf{E}} \times \vec{\mathbf{B}}$ is in the same direction as the velocity of the wave.)

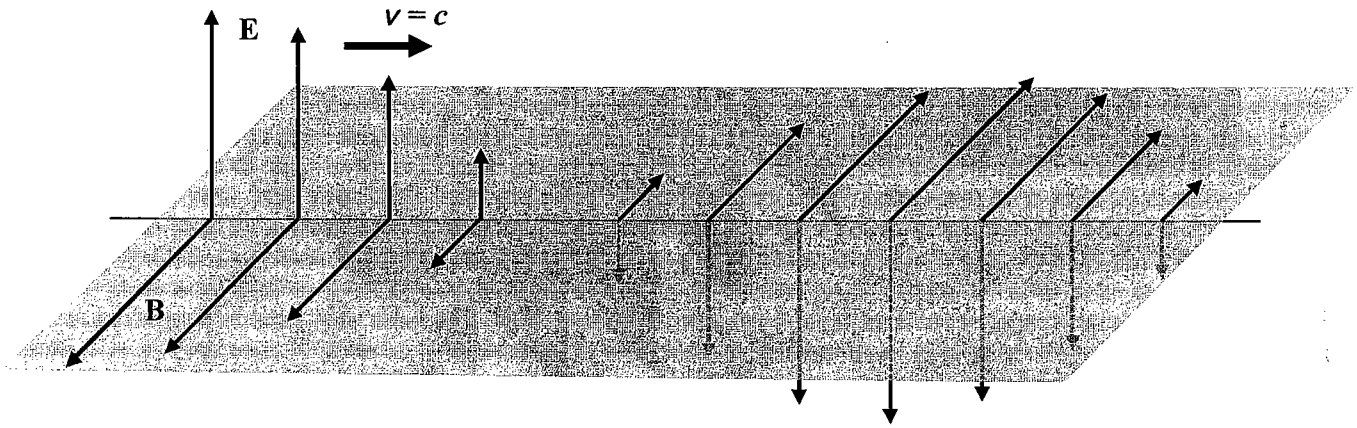
What is perhaps the most common graphical representation of a wave traveling to the right, characterizes the electric and magnetic fields along but a single line extending along the center of the beam in the direction of travel. An example of such a line would be the line along the center of the plane in the following copy of the diagram we have been working with:

An example of an electric field depiction, on a single line along the direction of travel, would be:

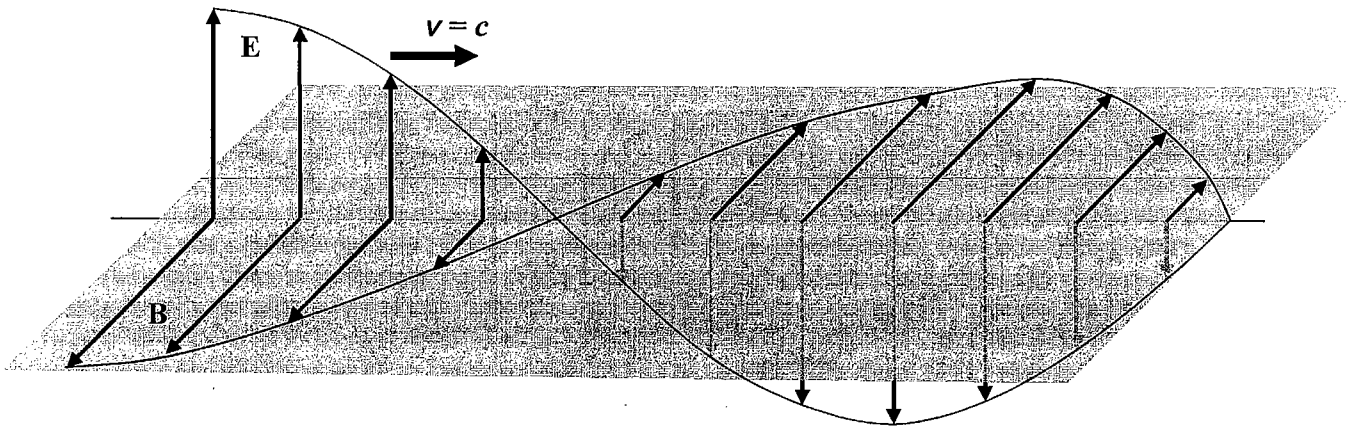


It is important to keep in mind that the arrows are characterizing the electric field at points on the one line, and, it's the *length* of the arrow that indicates the strength of the electric field at the tail of the arrow, *not* the spacing between arrows.

The simplicity of this field-on-a-line diagram allows for the inclusion of the magnetic field vectors in the same diagram:

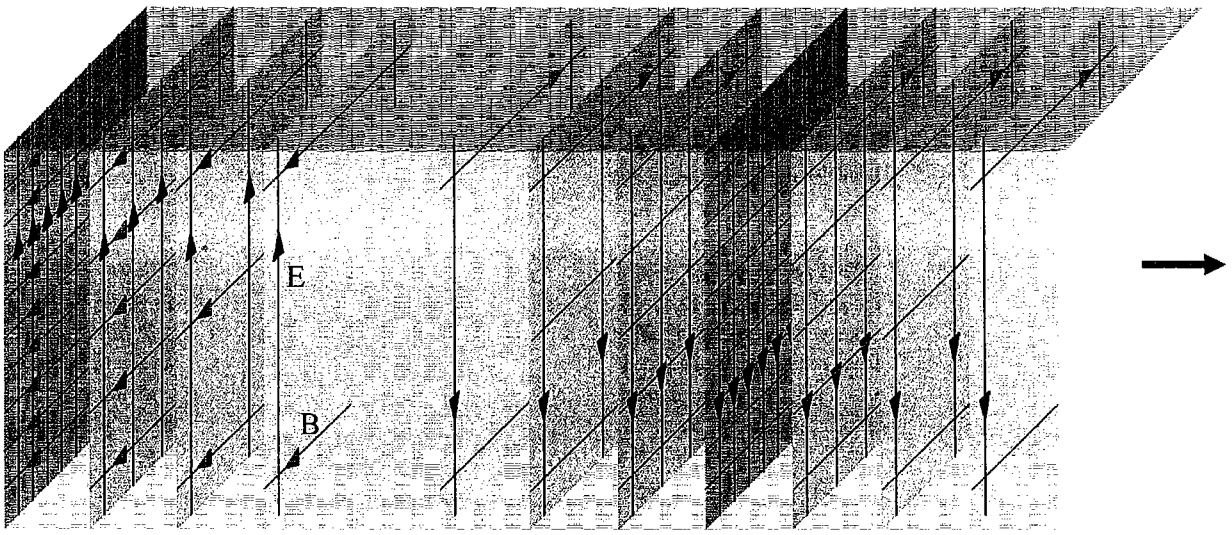


If one connects the tips of the arrows in this kind of diagram, the meaning of those Electric Field vs. Position sinusoidal curves presented in the last chapter becomes more evident:



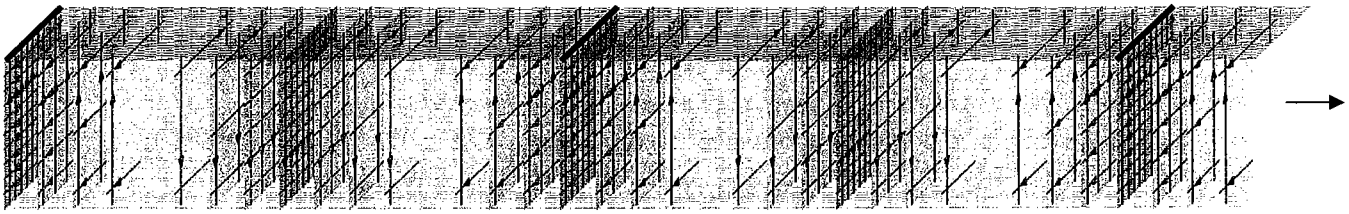
Huygens' Principle involves wavefronts. A wavefront is the part of a wave which is at a surface that is everywhere perpendicular to the direction in which the wave is traveling. If such surfaces are planes, the wave is called a *plane wave*. The kind of wave we have been depicting is a plane

wave. The set of fields on any one of the gray "sheets" on in the diagram:

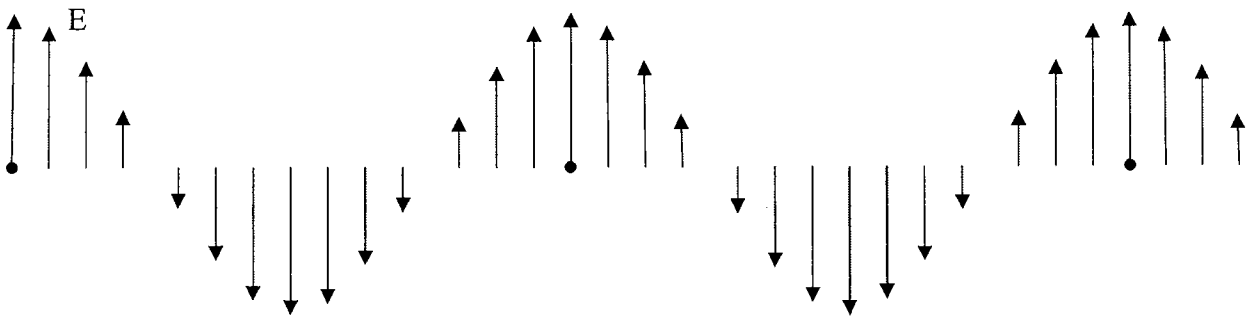


is a part of a wavefront. It is customary to focus our attention on wavefronts at which the electric and magnetic fields are a maximum in one direction. The rearmost sheet in the diagram above represents such a wavefront.

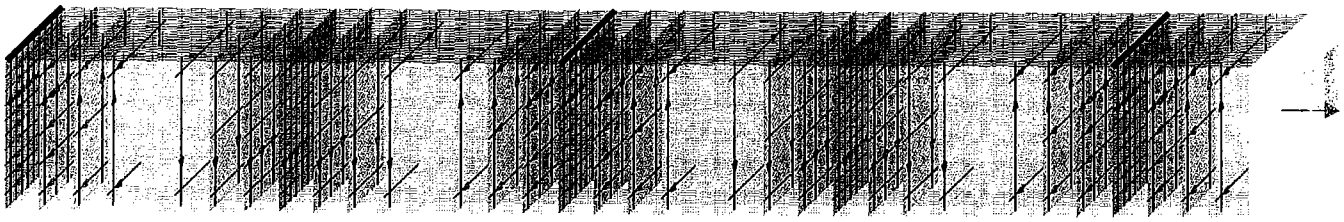
In the following diagram, you see black lines on the top of each sheet representing maximum-upward-directed-electric-field wavefronts.



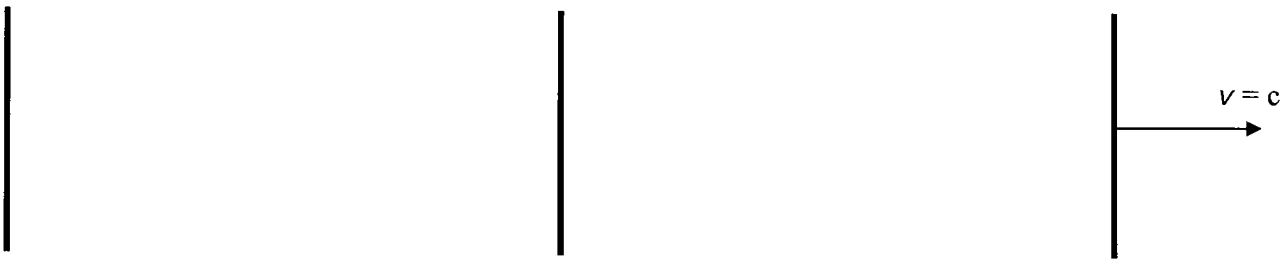
And, in the following, each such wavefront is marked with a black dot:



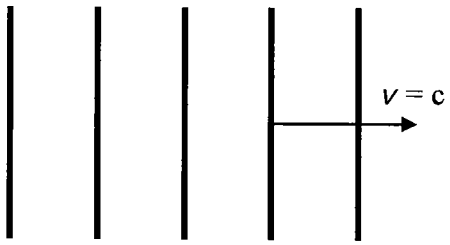
A common method of depicting wavefronts corresponds to a view from above, of the preceding electric field sheet diagram which I copy here:



Such a wavefront diagram appears as:



More commonly, you'll see more of them packed closer together. The idea is that the wavefronts look like a bird's eye view of waves in the ocean.



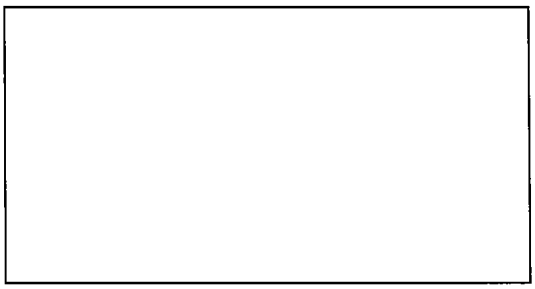
Note that the distance between adjacent maximum-field wavefronts, as depicted here, is one wavelength.

At this point we are ready to use the Huygens's Principle idea, the notion of a wavefront, and our understanding of the way physicists depict wavefronts diagrammatically, to explore the phenomenon of two slit interference.

Two-Slit Interference

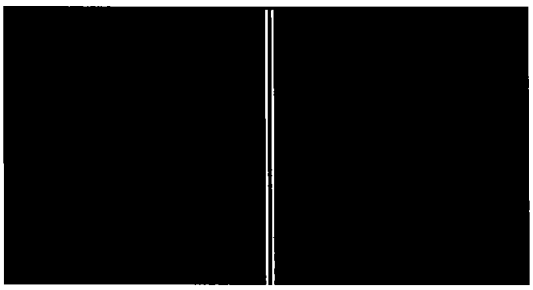
When you shine plane-wave visible light through a mask with two parallel slits cut into it, onto a screen, under certain circumstances (which we shall discuss) you see, on the screen, an extended pattern of bright and dark bands where you might expect to see two bright lines surrounded by darkness which we might call the shadow of the mask.

I use the kind of double slit mask that students make in a double-slit laboratory exercise to convey what we mean by a mask with a double slit in it. We obtain rectangular pieces of thin plate glass of about the size depicted in the following diagram:

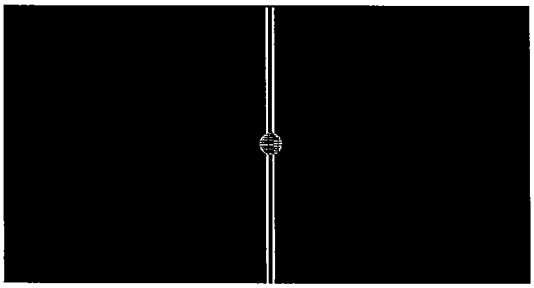


Prior to the laboratory session, we spray paint each mask with flat black spray paint which is given ample time to dry. Each student is given a razor-blade knife, a metal ruler, and a painted piece of glass that looks like this:

The students use the ruler and knife to etch two parallel lines in the paint. After they are done their masks looks like this:

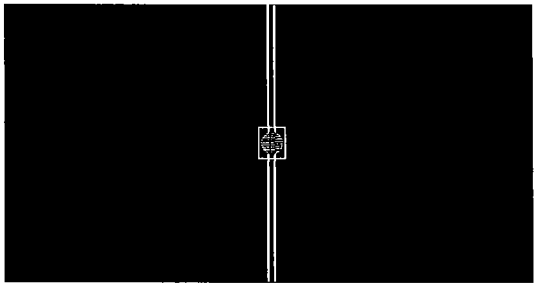


The double slit is illuminated with light from a laser. The laser beam hits the mask "head on". More specifically, the direction in which the light is traveling is at right angles to the surface of the mask. Such light is said to be *normally incident* upon the mask. (Recall that the word "normal" means "at right angles to." Make sure you know what we mean by the statement that light is *normally incident* upon a mask with a double slit in it.)

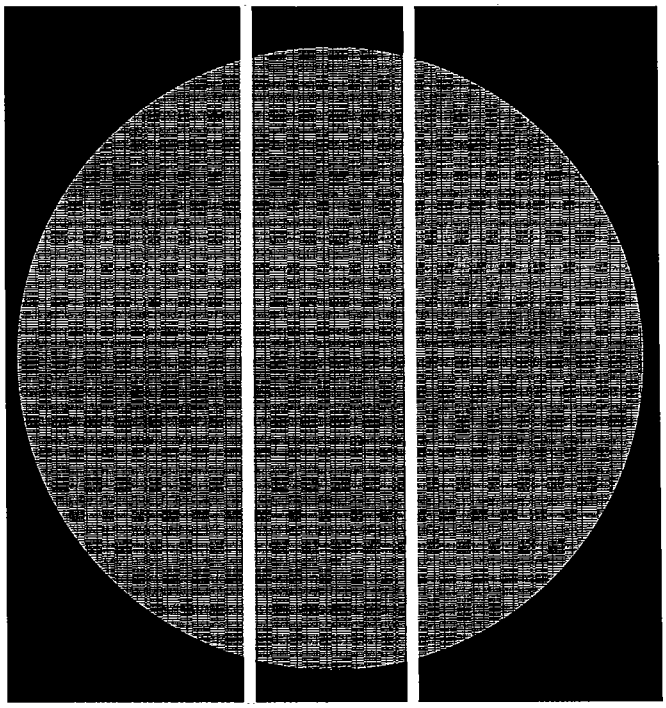


On a white paper screen behind the mask, aligned the same way the mask is, the students see a pattern of bright and dark bands distributed horizontally over the face of the screen. The bright bands are typically referred to as *fringes*. In cases (unlike the laboratory exercise under discussion) where the width of each fringe is small compared to its height, the fringes are often referred to as *lines*.

Huygens' Principle helps us understand this phenomenon. The phenomenon is an interference phenomenon. Aside from the need for the slits to be close enough together for the beam to "hit" both slits at the same time, the slits, for reasons that will soon be clear, must be close to each other. In analyzing the phenomenon here, I am going to "zoom in" on the small piece of the mask outlined in the following diagram:



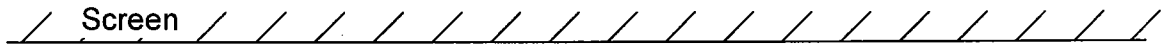
Upward ↑



← Leftward

Rightward →

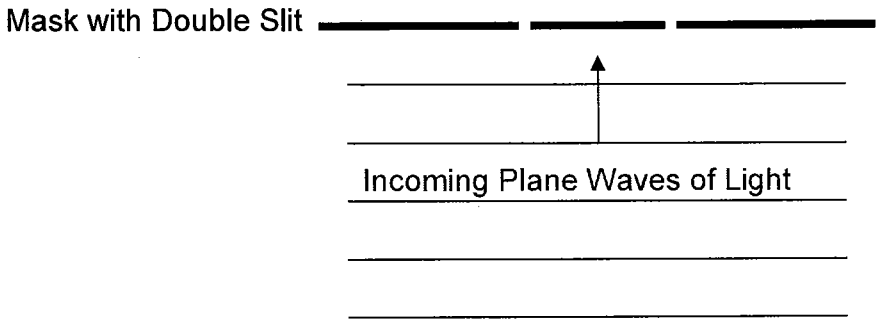
I'm going to call the direction in which the light is traveling, on its way to the mask (which is into the page in the diagram above) the *forward* direction. Now, I am going to depict the entire situation as viewed from above from above.



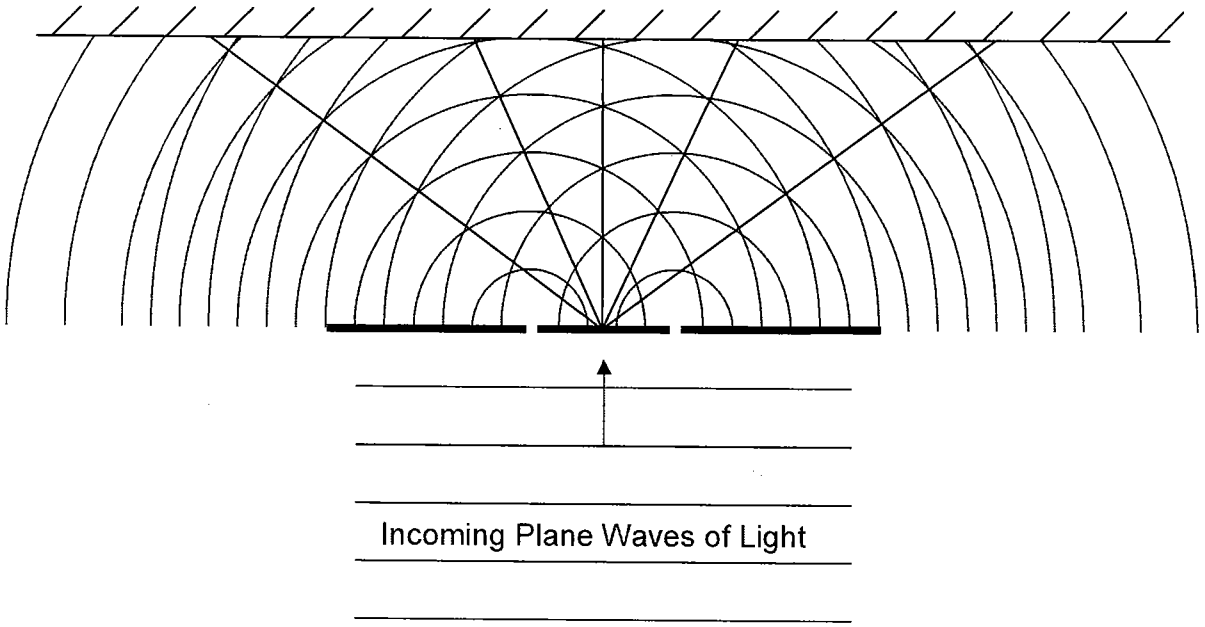
Forward ↑

← Leftward

Rightward →



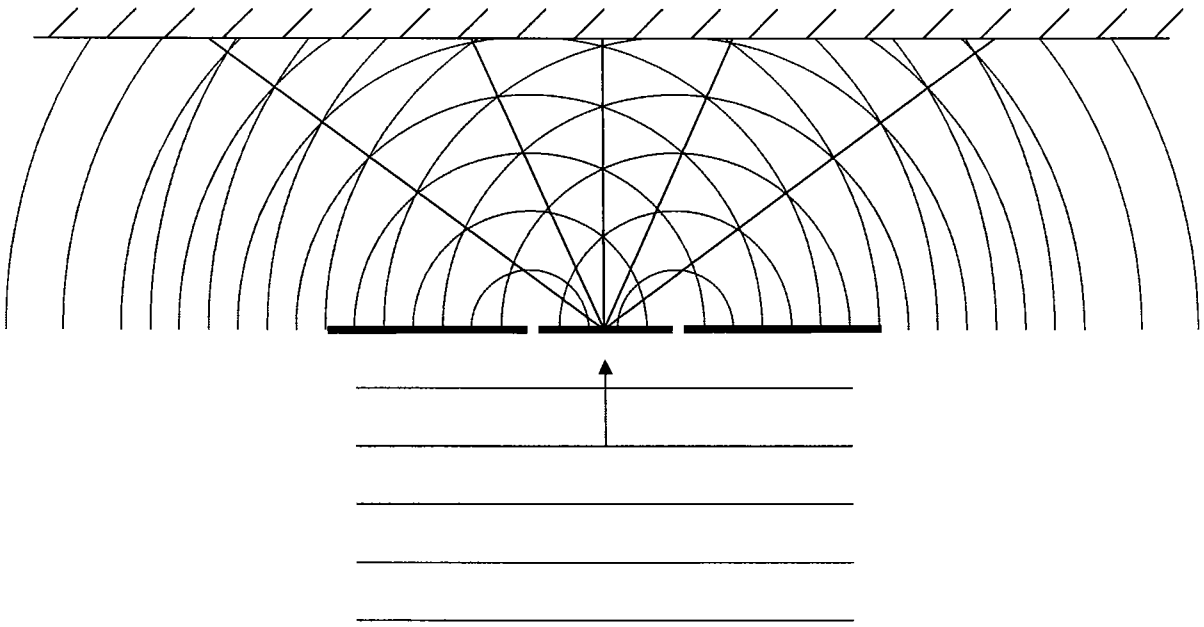
Treating the interface of each slit as a point source in accord with Huygens' Principle, and representing the intersections of the spherical wavefronts with the plane of the page, we have:



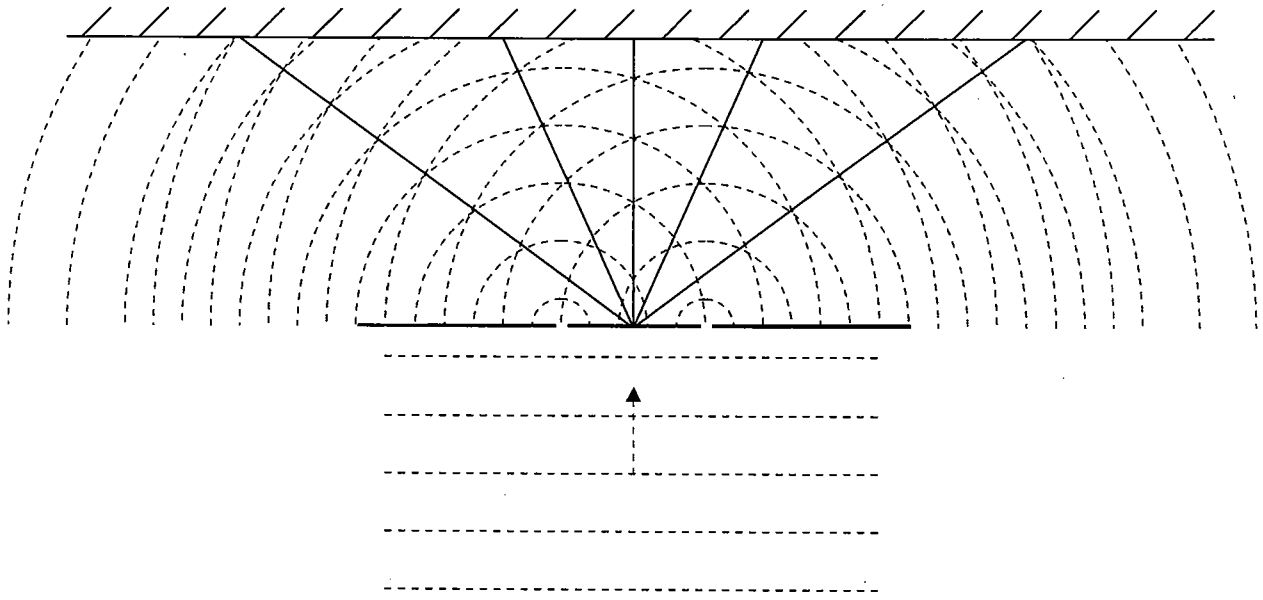
The line and circle segments representing the wavefronts correspond to points in space where, at the instant in question, the electric field is maximum in one particular direction. While the direction can be any direction perpendicular to the direction in which the waves are traveling, to make the discussion easier, let's assume that the electric field oscillations are vertical (into and out of the page) and that the wavefronts represent points in space at which, at the instant in question, the electric field is maximum *upward.* At those points in space forward of the source, the electric field due to each point-like source varies periodically from upward and maximum (call it E_{max} , see footnote[1]), to zero, to downward and maximum ($-E_{max}$), back to zero, and back to maximum upward (E_{max}), with a frequency that we call the frequency of the light waves. At those points in space forward of the source, the total electric field is the sum of the contribution to the electric field from the left slit and the contribution to the electric field from the right slit. Thus, for the instant depicted in the diagram, where the wavefronts cross, we have both contributions at a maximum in one and the same direction. So the total electric field at such points is, at the instant in question, twice that due to either source. As time elapses, the electric field varies at such points, from $2E_{max}$ to 0 to $-2E_{max}$ to 0, and back to $2E_{max}$; repeatedly. The intensity of the light at such a position is 4 times what the intensity due to either slit alone would be. The superposition of the time-varying electric fields is called interference, and when the various contributions add together to form an electric field that varies with an amplitude that is bigger than the electric field due to any individual contributor, we call the interference constructive.

[1] At any position through which the light under discussion travels, the electric field of that light varies sinusoidally with time. I'm using the symbol E_{max} to represent the amplitude of the oscillations due to one source at such a point. This amplitude is different at different locations, both because of diminishment with distance from the source, and, because both of the point-like sources under discussion here are actually infinite sets of point sources themselves.

Note that the points where the wavefronts cross lie on, or very near, straight lines extending out from the midpoint between the slits:



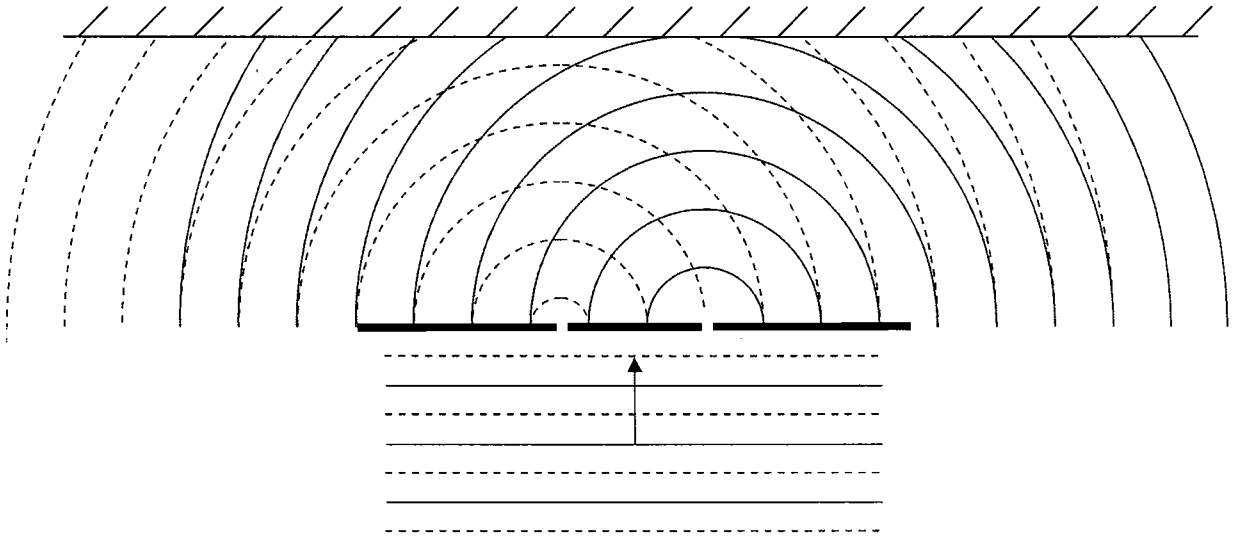
The big deal is, that at every point on those lines (except "too" close to the slits) the electric field is oscillating with an amplitude of $2E_{max}$. If we were to let time continually elapse, you would see those crossing points continually being created near the slits, moving onto the lines and outward along the lines toward the screen. At every point on the lines (except "too" close to the slits), the interference is always constructive at every instant in time. For instance, for the same instant depicted in the diagram above, consider the electric field on the wavefronts midway between the wavefronts depicted. On these "midway between" wavefronts, the electric field is maximum downward. I am going to redraw the wavefront diagram above, minus the maximum-upward (electric field vector) wavefronts, but *with* the maximum-downward wavefronts (depicted with dashed lines). Again, this diagram is for the same instant in time characterized by the diagram above. Also, I am going to leave those same straight lines along which the constructive interference is occurring, on the diagram.



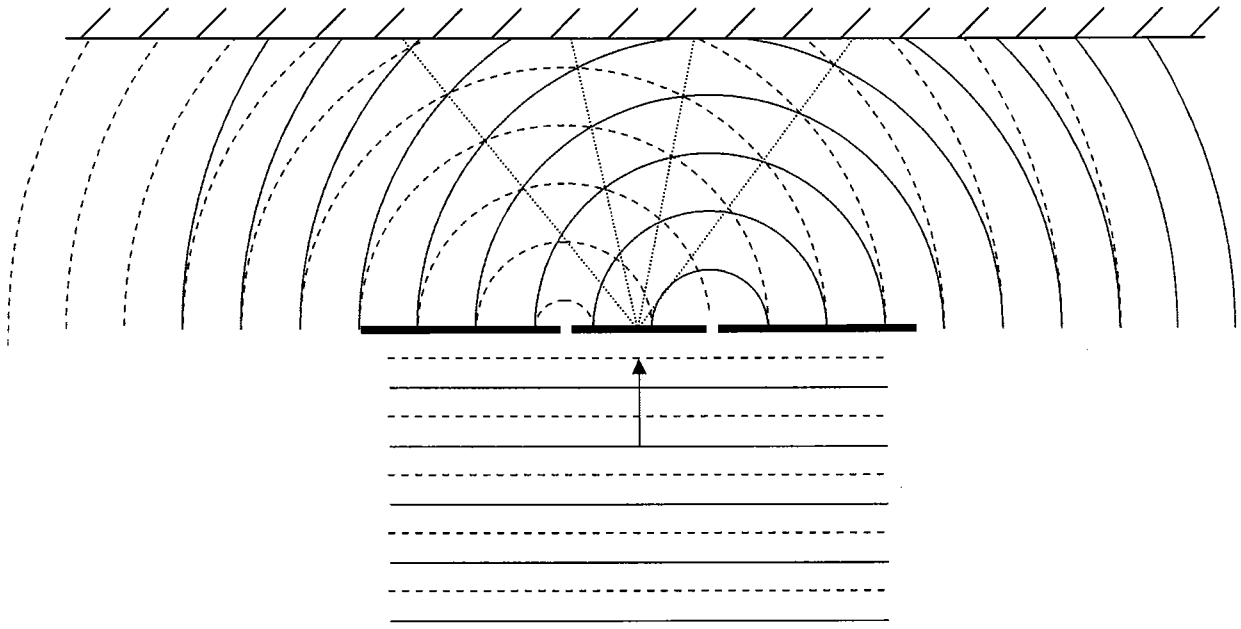
Note that, indeed, (except for "too close" to the slits), the maxima (wavefront crossing points) for the maximum-downward-electric-field vector wavefronts, occur along the *exact same lines* that the maximum-upward-electric-field vector wavefronts occurred along. Again, I want to stress that the amplitude of oscillations is maximal ($2E_{max}$) at every point on each of those lines—at points on the lines other than where the wavefronts are crossing, it is just that the electric field is at different stages of those maximum-amplitude oscillations.

Where the lines consisting of points of maximal constructive interference intersect the screen, we see a bright fringe. How about the dark fringes? Those result at points in space where the electric field from one slit always cancels the electric field from the other slit. We can find such points by depicting the maximum-upward-electric-field wavefronts from one slit at a particular instant, while, on the same diagram, depicting the maximum-downward-electric-field wavefronts from the other slit. Where they cross, we have perfect cancellation, not just for the instant depicted, but, for all time.

Here we show maximum-downward-electric-field wavefronts from the left slit interfering with maximum-upward-electric-field wavefronts from the right slit:

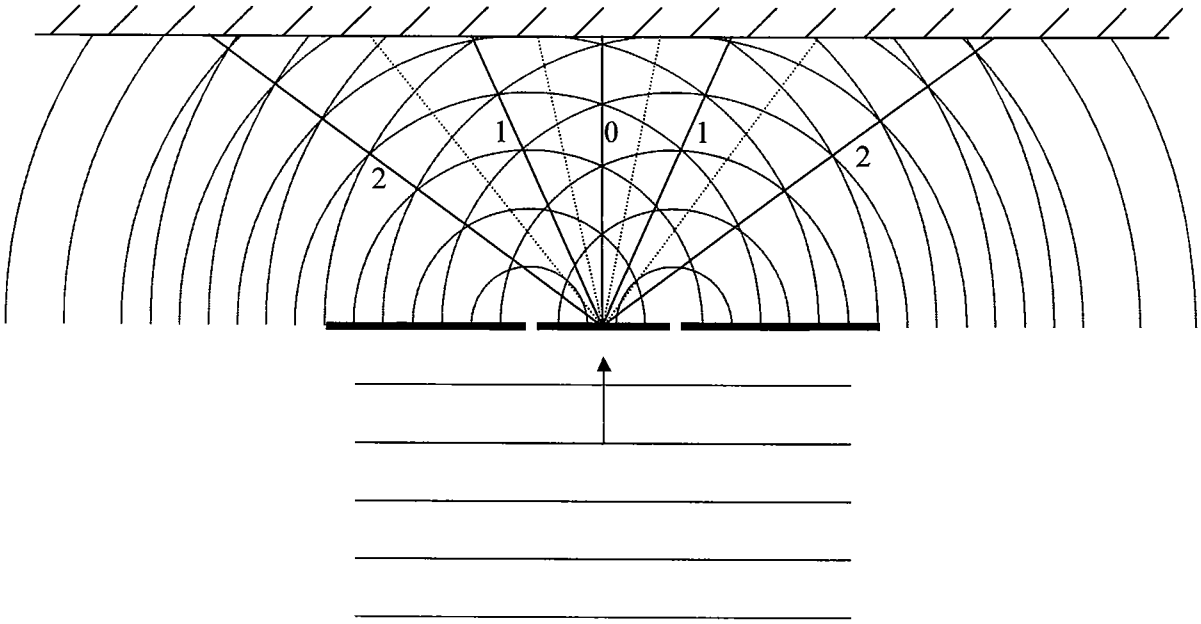


Again, except for very close to the slits, the points of intersection of the oppositely-directed electric field wavefronts, the points of destructive interference, lie on lines (*not* the same lines as before) extending outward from the point on the mask that is midway between the slits:



Along these lines, the interference is always completely destructive. Where they intersect the screen, we see dark (unilluminated) fringes.

If we put these lines on which destructive interference always occurs on the diagram with the lines along which constructive interference always occurs,



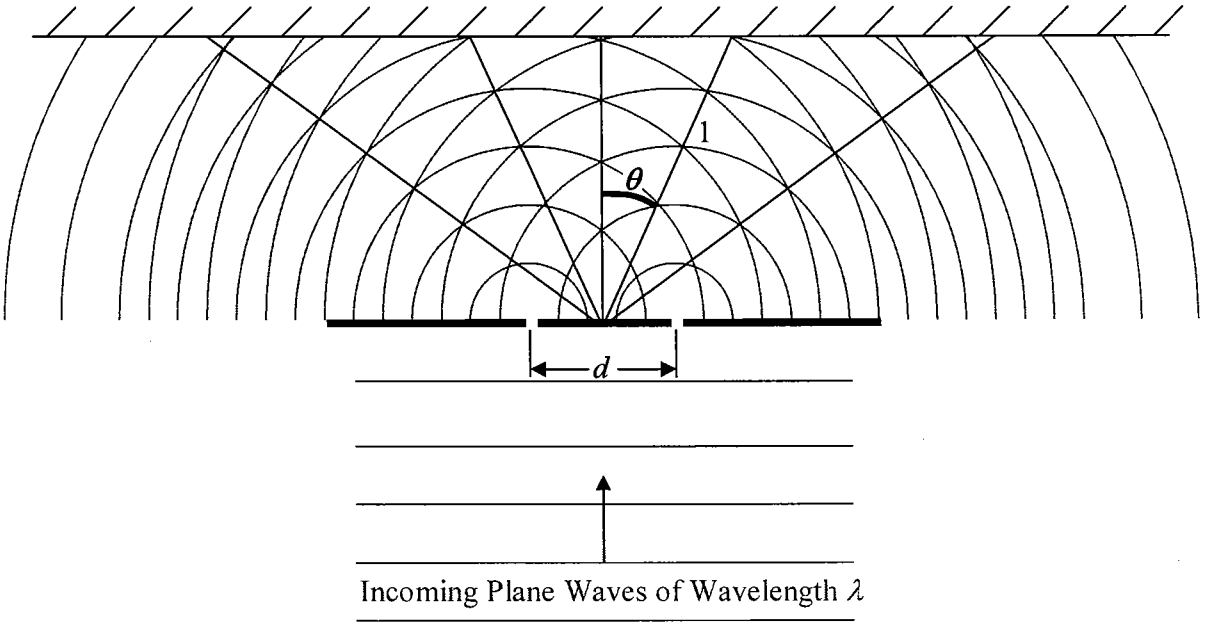
we see that the angles at which destructive interference always occurs look like they are about midway between the angles at which constructive interference always occurs.

Okay, it's time to get quantitative. It's probably pretty obvious to you that the pattern depends on the wavelength of the light and the spacing between the slits. What we need to do to put an end to this chapter is to find a mathematical relationship between the angles at which the maxima (bright light) and minima (darkness, a.k.a. zero light) occur. I am going to use the symbol θ to represent the angle that whichever maximum or minimum we are focusing on at the time, occurs. We number the maxima 0, 1, 2, ... working out from the middle in both directions as indicated on the diagram above. The minima are labeled 1, 2, 3 (there is no "0" minimum), also working out from the middle. To avoid clutter, I have not labeled the minima in the diagram. If necessary, use those numbers as subscripts on the angle θ to distinguish one angle at which a maximum occurs from another, and, to distinguish one angle at which a minimum occurs from another. If it is not clear from the context, the superscripts ^{MAX} and ^{MIN} may also be needed. The angle of a particular maximum or minimum to the left is always the same as the angle of the corresponding maximum or minimum to the right, so we don't need to differentiate between left and right.

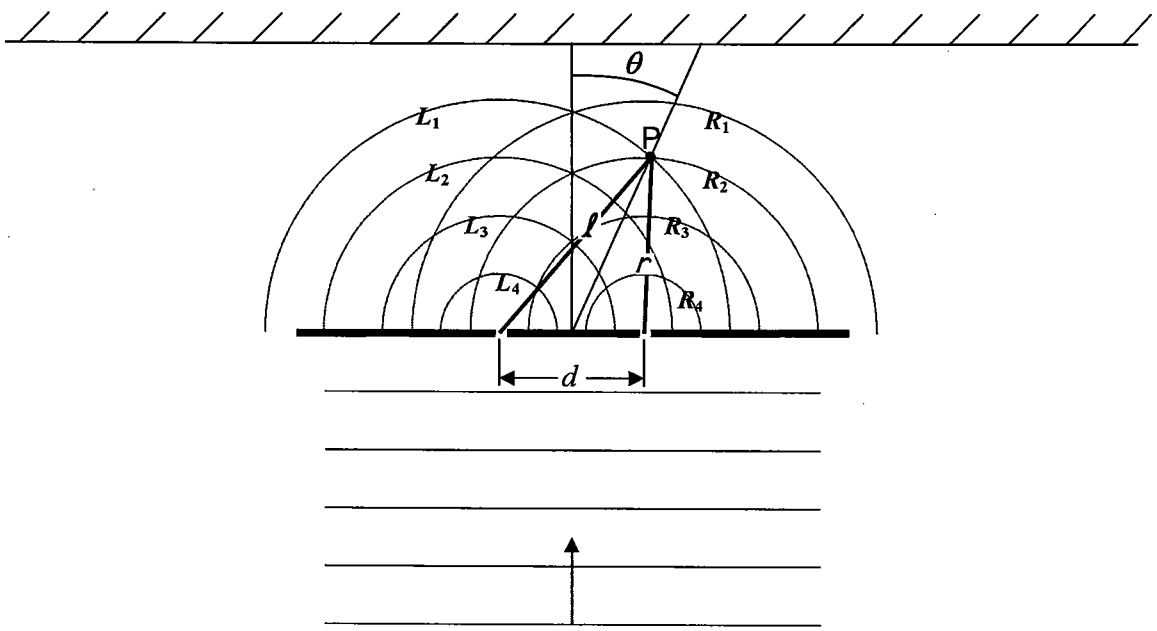
The two slits represent a special pair of sources. What's special about such a pair of sources is that the two sources are exactly in synchronization with each other. That is, for instance, when one is producing a maximum-upward electric field vector, so is the other. We say that the two sources are in *phase* with each other. Another thing that is special about the sources is that, because we consider each slit to be the same size as the other, and the light from the source to be of one and the same intensity across the face of the beam, the amplitude of the oscillations is the

same for each. This is why we get total cancellation along the lines on which the minima occur. By inspection of the diagram, it is clear that, to get to any point on the line in the straight-ahead direction, the distance that the light from one slit travels is the same as the distance that light from the other slit travels. As such, at any point along the center line (the line extending from the midpoint between the two slits, straight ahead) whenever a crest (maximum upward electric field) arrives from one slit, a crest arrives from the other slit; and; whenever a trough (maximum *downward* electric field) arrives from one slit, a trough arrives from the other slit. Thus, on the centerline, the interference is *always* constructive. So, we have one quantitative result already, a maximum (constructive interference, bright fringe) occurs at $\theta = 0°$.

Let's define the center-to-center slit spacing to be d, and, the wavelength of the incoming plane waves to be λ. We'll work on finding an expression for the angle θ at which the first maximum occurs, first.



Let's redraw the diagram with a little less clutter to see what's going on.



I removed some of the spherical wavefronts and labeled the wavefronts from the left slit L_1, L_2, L_3, and L_4 corresponding to the order in which they came from the left slit. Similarly, wavefronts from the right slit are labeled R_1, R_2, R_3, and R_4 corresponding to the order in which they came from the right slit. Furthermore, I have indicated the distance ℓ from the left slit to one point (point P) on the line of maxima, and, the distance r from the right slit to the same point P on the line of maxima.

Look at the diagram and note that, although R_1 and L_1 originate from their respective slits at one and the same instant, as do R_2 and L_2, it is R_2 and L_1 that arrive at point P together. By the time it arrives at point P, L_1 has been traveling for a greater amount of time than R_2 has, but, L_1 arrives at point P at the same time as R_2 does because it (L_2) has a greater distance to travel. This is the whole key to constructive interference. At any point illuminated by two in-phase sources there will be constructive interference if the point is the same distance from both sources, or, when the distance is different, if that difference is one wavelength, two wavelengths, three wavelengths, or, for that matter, any integer number of wavelengths. If the path difference *is* an integer number of wavelengths, then, whatever part of the wave is arriving from one source, the same part of the wave will be arriving from the other source. So, for instance, crest arrives with crest, and trough arrives with trough. Defining

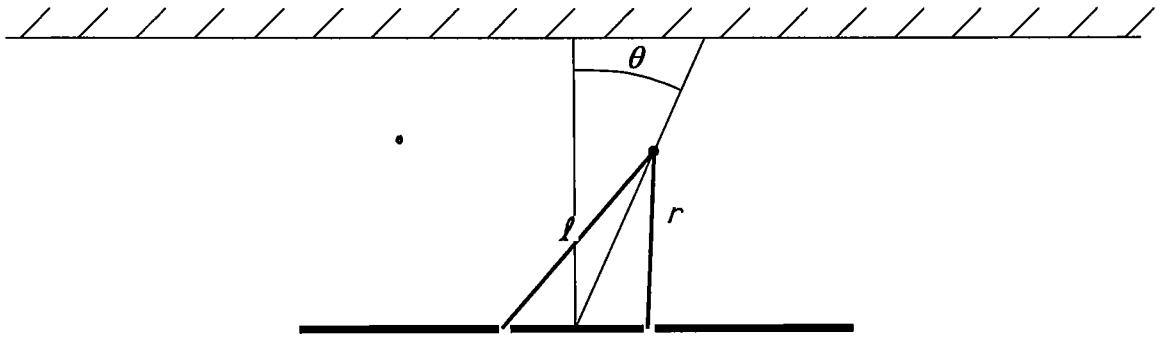
$$\Delta s = \ell - r$$

we have:

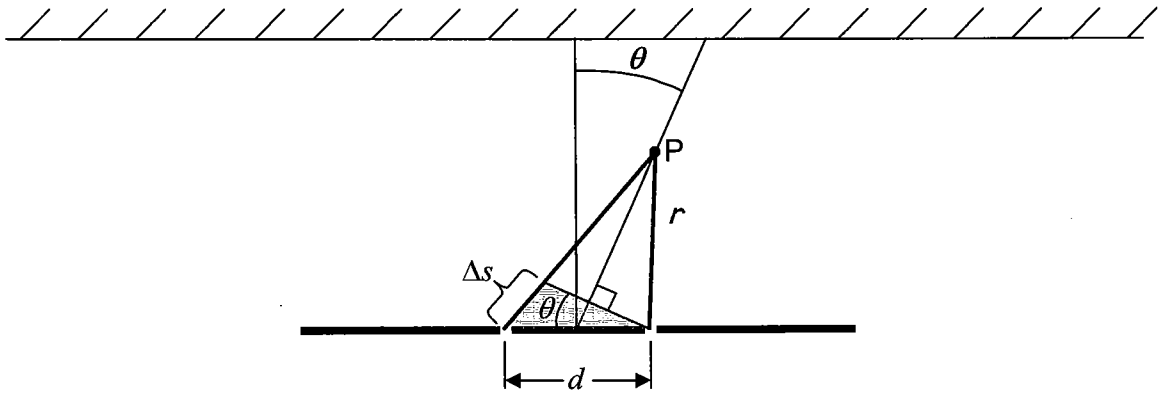
$$\Delta s = m\lambda \qquad (m = 0, 1, 2, \ldots)$$

as our condition for constructive interference.

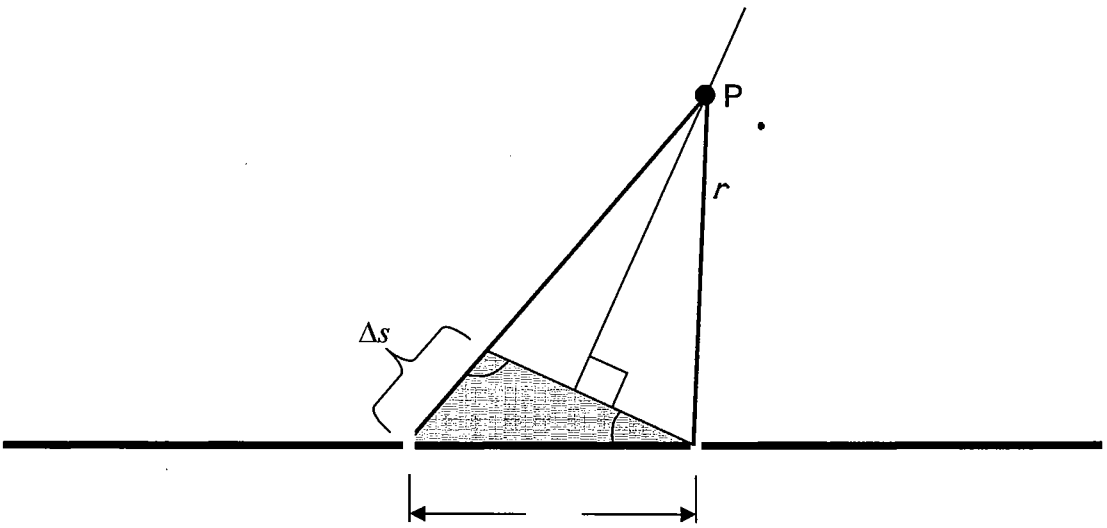
Now I provide the diagram we have been working with, without the wavefronts, so that we can do the geometry needed to relate Δs to the angle θ.



First I want to draw a line segment perpendicular to our line of maximum interference, ending on the right slit.



The new line segment breaks up the path of length ℓ into a part identical in length to the one of length r, and a part whose length is the path-length difference Δs. The new line segment also forms a *triangle*, the one that is shaded in the diagram. Using plane geometry, I have identified the angle labeled θ in the small triangle as the same angle θ that the line of maximal constructive interference makes with the straight-ahead direction. I need to blow up that triangle so that we can analyze it:



Okay, now we make an approximation. See angle ϕ in the shaded triangle? It is approximately equal to 90°. Indeed, the farther point P is from the slits, the closer ϕ is to 90°. All we need is for the distance from the slits to point P to be large compared to the distance between the slits. In practice, this is realized in an actual double-slit experiment. The distance to point P is typically thousands to millions of times greater than the distance between the slits. The approximation in question is thus, typically, a fantastic approximation. Treating ϕ as a right *angle* makes the shaded triangle a right *triangle*, meaning that the path difference can be expressed as:

$$\Delta s = d \sin\theta$$

This is the relation we've been looking for. Combining it with the fact that the path difference has to be an integer number of wavelengths for interference *maxima* we have:

$$m\lambda = d\sin\theta \qquad (m = 0, 1, 2, \ldots) \qquad (22\text{-}1)$$

as the condition for maximum constructive interference.

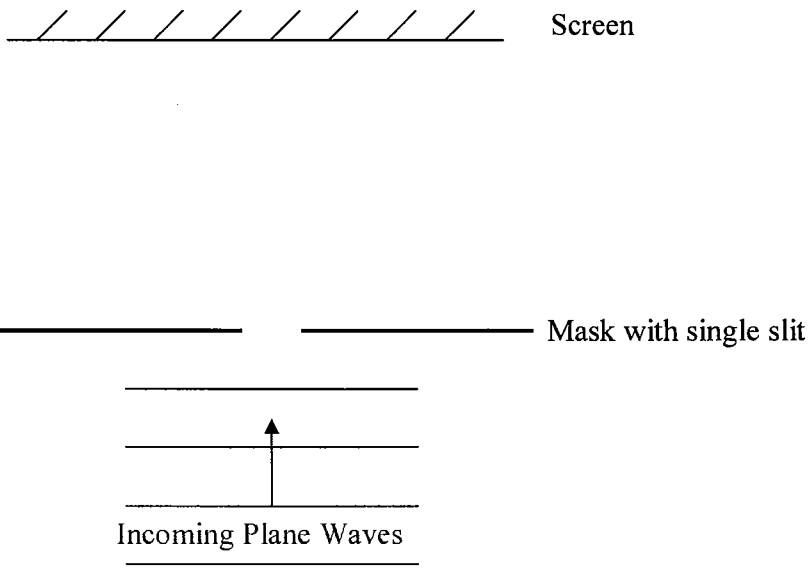
For perfectly *destructive* interference, the path difference must be half a wavelength or, half a wavelength more than any integer number of wavelengths. So, for interference *minima* we have:

$$(m + \tfrac{1}{2})\lambda = d\sin\theta \qquad (m = 0, 1, 2, \ldots) \qquad (22\text{-}2)$$

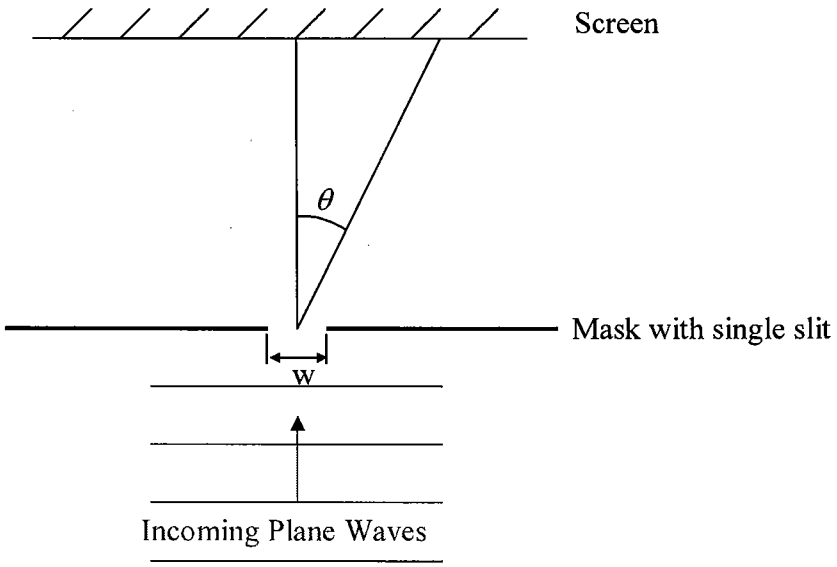
The value of the integer m is referred to as the *order* of interference. Thus, the first maximum to either side of the straight-ahead direction is referred to as the *first order* maximum, the next one is called the *second order* maximum, etc.

23 Single-Slit Diffraction

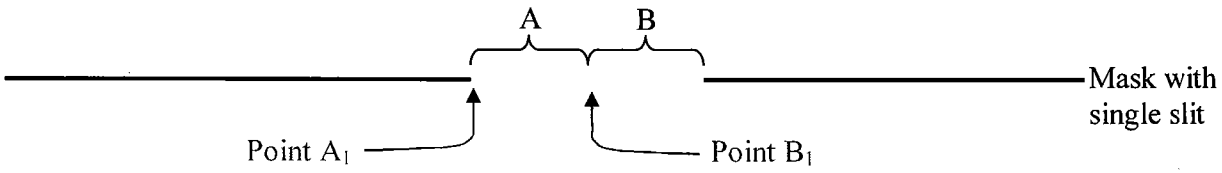
Single-slit diffraction is another interference phenomenon. If, instead of creating a mask with two slits, we create a mask with one slit, and then illuminate it, we find, under certain conditions, that we again get a pattern of light and dark bands. It is not the same pattern that you get for two-slit interference, but, it's quite different from the single bright line in the straight-ahead direction that you might expect. Here's how it comes about. Firstly, here's the setup:



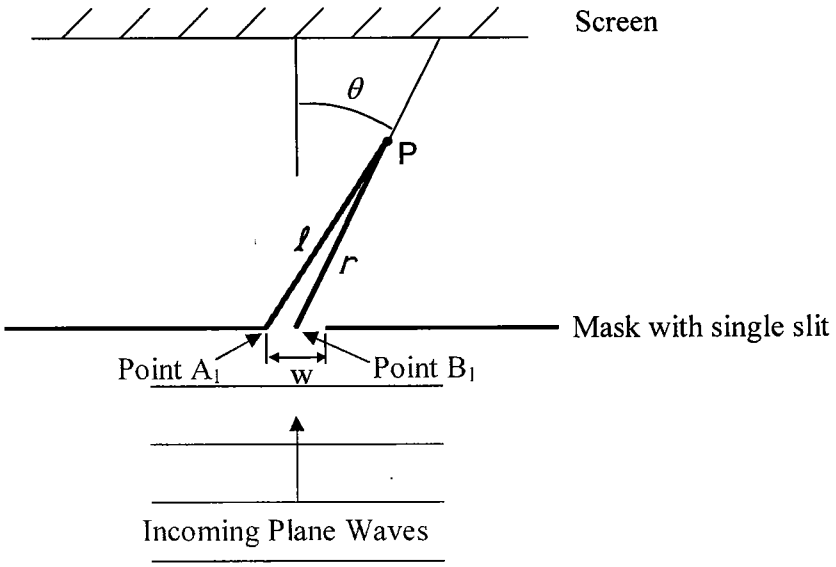
Again, we get a bright fringe in the straight-ahead position on the screen. From there, working out to either side, we get bands that alternate between dark and bright. The first maximum to the right or left of the central maximum is not nearly as bright as the central maximum. And each maximum after that is less bright than the maximum preceding it. As far as the analysis goes, I want to start with the minima. Consider an imaginary line extending out from the midpoint of the slit all the way to the screen.



So now the question is, "Under what conditions will there be completely destructive interference along a line such as the one depicted to be at angle θ above?" To get at the answer, we first divide the slit in half. I'm going to enlarge the mask so that you can see what I mean.



Now I imagine dividing side A up into an infinite number of pieces and side B up the same way. When the slit is illuminated by the light, each piece becomes a point source. Consider the first point source (counting from the left) on side A and the first point source (again counting from the left) on side B. These two point sources are a distance $w/2$ apart, where w is the width of the slit. If the light from these two point sources (which are in phase with each other because they are really both part of the same incoming plane wave), interferes completely destructively, at some angle θ with respect to the straight-ahead direction, then the light from the second point source on side A and the second point source on side B will also interfere with each other completely destructively because these two point sources are also $w/2$ apart. The same goes for the third-from-the-left point sources on both sides, the fourth, the fifth, and so on, ad infinitum. So, all we need is to establish the condition that makes the light from the leftmost point source on side A (overall, the leftmost point of the slit) interfere completely destructively with the leftmost point source on side B (overall, essentially the midpoint of the slit). So, consider any point P on a proposed line of minima.



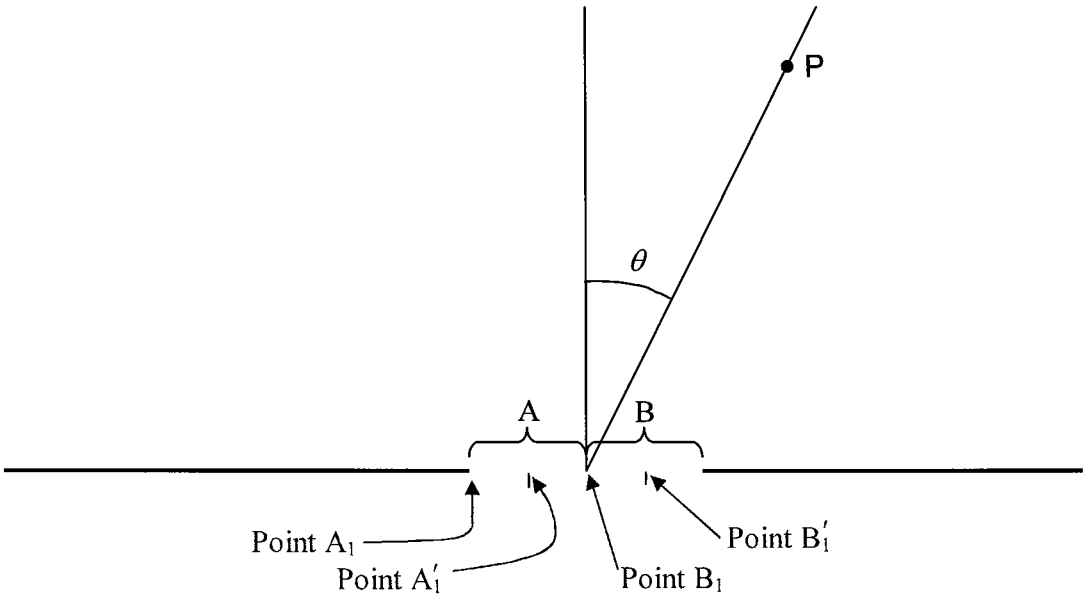
The distance between the two point sources is $w/2$. From the analysis done for the case of two-slit interference, we know that this results in a path difference $\ell - r = \frac{w}{2}\sin\theta$. And, as you know, the condition for completely destructive interference is that the path difference is half a wavelength, or, any integer number of wavelengths plus a half a wavelength. So, we have a minimum along any angle θ (less than 90°) such that:

$$\left(m+\frac{1}{2}\right)\lambda = \frac{w}{2}\sin\theta \qquad (m = 0, 1, 2, \ldots)$$

Now we turn our attention to the question of diffraction *maxima*. I should warn you that this analysis takes an unexpected turn. We do the exact same thing that we did to locate the minima, except that we set the path difference $\ell - r$ equal to an *integer* number of wavelengths (instead of a half a wavelength plus an integer number of wavelengths). This means that the path to point P from the leftmost point on side A (position A_1) of the slit, differs by an integer number of wavelengths, from the path to point P from the leftmost point on side B (position B_1). This will also hold true for the path from A_2 vs. the path from B_2. It will hold true for the path from A_3 vs. the path from B_3 as well. Indeed, it will hold true for any pair of corresponding points, one from side A and one point from side B. So, at point P, we have maximally constructive interference for every pair of corresponding points along the width of the slit. There is, however, a problem. While, for any pair of points, the oscillations at P will be maximal; that just means that P is at an angle that will make the *amplitude* of the *oscillations* of the electric field due to the pair of points maximal. But the electric field due to the pair of points will still be oscillating, e.g. from max up, to 0, to max down, to 0, and back to max up. And, these oscillations will not be in synchronization with the maximal oscillations due to other pairs of points. So, the grand total will not necessarily correspond to an intensity maximum. The big difference between this case and the minima case is that, in contrast to the time varying maximal oscillations just discussed,

when a pair of contributions results in an electric field amplitude of zero, the electric field due to the pair is always zero. It is constant at zero. And, when every pair in an infinite sum of pairs contributes zero to the sum, at every instant in time, the result is zero. In fact, in our attempt to locate the angles at which maxima will occur, we have actually found some more minima. We can see this if we sum the contributions in a different order.

Consider the following diagram in which each half of the slit has itself been divided up into two parts:



If the path difference between "A_1 to P" and "B_1 to P", is one wavelength, then the path difference between "A_1 to P" and "A_1' to P" must be half a wavelength. This yields completely destructive interference. Likewise for the path difference between "B_1 to P" and "B_1' to P". So, for each half of the slit (with each half itself being divided in half) we can do the same kind of pair-wise sum that we did for the whole slit before. And, we get the same result—an infinite number of *zero* contributions to the electric field at P. All we have really done is to treat each half of the slit the way we treated the original slit. For the entire slit we found

$$\left(m+\frac{1}{2}\right)\lambda=\frac{w}{2}\sin\theta \qquad (m=0,\,1,\,2,\,\ldots)$$

Here we get the same result but with w itself replaced by $w/2$ (since we are dealing with half the slit at a time.) So now we have:

$$\left(m+\frac{1}{2}\right)\lambda=\frac{w}{4}\sin\theta \qquad (m=0,\,1,\,2,\,\ldots)$$

Let's abandon our search for maxima, at least for now, and see where we are in terms of our search for minima. From our consideration of the entire slit divided into two parts, we have $\left(m+\frac{1}{2}\right)\lambda = \frac{w}{2}\sin\theta$ which can be written as $(2m+1)\lambda = w\sin\theta$ meaning that we have a minimum when:

$$w\sin\theta = 1\lambda, 3\lambda, 5\lambda, \ldots$$

From our consideration of each half divided into two parts (for a total of four parts) we have $\left(m+\frac{1}{2}\right)\lambda = \frac{w}{4}\sin\theta$ which can be written $(4m+2)\lambda = w\sin\theta$ meaning that we have a minimum when:

$$w\sin\theta = 2\lambda, 6\lambda, 10\lambda, 14\lambda, \ldots$$

If we cut each of the four parts of the slit in half so we have four pairs of two parts, each $\frac{w}{8}$ in width, we find minima at $\left(m+\frac{1}{2}\right)\lambda = \frac{w}{8}\sin\theta$ which can be written $(8m+4)\lambda = w\sin\theta$ meaning that we have a minimum when:

$$w\sin\theta = 4\lambda, 12\lambda, 20\lambda, 28\lambda, \ldots$$

If we continue this process of splitting each part of the slit in two and finding the minima for each adjacent pair, ad infinitum, we eventually find that we get a minimum when $w\sin\theta$ is equal to any integer number of wavelengths.

$$w\sin\theta = 1\lambda, 2\lambda, 3\lambda, 4\lambda, \ldots$$

a result which write as

$$m\lambda = w\sin\theta \qquad (m = 1, 2, 3, \ldots) \tag{23-1}$$

We still haven't found any maxima. The only analytical way to determine the angles at which maxima occur is to do a full-fledged derivation of the intensity of the light as a function of position, and then mathematically solve for the maxima. While, this is not really as hard as it sounds, let's save that for an optics course and suffice it to say that, experimentally, we find maxima approximately midway between the minima. This includes the straight-ahead (0°) direction except that the straight-ahead maximum, a.k.a. the central maximum, is exactly midway between its neighboring minima.

Conditions Under Which Single-Slit Diffraction and Two-Slit Interference Occurs

To see the kinds of interference patterns that we have been talking about in this and the preceding chapter, certain conditions need to be met. For instance, in order to see *one* set of bright fringes in the two-slit interference experiment we need *monochromatic* light. Translated literally, from the Latin, monochromatic means one-color. Monochromatic light is single-frequency light. Strictly monochromatic light is an idealization. In practice, light that is

classified as being essentially monochromatic, actually consists of an infinite set of frequencies that are all very close to the nominal frequency of the light. We refer to the set of frequencies as a band of frequencies. If all the frequencies in the set are indeed very close to the nominal frequency of the light, we refer to the light as narrow-band radiation.

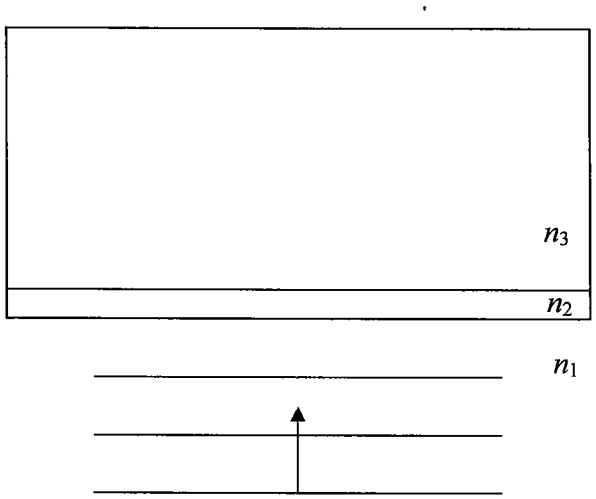
If you illuminate a single or double slit with light consisting of several discrete (individual) wavelengths of light, you get a mix of several interference/diffraction patterns. If you illuminate a single or double slit with a continuum of different frequencies, you find that minima from light of one wavelength are "filled in" by maxima and/or intermediate-amplitude oscillations of light of other wavelengths. Depending on the slit width and (in the case of two-slit interference) slit separation, and the wavelengths of the light, you may see a spectrum of colors on the screen.

In order for the kind of interference that we have been talking about to occur, the light must be coherent. The light must be temporally coherent (coherent with respect to time). While it applies to any part of a wave, I am going to talk about it in terms of crests. In temporally coherent light, one wave crest is part of the same wave that the preceding wave crest is a part of. In light having a great deal of temporal coherence this holds true for thousands of crests in a row. In light with very little temporal coherence this may hold true for only one or two crests in a row. Another way of stating it is to say that light that consists of a bunch of little wave pulses is temporally incoherent and light that consists of long continuous waves is temporally coherent. The long continuous wave can be called a "wave train". In terms of wave trains, light that is temporally incoherent consists of lots of short wave trains, whereas light that is temporally coherent consists of a relatively small number of long wave trains. The kinds of interference we have been talking about involve one part of a wave passing through a slit or slits in a mask, interfering with another part of the same wave passing through the same mask at a later time. In order for the latter to indeed be part of the *same wave*, the light must consist of long wave trains, in other words, it must be temporally coherent. Now, if the wave crest following a given wave crest is not part of the same wave, the distance from one wave crest to the next will be different for different crests. This means the wavelengths are different and hence the frequencies are different. Thus the light is not monochromatic. Under the opposite circumstances, the light is monochromatic. So, monochromatic light is temporally coherent.

The other condition is that the light must be spatially coherent. In the context of light that is normally incident on plane masks, this means that the wavefronts must be plane and they must have extent transverse to the direction in which the light is traveling. In the case of two-slit interference for instance, spatial coherence means that the light at one slit really is in phase with the light at the other slit. In the case of single-slit diffraction, spatial coherence means that light passing through the right half of the slit is in phase with light passing through the left half of the slit.

24 Thin Film Interference

As the name and context imply, thin-film interference is another interference phenomenon involving light. Here's the picture, as viewed from above:



Incoming Plane Waves of Monochromatic Light

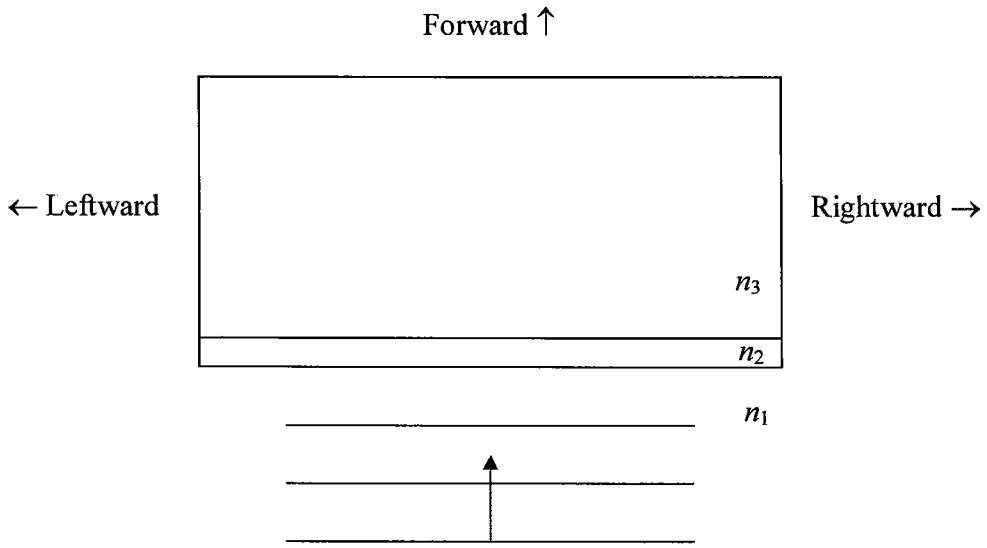
Involved are three transparent media: medium 1, medium 2, and medium 3, of index of refraction n_1, n_2, and n_3, respectively. (In general, a medium is a substance, but, in this context, vacuum is also considered a medium. The index of refraction n of a medium is the ratio of the speed of light in vacuum to the speed of light in that medium.) The phenomenon occurs whether or not $n_1 = n_3$, but, n_2 must be different from n_1 and n_3. Medium 2 is the "thin film." For thin-film interference to occur, the thickness of medium 2 must be on the order of the wavelength of the light. (The actual maximum thickness for which thin-film interference can occur depends on the coherence of the light.)

Here's the deal: Under most circumstances, when light encounters a smooth interface between two transparent media, some of the light goes through (transmitted light) and some of the light bounces off (reflected light). In the thin-film arrangement of three transparent media depicted above, for certain thicknesses of the thin film (medium 2) all the light can be reflected, and, for certain other thicknesses, all the light can be transmitted. You see this phenomenon when looking at soap bubbles, and sometimes when looking at puddles in the road (when there is a thin layer of oil on top of the water). Humans take advantage of the phenomenon by putting a thin coating of a transparent substance on lenses such as camera lenses and binocular lenses, a layer of just the right thickness for maximum transmission.

Based on the situations in which it occurs, it should be clear that we do not need monochromatic light to make thin-film interference happen. However, I am going to discuss it in terms of monochromatic light to get the idea across. Once you understand it in terms of monochromatic

light, you can apply it to white light (a mixture of all the visible frequencies) to answer questions such as, "What wavelength of incoming white light will experience maximum reflection?" The answer helps one understand the rainbow of colors you might see on the surface of a puddle in broad daylight. You put a clear layer of gasoline on top of a clear puddle of water and thin-film interference results in maximal constructive interference of the reflected light, at certain wavelengths.

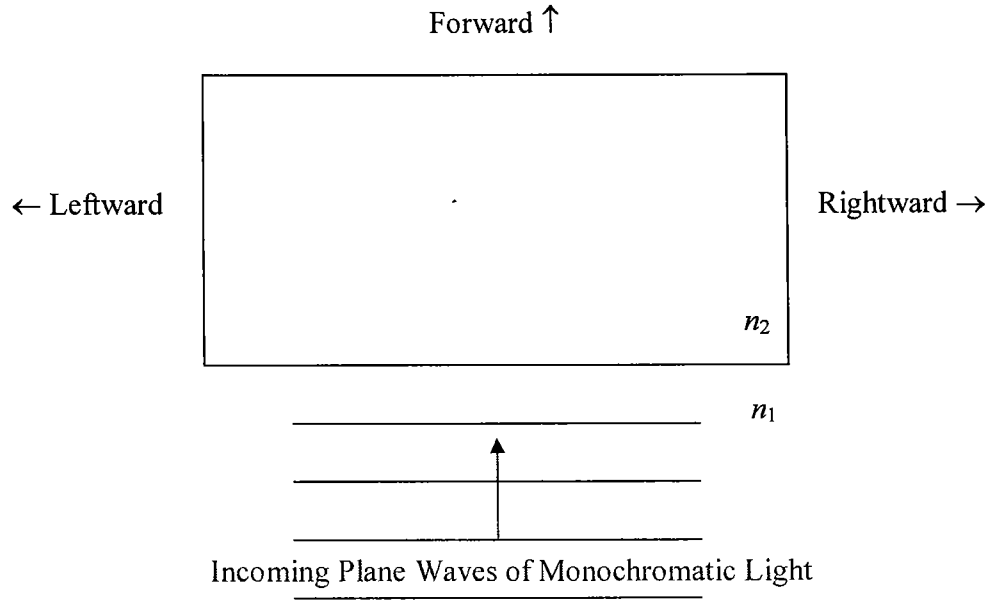
Based on your experience with soap bubbles and puddle surfaces, you know that the light does not have to be normally incident upon the interface between transparent media in order for thin-film interference to occur. However, the analysis is easier for the case of normal incidence, so, in this chapter, I am going to limit our analysis to the case of normal incidence.



Incoming Plane Waves of Monochromatic Light

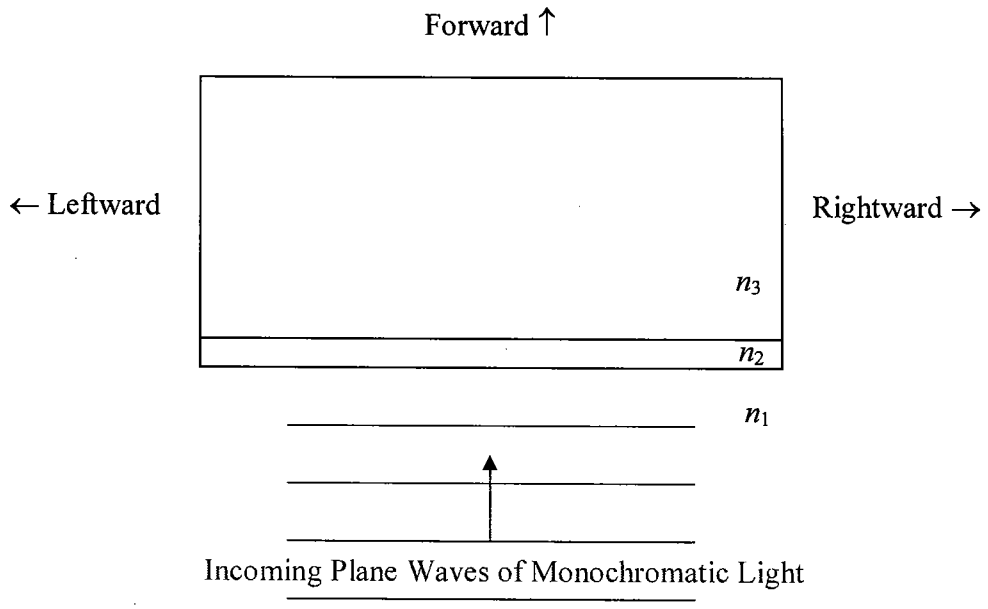
Here's the gross idea: In going through a thin transparent film, light encounters two interfaces, the n_1 abutting n_2 interface, and, the n_2 abutting n_3 interface. At each interface, some of the light gets through and some is reflected. We can say all that we need to say about thin-film interference, just by talking about the reflected light. The thing is, light reflected off the second interface interferes with light reflected off the first interface. The reflected light can be thought of as being from two sources at two different locations, one source being the n_1 abutting n_2 interface, and the other being the n_2 abutting n_3 interface. But, there is a fixed phase difference between the light from the two sources because the light was originally part of one and the same source of incoming light. The light reflected from the second interface travels farther, to arrive back at the same backward position, than the light reflected from the first interface does. If you figure, that, when that extra distance is one wavelength, the interference of reflected light is constructive, and that, when that extra distance is half a wavelength, the interference of reflected light is destructive, then you've got the right idea, but, there are two "complications" that need to be taken into account.

The first complication has to do with phase reversal upon reflection. Consider a single interface (forget about the thin film for a moment) between two transparent medium. Assume light to be incident upon the interface. Call the index of refraction of the medium in which the light is initially traveling, n_1, and call the index of refraction of the medium in which the transmitted light travels, n_2. Experimentally we find that if $n_2 > n_1$, then the reflected light is phase reversed, but, that if $n_2 < n_1$, the reflected light experiences no phase change at all.



Regarding what we mean by phase reversal: Think of a crest of a wave hitting the interface. More specifically, let the electric field be oscillating along the vertical (into and out of the page in the diagram) so that at the instant under consideration, we have a forward-moving maximum upward-directed electric field vector at the location of the interface. An infinitesimal time dt later, we will find a forward-moving maximum-upward electric field vector at a point $v_2\,dt$ forward of the interface. If $n_2 > n_1$ (phase reversal condition met), then, at the same instant in time (dt after the forward-moving maximum upward-directed electric field vector hits the interface) we have a maximum *downward*-directed electric field vector, traveling backward, at a point $v_1\,dt$ behind the interface. This is what we mean by phase reversal. An incoming electric field vector pointing in one direction, bounces off the interface as an electric field vector pointing in the *opposite* direction. If there is *no phase reversal*, then, at the specified instant in time, we would have a maximum *upward*-directed electric field vector, traveling backward, at a point $v_1\,dt$ behind the interface. With no phase reversal, an electric field vector pointing in one direction, bounces off the interface as an electric field vector pointing in the same direction.

Now back to the thin-film setup:



Recall that to get back to some specified point in space, light reflecting off the second interface (between medium 2 and medium 3) travels farther than the light reflecting off the first interface (between medium 1 and medium 2). Before, we hypothesized that if the path difference was a half a wavelength, the light from the two "sources" would interfere destructively, but, that if it was a full wavelength, the interference would be constructive. Now, if there is *no* phase reversal from either surface (because $n_2 < n_1$ and $n_3 < n_2$), or, if there *is* phase reversal from *both* surfaces (because $n_2 > n_1$ and $n_3 > n_2$) then our original hypothesis is still viable. But, if we have phase reversal at one of the interfaces but not the other ($n_2 > n_1$ but $n_3 < n_2$, or, $n_3 > n_2$ but $n_2 < n_1$), then the situation is reversed. A path difference of one wavelength would result in a crest interfering with a "crest that upon reflection turned into a trough" meaning that a path difference of one wavelength would result in *destructive* interference. And, a path difference of *half a wavelength* would result in a crest interfering with a "trough that upon reflection turned into a crest" meaning that a path difference of half a wavelength would result in *constructive* interference. Okay, we have addressed the phase reversal issue. We have one more complication to deal with. The thing is, the light that bounces off the *second* interface, not only travels a greater distance, but, it travels at a different *speed* while it is traveling that extra distance because it is in a different medium. Let's see how this complication plays out.

The phenomenon holds true for every part of the wave. I focus the attention on crests, just because I find them easier to keep track of. For now, I also want to focus attention on the no-phase-reversal constructive interference case. Consider an instant when a crest of the forward-traveling incoming wave hits the first interface. The crest of the transmitted wave travels through the interface, proceeds through the second medium at speed $v_2 = c/n_2$, bounces off the interface with the third medium, and travels back through the second medium, completing its

round trip (of distance two times the thickness of the second medium) through the second medium in time:

$$t_2 = \frac{2(thickness)}{v_2}$$

Now, while that is going on, the next crest from the incoming wave is traveling forward at speed $v_1 = c/n_1$. It arrives at the same interface (between medium 1 and medium 2) at time

$$t_1 = \frac{\lambda_1}{v_1}$$

where λ_1 is the wavelength of the light while it is traveling in medium 1. (Remember, the source establishes the frequency of the light and *that* never changes, but, from $v = \lambda f$, the *wavelength* depends on the speed of the wave in the medium in which the light is traveling.) For constructive interference (under no phase-reversal conditions), we must have

$$t_1 = t_2$$

which, from the expressions for t_1 and t_2 above, can be written as:

$$\frac{\lambda_1}{v_1} = \frac{2(thickness)}{v_2}$$

Substituting $v_1 = \lambda_1 f$ and $v_2 = \lambda_2 f$ yields:

$$\frac{\lambda_1}{\lambda_1 f} = \frac{2(thickness)}{\lambda_2 f}$$

$$\lambda_2 = 2(thickness)$$

I'm going to leave the result in this form because twice the thickness of the thin film is the path difference. So the equation is saying that, under no-phase-reversal conditions, there will be constructive interference of the light reflected from the two interfaces, when the wavelength that the light has in the material of which the thin film consists, is equal to the path difference. Of course, if the path difference is $2\lambda_2$, $3\lambda_2$, $4\lambda_2$, etc. we will also get constructive interference. We can write this as:

$$m\lambda_2 = 2(thickness) \qquad (m = 1,2,3,\ldots) \qquad (24\text{-}1)$$

where:
m is an integer,
λ_2 is the in-the-thin-film wavelength of the light, and,
thickness is the thickness of the thin film.

This condition is also appropriate for the case of maximal constructive interference when phase reversal occurs at both interfaces. But, this condition yields completely *destructive* interference of reflected light when there is phase reversal at one interface but not the other.

The in-the-film wavelength λ_2 of the light can be expressed most simply in terms of the wavelength λ_1 of the light in the medium in which it is originally traveling, by setting the two expressions for the frequency equal to each other. From $v_1 = \lambda_1 f$ we have $f = v_1/\lambda_1$. but from $n_1 = c/v_1$ we have $v_1 = c/n_1$. Replacing v_1 in $f = v_1/\lambda_1$ with c/n_1 yields $f = \dfrac{c}{n_1\lambda_1}$. In a similar manner, we find that f can be expressed as $f = \dfrac{c}{n_2\lambda_2}$. Setting the two expressions for f equal to each other yields:

$$\frac{c}{n_1\lambda_1} = \frac{c}{n_2\lambda_2}$$

which can be written as:

$$\lambda_2 = \frac{n_1}{n_2}\lambda_1 \qquad (24\text{-}2)$$

To get a maximum when we have phase reversal at one, and only one, interface, we need the path difference (twice the thickness) to be *half* a wavelength-in-the-film,

$$\tfrac{1}{2}\lambda_2 = 2(thickness)$$

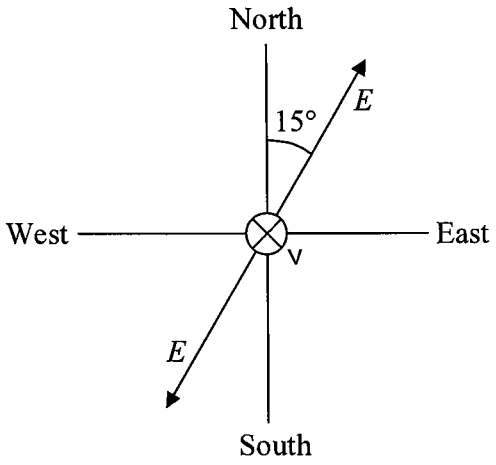
or that, plus, an integer number of full wavelengths-in-the-film:

$$(m+\tfrac{1}{2})\lambda_2 = 2(thickness) \qquad (m = 1,2,3,\ldots) \qquad (24\text{-}3)$$

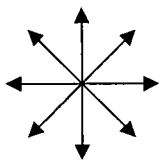
This is also the condition for completely destructive interference for the case of no phase reversal, or, phase reversal at both interfaces. This is exactly what we want for a camera lens that is to be used in air ($n_1 = 1.00$). Consider a clear *plastic* medium of index of refraction $n_2 = 1.3$. Now consider the *wavelength* that light from the middle of the visible spectrum (green light) would have in that medium. Put a coating of the plastic, one quarter as thick as that wavelength is long, on a lens made of glass having an index of refraction $n_3 = 1.5$. (Note that we have phase reversal at both interfaces.) With that coating, the lens reflects none of the light of the specified wavelength that is normally incident on the lens (and a reduced amount of light of nearby wavelengths). That is, it transmits more of the light than it would without the coating. This is the desired effect.

25 Polarization

The polarization direction of light refers to the two directions or one of the two directions in which the electric field is oscillating. For the case of completely polarized light there are always two directions that could be called the polarization direction. If a single direction is specified, then that direction, and the exact opposite direction, are both the directions of polarization. Still, specifying one direction completely specifies the direction of polarization. For instance for light that is traveling straight downward near the surface of the earth, if the polarization direction is said to be a compass heading of 15°, that unambiguously means that the electric field oscillates so that it is at times pointing in the direction with a compass heading of 15°, and at times pointing in the direction with a compass heading of 195° (15° west of south).



Randomly polarized light, a.k.a. *unpolarized light*, has electric field oscillations in each and every direction perpendicular to the direction in which the light is traveling. Such light is often depicted, as viewed from behind, (where forward is the direction in which the light is traveling) as:



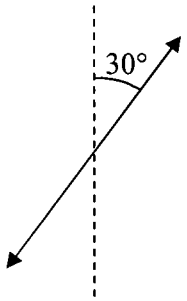
Unpolarized Light, a.k.a.
Randomly Polarized Light

Vertically polarized light traveling horizontally away from you is typically depicted as:

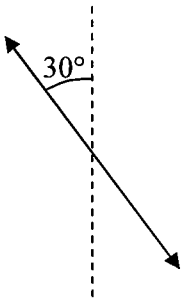
Vertically Polarized Light

where the direction in which the light is traveling is "into the page" and upward is "toward the top of the page." At a particular position through which the light is traveling, starting at an instant when the electric field vector at that position is upward and maximum, the electric field will decrease to zero, then be downward and increasing, reach a maximum downward, then be downward and decreasing, become zero, then be upward and increasing, then reach a maximum upward, and repeat, continually. The diagram depicting the polarization indicates the directions that the electric field does point, at some time during its oscillations. It in no way is meant to imply that the electric field is pointing in two directions at the same time.

Light that is traveling horizontally away from you that is polarized at 30° with respect to the vertical could be either:

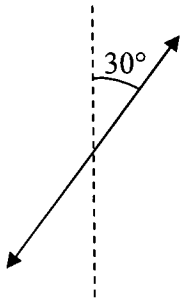


or



If you encounter such an ambiguous specification of polarization in a problem statement then the answer is the same for either case, so, it doesn't matter which of the two possible polarization directions you pick. Pick one arbitrarily and work with it.

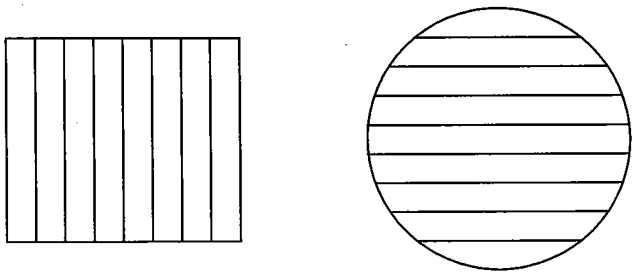
Light that is traveling horizontally away from you and is polarized, from your point of view, at 30° clockwise from the vertical is, however, unambiguously:



Polarizers

A plastic material is manufactured in the form of flat sheets that polarize light that travels through them. A sample of such a flat sheet is called a polarizer. In use, one typically causes light to travel toward a polarizer along a direction that is perpendicular to the polarizer. In other words, one causes the light to be normally incident upon the polarizer.

Schematically, one typically depicts a polarizer by means of a rectangle or a circle filled with parallel line segments.



The orientation of the lines is referred to as the polarization direction of the polarizer. The effect of a polarizer is to transmit light that is polarized in the same direction as that of the polarizer, and to block (absorb or reflect) light that is polarized at right angles to the direction of the polarizer.

The polarization direction of the rectangular polarizer depicted above is vertical. So, it lets vertically-polarized light through and blocks horizontally-polarized light.

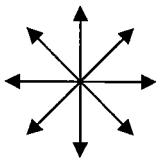
The polarization direction of the circle-shaped sample of polarizing material depicted above is horizontal. So, it lets horizontally-polarized light through and blocks vertically-polarized light.

When unpolarized light (a.k.a. randomly-polarized light) is normally incident on any polarizer, half the light gets through. So, if the intensity of the incoming light is I_0, then the intensity of the light that gets through, call it I_1, is given by:

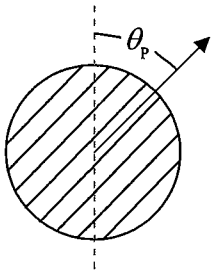
$$I_1 = \tfrac{1}{2} I_0 \tag{25-1}$$

In completely unpolarized light, the electric field vectors are oscillating in every direction that is perpendicular to the direction in which the light is traveling. But all the electric field vectors are, as the name implies, vectors. As such, we can break every single one of them up into a component along the direction of polarization of the polarizer and a component that is perpendicular to the polarization direction of the polarizer. A polarizer will let every component that is along the direction of polarization of the polarizer through, and block every component that is perpendicular to the polarization direction. In completely unpolarized light, no matter what the direction of polarization of the polarizer is, if you break up all the electric field vectors into components parallel to and perpendicular to the polarizer's polarization direction, and add all the parallel components together, and then separately add all the perpendicular components together, the two results will have the same magnitude. This means that we can view completely unpolarized light as being made up of two halves: half polarized *parallel* to the polarizer's polarization direction, and half polarized *perpendicular* to the polarizer's polarization direction. The half that is polarized parallel to the polarizer's polarization direction gets through the polarizer, and the other half doesn't.

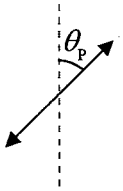
Unpolarized Light Traveling Directly Away From You



When completely unpolarized light of intensity I_0 ...



... is normally incident on a polarizer whose polarization direction makes an angle θ_P with the vertical...

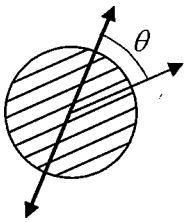


... the light that gets through is polarized in the polarizer's direction of polarization, and, has an intensity $I_1 = \frac{1}{2} I_0$.

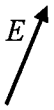
Note how the effect of a polarizer on the intensity of normally-incident unpolarized light does not depend on the orientation of the polarizer. You get the same intensity $I_1 = \frac{1}{2} I_0$ of light getting through the polarizer, no matter what the direction of polarization of the polarizer is.

Now suppose that we have some light that, for whatever reason, is *already polarized.* When polarized light is normally incident on a polarizer, the intensity of the light that gets through *does* depend on the direction of polarization of the polarizer (relative to that of the incoming light). Suppose for instance, that the incoming light,

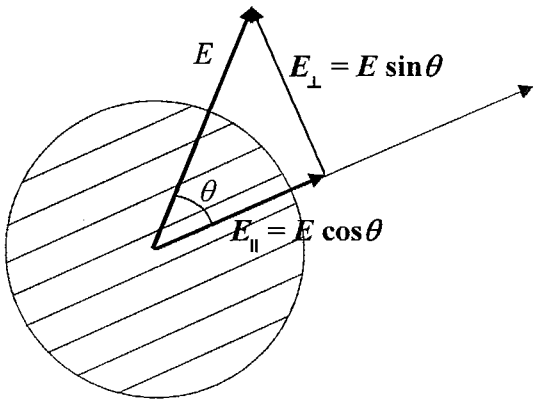
is polarized at an angle θ with respect to the polarization direction of a polarizer upon which the light is normally incident:

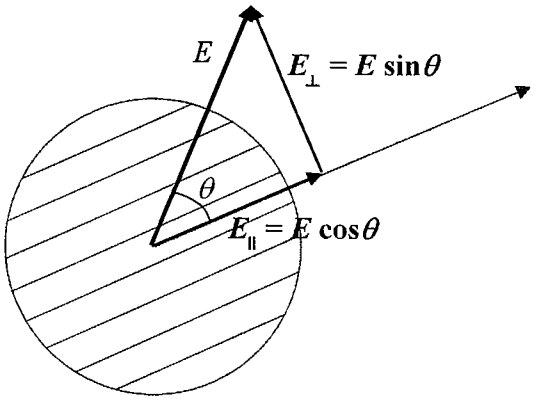


Before it hits the polarizer, the light's electric-field-oscillations-amplitude vector,



can be broken up into a component parallel to the polarizer's polarization direction and a component perpendicular to the polarizer's polarization direction.





The parallel component $E_{||} = E\cos\theta$ gets through the polarizer, the perpendicular component does not.

Now the intensity of polarized light is proportional to the square of the amplitude of the oscillations of the electric field. So, we can express the intensity of the *incoming light* as

$$I_0 = (constant)\, E^2$$

and the intensity of the *light that gets through* as:

$$I_1 = (constant)\, E_{||}^2$$

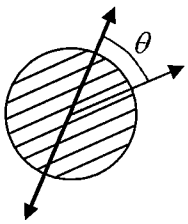
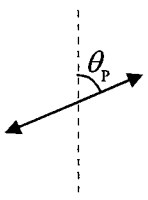
$$I_1 = (constant)\, (E\cos\theta)^2$$

$$I_1 = (constant)\, E^2 (\cos\theta)^2$$

$$I_1 = I_0 (\cos\theta)^2 \qquad (25\text{-}2)$$

Summarizing:

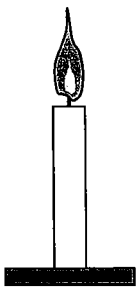
Polarized Light Traveling Directly Away From You

 When polarized light of intensity I_0...	 ... is normally incident on a polarizer whose polarization direction makes an angle θ with the polarization direction of the light...	 ... the light that gets through is polarized in the polarizer's direction of polarization, and, has an intensity $I_1 = I_0(\cos\theta)^2$.

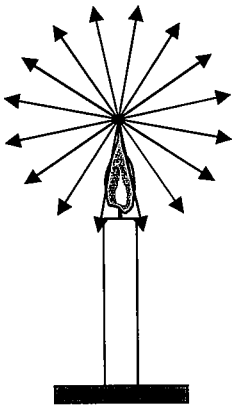
26 Geometric Optics, Reflection

We now turn to a branch of optics referred to as geometric optics and also referred to as ray optics. It applies in cases where the dimensions of the objects (and apertures) with which the light interacts are so large as to render diffraction effects negligible. In geometric optics we treat light as being made up of an infinite set of narrow beams of light, called *light rays*, or simply rays, traveling through vacuum or transparent media along straight line paths. Where a ray of light encounters the surface of a mirror, or the interface between the transparent medium in which it (the light) is traveling and another transparent medium, the ray makes an abrupt change in direction, after which, it travels along a new straight line path.

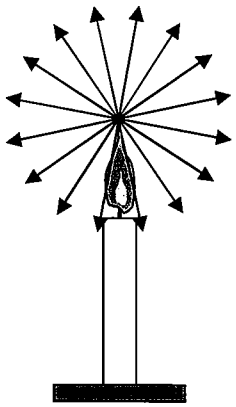
In the geometric optics model of light, we see light emitted by sources of light because the light enters our eyes. Consider for instance, a candle.



Every point of the flame of the candle emits rays of light in every direction.



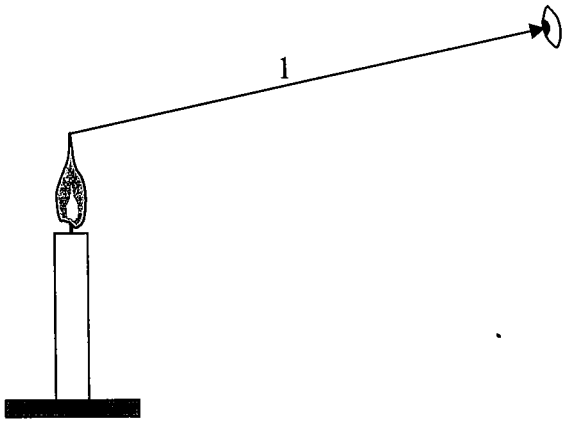
While the preceding diagram conveys the idea in the statement preceding the diagram, the diagram is not the complete picture. To get a more complete picture of what's going on, what I want you to do is to look at the diagram provided, form a picture of it in your mind, and, to the picture in your mind, add the following embellishments:



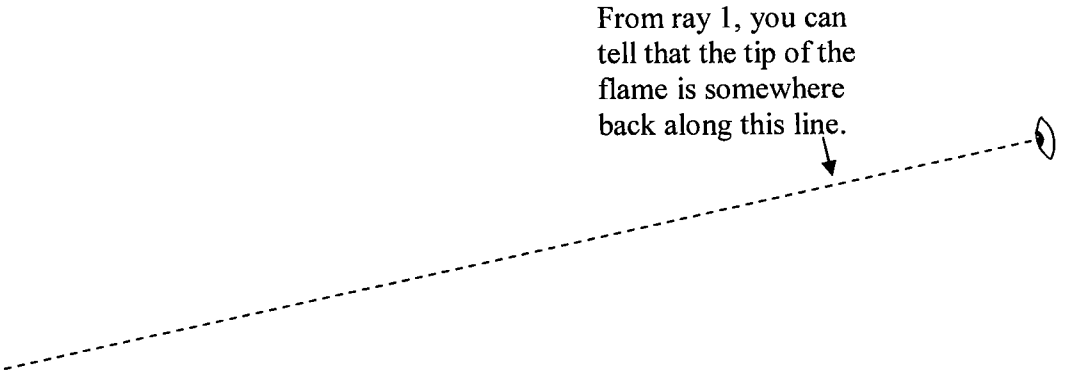
1. First off, I need you to imagine it to be a real candle extending in *three dimensions*. Our set of rays depicted as arrows whose tips are all on a circle becomes a set of rays depicted as arrows whose tips all end on a *sphere*. Thus, in addition to rays going (at various angles) upward, downward and to the sides, you've got rays proceeding (at various angles) away from you and toward you.
2. Now I need you to add more rays to the picture in your mind. I included 16 rays in the diagram. In three dimensions, you should have about 120 rays in the picture in your mind. I need you to bump that up to infinity.
3. In the original diagram, I showed rays coming only from the tip of the flame. At this point, we have an infinite number of rays coming from the tip of the flame. I need you to picture that to be the case for each point of the flame, not just the tip of the flame. In the interest of simplicity, in the picture in your mind, let the flame of the candle be an opaque solid rather than gaseous, so that we can treat all our rays as coming from points on the surface of the flame. Neglect any rays that are in any way directed into the flame itself (don't include them in the picture in your mind). Upon completion of this step, you should have, in the picture in your mind, an infinite number of rays coming from each of the infinite number of points making up the surface of the flame.
4. For this next part, we need to establish the setting. I'm concerned that you might be reading this in a room in which lit candles are forbidden. If so, please relocate the candle in the picture in your mind to the dining room table in your home, or, replace the candle with a fake electric-powered candle such as you might see in a home around Christmastime. Now I need you to extend each of the rays in the picture in your mind all the way out to the point where they bump into something. Please *end* each ray at the point where it bumps into something. (A ray that bumps into a non-shiny surface, bounces off in all directions [diffuse reflection]. Thus, each ray that bumps into a non-shiny surface creates an infinite set of rays coming from the point of impact. A ray bumping into perfectly shiny surfaces continues as a single ray in one particular, new, direction [specular reflection]. To avoid clutter, let's omit all the reflected rays from the picture in your mind.)

If you have carried out steps 1-4 above, then you have the picture, in your mind, of the geometric optics model of the light given off by a light-emitting object. When you are in a room with a candle such as the one we have been discussing, you can tell where it is (in what direction and how far away—you might not be able to give very accurate values, but you can tell where it is) by looking at it. When you look at it, an infinite number of rays, from each part of the surface of

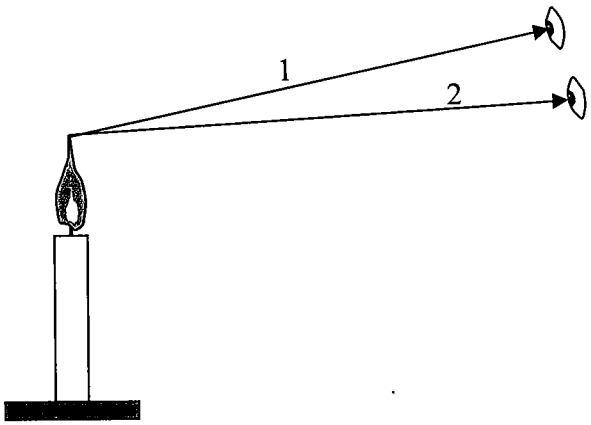
the flame, are entering your eyes. What is amazing is how few rays you need to determine where, for instance, the tip of the flame is. Of the infinite number of rays available to you, you only need two! Consider what you can find out from a single ray entering your eye:

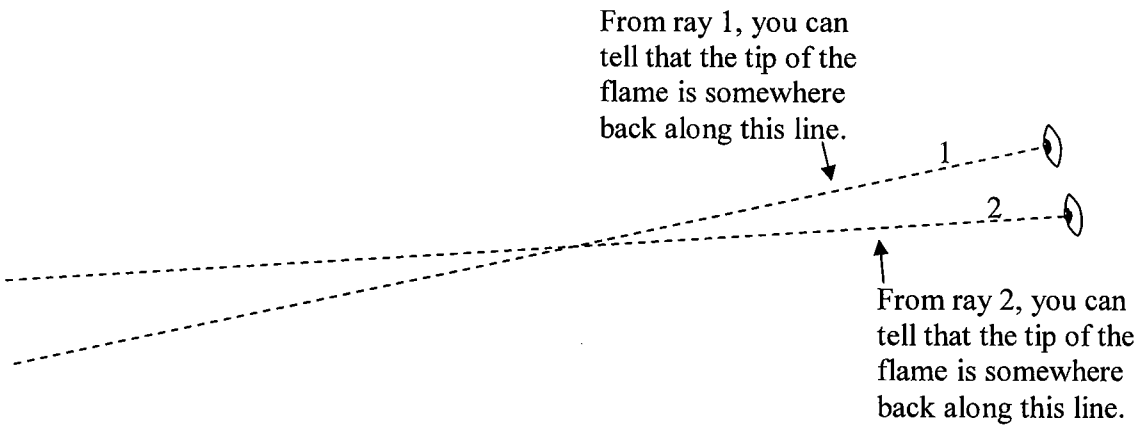


From just one of the infinite number of rays, you can deduce the direction that the tip of the flame is in, relative to you. In other words, you can say that the tip of the flame lies somewhere on the line segment that both contains the ray that enters your eye, and, that ends at the location of your eye.



From one other ray, ray 2 in the diagram at right, you can deduce that the tip of the flame must lie somewhere back along one other line.



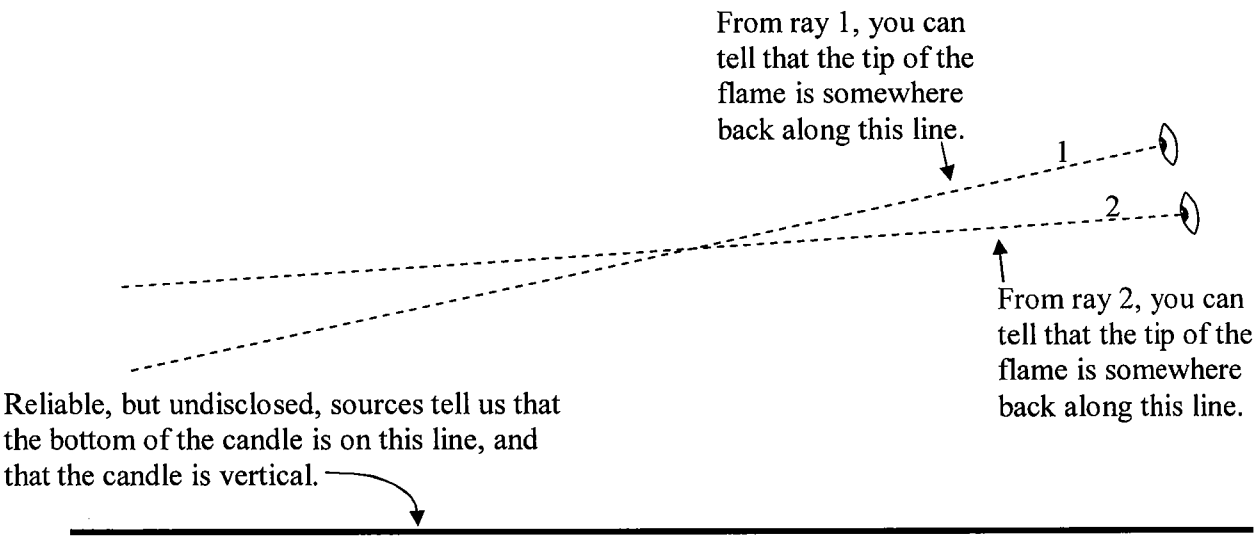


There is only one point in space that is both "somewhere back along line 1" and "somewhere back along line 2." That one point is, of course, the point where the two lines cross. The eye-brain system is an amazing system. When you look at something, your eye-brain system automatically carries out the "trace back and find the intersection" process to determine how far away that something is. Again, you might not be able to tell me how many centimeters away the candle, for instance, is, but you must know how far away it is because you *would* know about how hard to throw something to hit the candle[1].

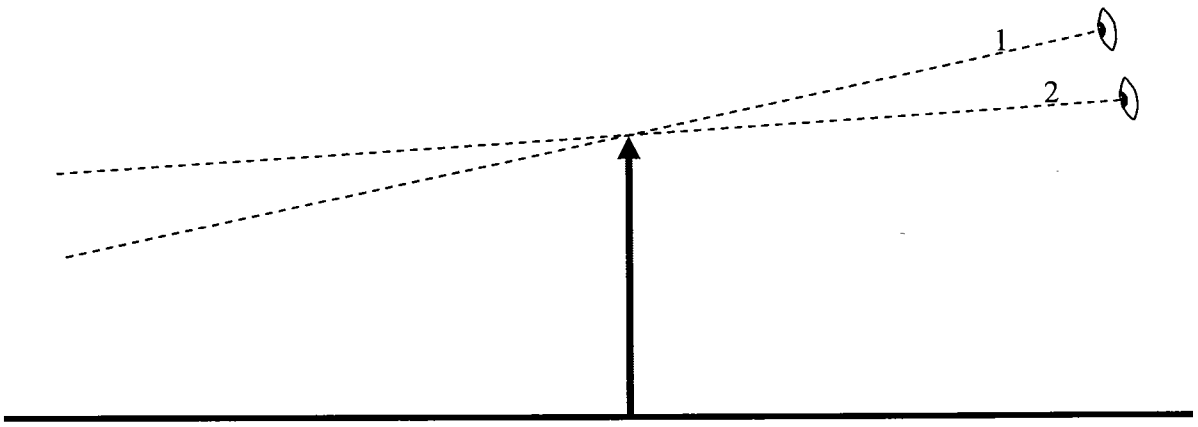
This business of tracing rays back to see where they come from is known as ray tracing and is what geometric optics is all about.

At this point I want to return our attention to the candle to provide you with a little bit more insight into the practice of ray tracing. Suppose that when you were determining the location of the tip of the flame of the candle, you already had some additional information about the candle. For instance, assume: You know that the rays are coming from the upper extremity of the candle; you know that the bottom of the candle is on the plane of the surface of your dining room table; and you know that the candle is vertical. We'll also assume that the candle is so skinny that we are not interested in its horizontal extent in space, so, we can think of it as a skinny line segment with a top (the tip of the candle) and a bottom, the point on the candle that is at table level. The intersection of the plane of the table surface with the plane of the two rays is a line, and, based on the information we have, the bottom of the candle is on that line.

[1] Note: Throwing things at lit candles is a dangerous practice in which I urge you *not* to engage.



Taken together with the information gleaned from the rays, we can draw in the entire (skinny) candle, on our diagram, and from the diagram, determine such things as the candle's height, position, and orientation (whether it is upside down [inverted] or right side up [erect]). In adding the candle to the diagram, I am going to draw it as an arrow. Besides the fact that it is conventional to draw objects in ray tracing diagrams as arrows, we use an arrow to represent the candle to avoid conveying the impression that, from the limited facts we have at our disposal, we have been able to learn more about the candle (diameter, flame height, etc) than is possible. (We can only determine the height, position, and orientation.)

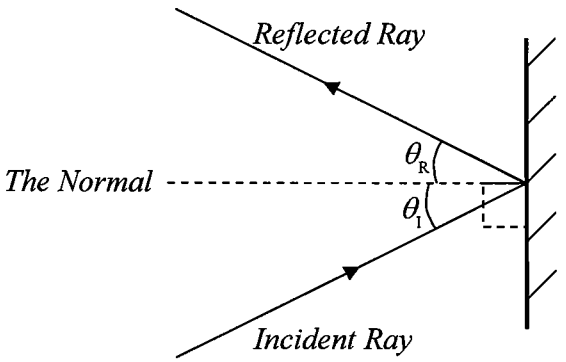


The trace-back method for locating the tip of the candle flame works for any two rays, from among the infinite number of rays emitted by the tip of the candle flame. All the rays come from the same point and they all travel along different straight line paths. As such, the rays are said to diverge from the tip of the candle flame. The trace-back method allows us to determine the point from which the rays are diverging.

By means of lenses and mirrors, we can redirect rays of light, infinite numbers of them at a time, in such a manner as to fool the eye-brain system that is using the trace-back method into perceiving the point from which the rays are diverging as being someplace other than where the object is. To do so, one simply has to redirect the rays so that they *are* diverging from someplace other than their point of origin. The point, other than their point of origin, from which the rays diverge (because of the redirection of rays by mirrors and/or lenses), is called the *image* of the point on the object from which the light actually originates.

The Law of Reflection

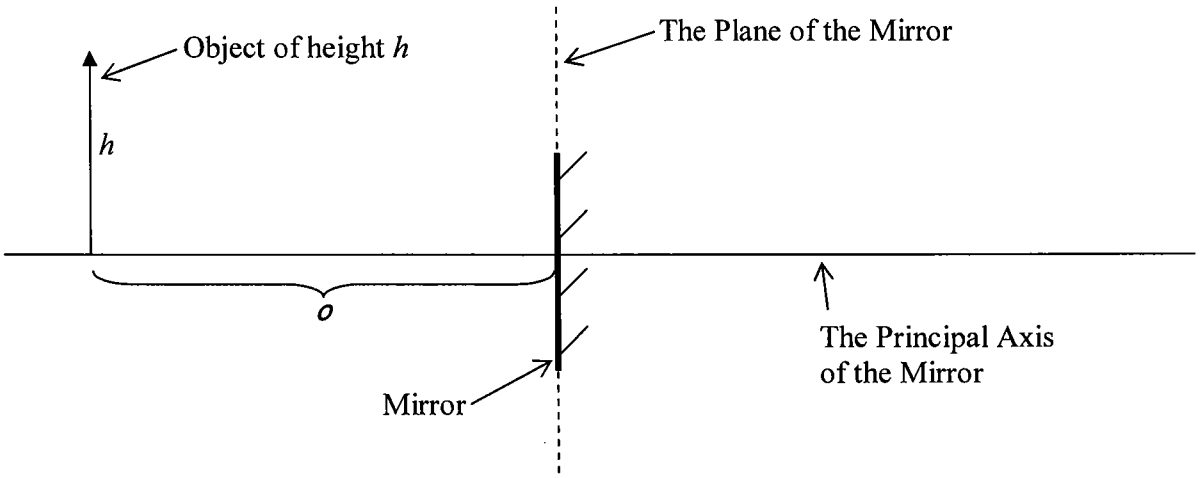
I have mentioned specular reflection. In specular reflection, a ray of light traveling along one straight line path, hits a smooth shiny surface and continues along a new straight line path. The adoption of the new path, at the smooth shiny surface, by the incoming ray is called reflection. Where the ray travels along the new path, we call the ray, the reflected ray. The smooth shiny surface is typically called a mirror. The law of reflection, derived originally directly from experimental evidence and, by Huygens, from the principle now known as Huygens' principle, states that *the angle that the reflected ray makes with an imaginary line that is perpendicular to the mirror, and, passes through the point where the incoming ray hits the mirror, is equal to the angle that the incoming ray makes with the same imaginary line*. The point where the incoming ray hits the mirror is called the *point of incidence*. The imaginary line that is perpendicular to the surface of the mirror and passes through the point of incidence is called *the normal*. The angle that the incoming ray makes with the normal is called the *angle of incidence* θ_I. The angle that the reflected ray makes with the normal is called the *angle of reflection* θ_R. In terms of this jargon, the law of reflection can be stated as: The angle of reflection θ_R is equal to the angle of incidence θ_I.



The Law of Reflection states that $\theta_R = \theta_I$.

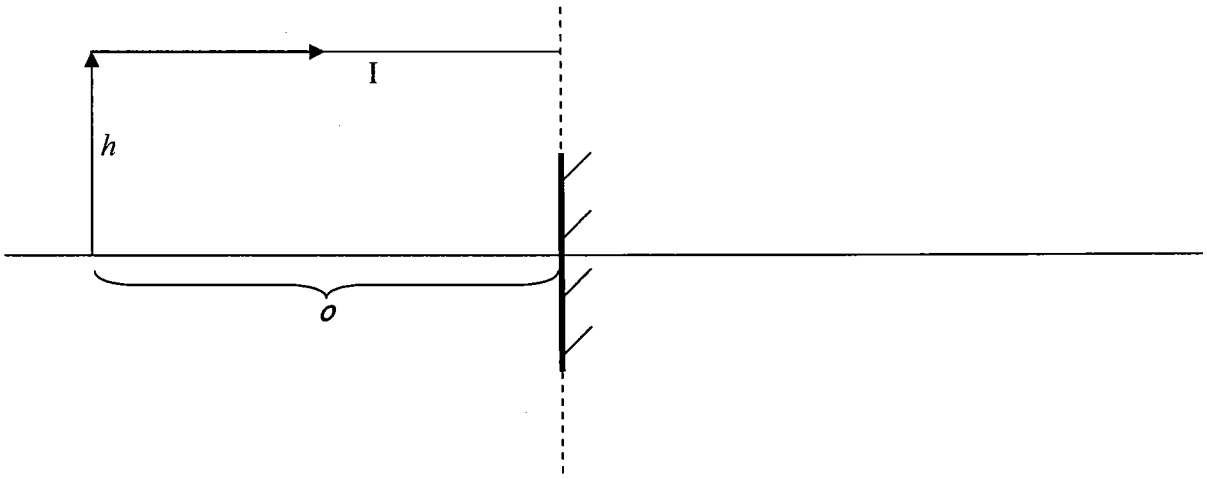
Geometric Optics Applied to a Plane Mirror

Let's apply our ray-tracing methods to the case of an object in front of a plane mirror in order to determine the position of the image of that object. Here's the configuration.

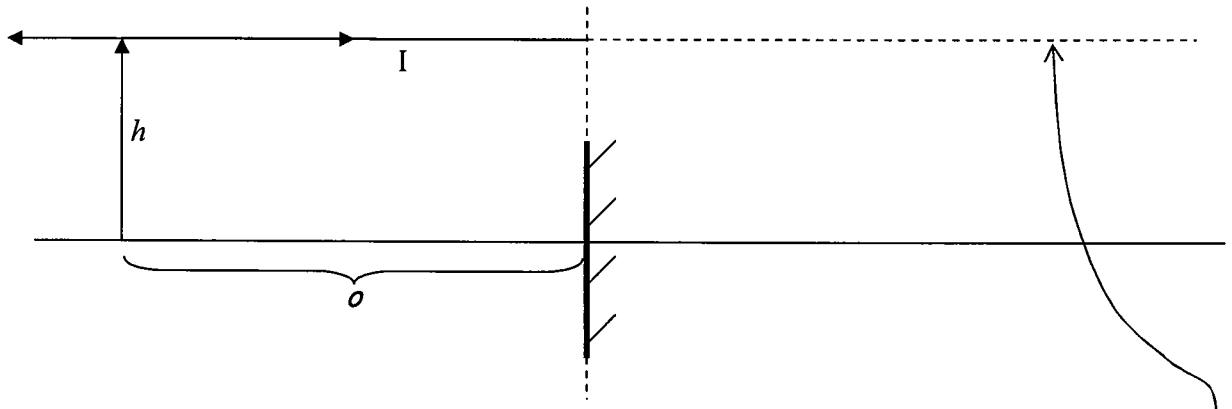


We have an object of height h a distance o from the plane of the mirror. Our object is represented by an arrow. The tail of the arrow is on a reference line that is perpendicular to the plane of the mirror. I am calling the reference line "the principal axis of the mirror." The plane of the mirror is the infinite plane that contains the surface of the mirror.

We use the method of principal rays to determine the position of the image of the object. In the method of principal rays, we consider only a few incident rays for which the reflected rays are particularly easy to determine. Experimentally, we find that the position of the image is independent of the size of the mirror, so we consider the mirror to be as large as it needs to be for the principal rays to hit it. In particular, if a principal ray appears to miss the mirror in our diagram, we show the ray as reflecting off the plane of the mirror nevertheless. Our Principal Ray I for the case at hand is one that approaches the plane of the mirror along a line that is parallel to the principal axis of the mirror.

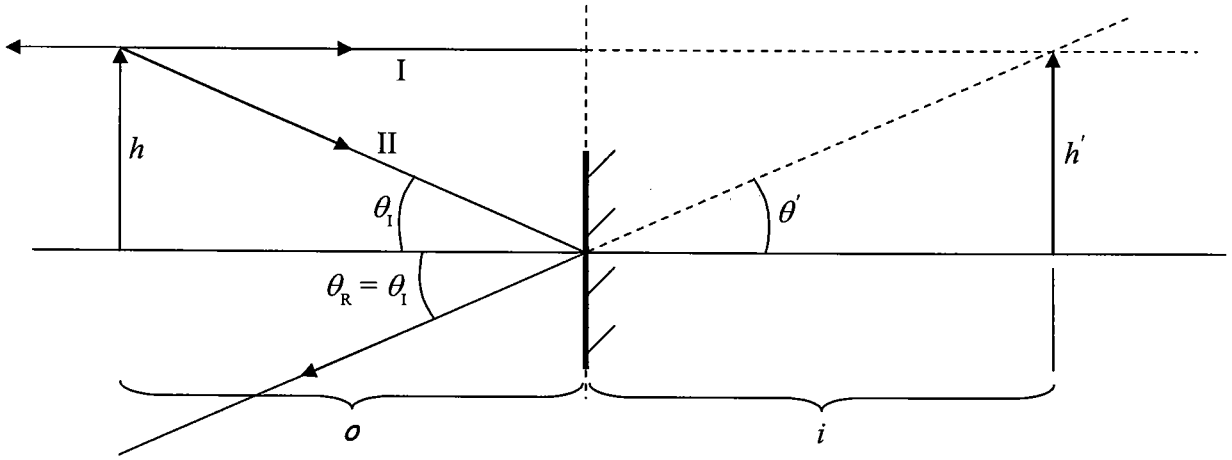


According to the law of reflection, Principal Ray I is reflected straight back on itself as depicted in the following diagram:



Using the trace-back method we know that the tip of the object lies somewhere along this line.

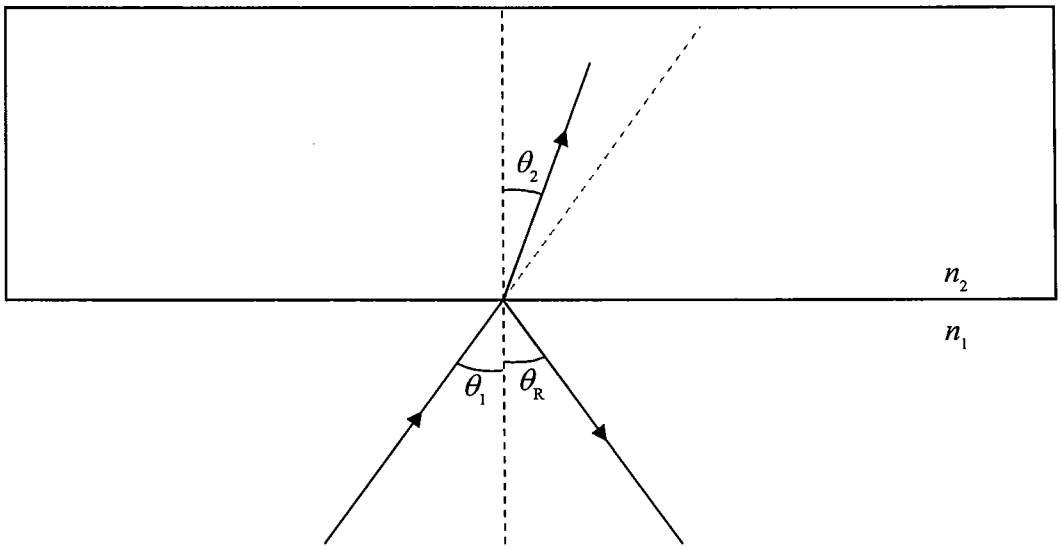
Principal Ray II hits the mirror right where the principal axis of the mirror intersects the mirror. In accord with the Law of Reflection, with, for the ray in question, the principal axis of the mirror being the normal, the reflected ray makes the same angle with the principal axis of the mirror as the incident ray does.



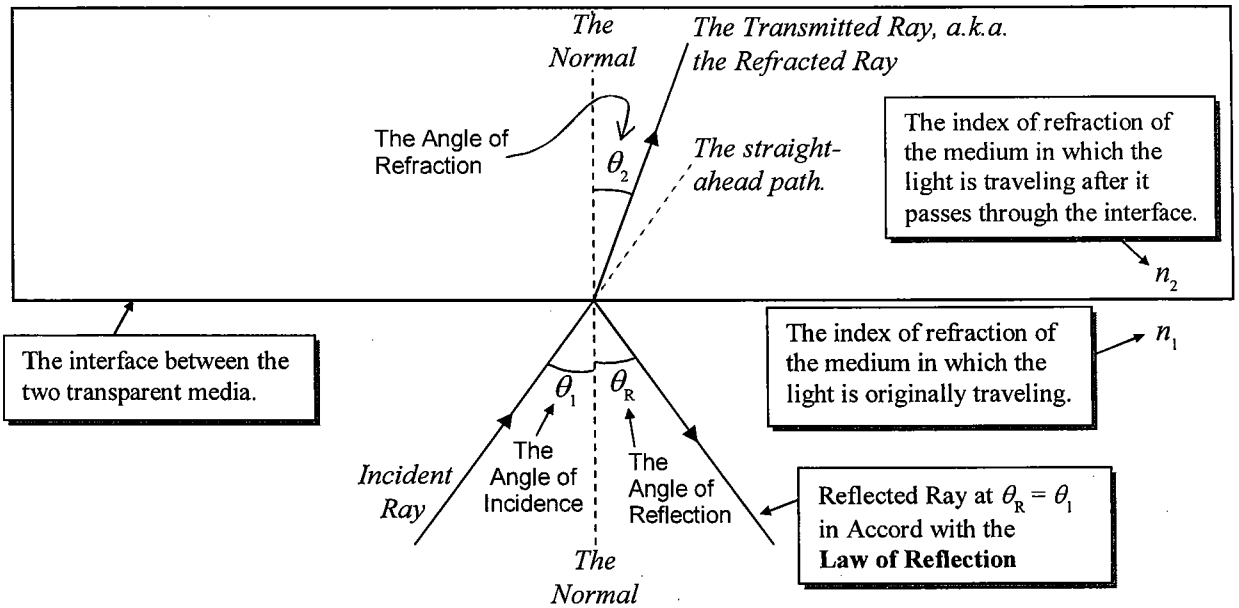
Tracing back the second reflected ray, to the point where it intersects the first reflected ray trace-back line, yields the position of the image of the tip of the arrow. I have drawn the shaft of the image of the arrow in so that it is perpendicular to the principal axis of the mirror. The question is, what is the image height h' and what is the distance of the image from the plane of the mirror? Well, the image height h' is the distance between the same two parallel lines that the object height h is the distance between. So, $h' = h$. Since vertical angles are equal, we have θ' in the diagram above being equal to θ_R which we know to be equal to θ_I from the law of reflection. Thus the right triangle of side h' and angle θ' is congruent to the triangle of height h and angle θ_I. Hence, since corresponding sides of congruent triangles are equal, we have $i = o$. That is to say that the image distance, from the plane of the mirror, is equal to the object distance.

27 Refraction, Dispersion, Internal Reflection

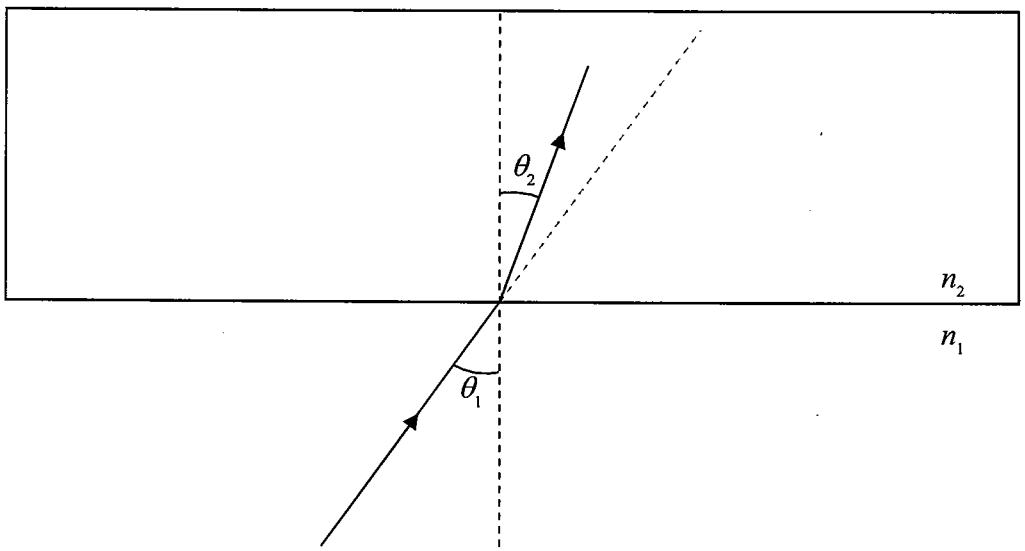
When we talked about thin film interference, we said that when light encounters a smooth interface between two transparent media, some of the light gets through, and some bounces off. There we limited the discussion to the case of normal incidence. (Recall that *normal* means *perpendicular to* and normal incidence is the case where the direction in which the light is traveling is perpendicular to the interface.) Now we consider the case in which light shining on the smooth interface between two transparent media, is *not* normally incident upon the interface. Here's a "clean" depiction of what I'm talking about:



and here's one that's all cluttered up with labels providing terminology that you need to know:



As in the case of normal incidence, some of the light is reflected and some of it is transmitted through the interface. Here we focus our attention on the light that gets through.



Experimentally we find that the light that gets through travels along a different straight line path than the one along which the incoming ray travels. As such, the transmitted ray makes an angle θ_2 with the normal that is different from the angle θ_1 that the incident ray makes with the normal.

The adoption of a new path by the transmitted ray, at the interface between two transparent media is referred to as *refraction*. The transmitted ray is typically referred to as the *refracted ray*, and the angle θ_2 that the refracted ray makes with the normal is called the *angle of refraction*. Experimentally, we find that the angle of refraction θ_2 is related to the angle of incidence θ_1 by Snell's Law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad (27\text{-}1)$$

where:

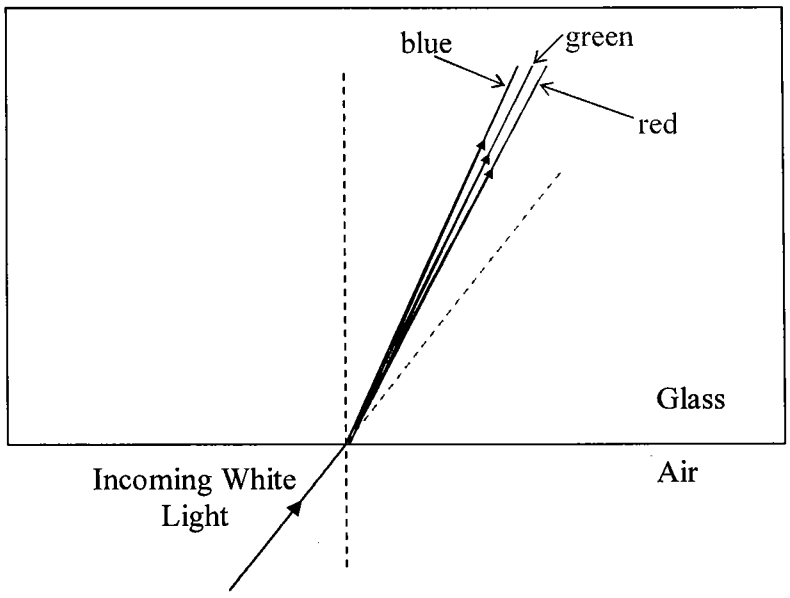
- n_1 is the index of refraction of the *first* medium, the medium in which the light is traveling before it gets to the interface,
- θ_1 is the angle that the incident ray (the ray in the *first* medium) makes with the normal,
- n_2 is the index of refraction of the *second* medium, the medium in which the light is traveling after it goes through the interface, and,
- θ_2 is the angle that the refracted ray (the ray in the *second* medium) makes with the normal.

Dispersion

On each side of the equation form of Snell's law we have an *index of refraction.* The index of refraction has the same meaning as it did when we talked about it in the context of thin film interference. It applies to a given medium. It is the ratio of the speed of light in that medium to the speed of light in vacuum. At that time, I mentioned that different materials have different indices of refraction, and in fact, provided you with the following table:

Medium	***Index of Refraction***
Vacuum	1
Air	1.00
Water	1.33
Glass (Depends on the kind of glass. Here is one typical value.)	1.5

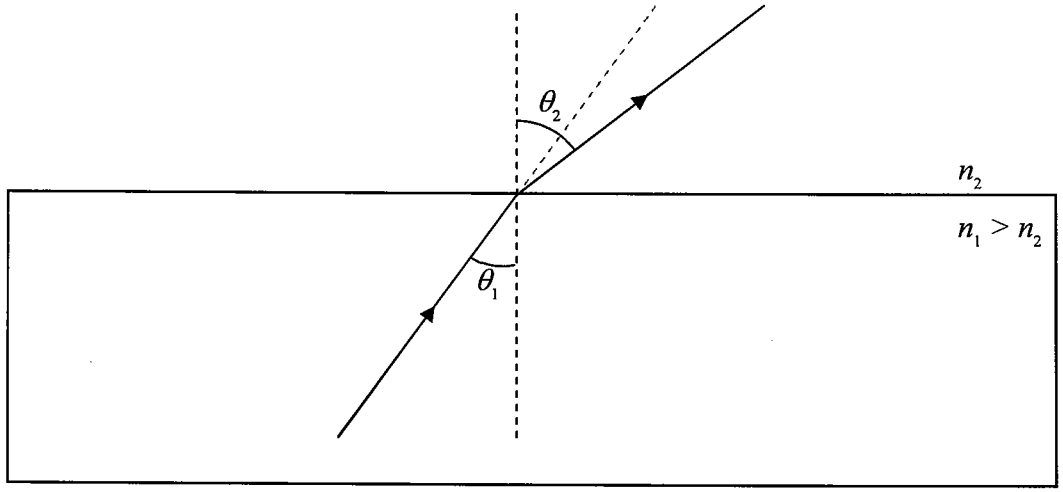
What I didn't mention back then is that there is a slight dependence of the index of refraction on the wavelength of the visible light, such that, the shorter the wavelength of the light, the greater the index of refraction. For instance, a particular kind of glass might have an index of refraction of 1.49 for light of wavelength 695 nm (red light), but an index of refraction that is greater than that for shorter wavelengths, including an index of refraction of 1.51 for light of wavelength 405 nm (blue light). The effect in the case of a ray of white light traveling in air and encountering an interface between air and glass is to cause the different wavelengths of the light making up the white light to refract at different angles.



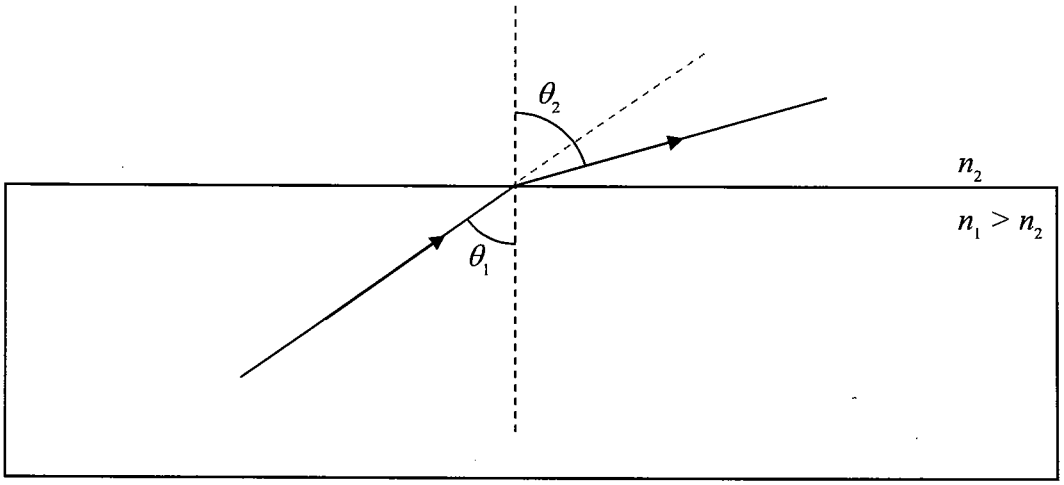
This phenomena of white light being separated into its constituent wavelengths because of the dependence of the index of refraction on wavelength, is called *dispersion.*

Total Internal Reflection

In the case where the index of refraction of the first medium is greater than the index of refraction of the second medium, the angle of refraction is greater than the angle of incidence.

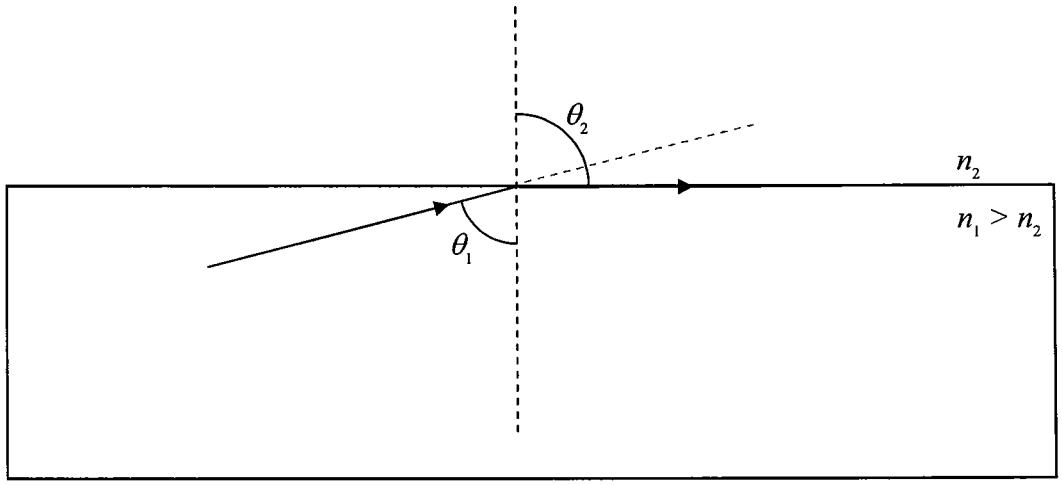


For such a case, look what happens when we increase the angle of incidence, θ_1:



The angle of refraction gets bigger...

until eventually it (the angle of refraction) gets to be 90°.



It can't get any bigger than that, because, beyond that, the light is not going through the interface. An angle of refraction greater than 90° has no meaning. But, note that we still have room to increase the angle of incidence. What happens if we continue to increase the angle of incidence? Well, indeed, no light gets through the interface. But, remember at the beginning of this chapter where we talked about how, when light is incident on the interface between two transparent media, some of the light gets through and some of it is reflected? Well, I haven't been including the reflected ray on our diagrams because we have been focusing our attention on the transmitted ray, but, it is always there. The thing is, at angles of incidence bigger then the angle that makes the angle of refraction 90°, the reflected ray is all there is. The phenomenon, in which there is no transmitted light at all, just reflected light, is known as *total internal reflection.* The angle of incidence that makes the angle of refraction 90° is known as *the critical angle.* At any angle of incidence greater than the critical angle, the light experiences total internal reflection. Note that the phenomenon of total internal reflection only occurs when the light is initially in the medium with the bigger index of refraction.

Let's investigate this phenomenon mathematically. Starting with Snell's Law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

solved for the sine of the angle of refraction:

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

we note that, since it was stipulated that $n_1 > n_2$, the ratio n_1/n_2 is greater than 1. The $\sin\theta_1$ is always less than 1, but, if θ_1 is big enough, $\sin\theta_1$ can be so close to 1 that the right-hand side of

the equation $\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1$ is greater than 1. In that case, there is no θ_2 that will solve the equation because there is no angle whose sine is greater than 1. This is consistent with the experimental fact that, at angles of incidence greater than the critical angle, no light gets through the interface. Let's solve for the critical angle. At the critical angle, the angle of refraction θ_2 is 90°. Let's plug that into the equation we have been working with and solve for θ_1:

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

evaluated at $\theta_2 = 90°$ yields:

$$\sin 90° = \frac{n_1}{n_2}\sin\theta_1$$

$$1 = \frac{n_1}{n_2}\sin\theta_1$$

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

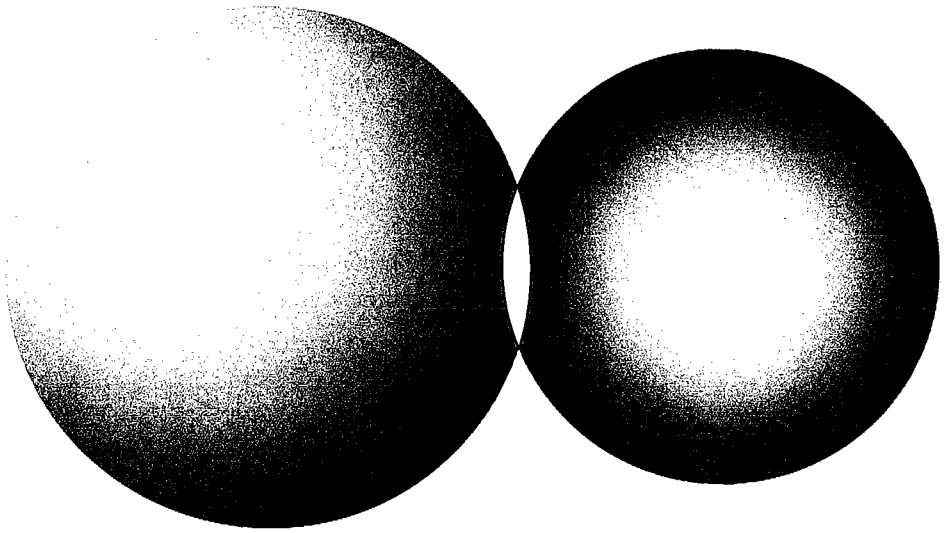
$$\theta_1 = \sin^{-1}\frac{n_2}{n_1}$$

This is such a special angle of incidence that we not only give it a name (as mentioned, it is called the critical angle), but, we give it its own symbol. The critical angle, that angle of incidence beyond which there is no transmitted light, is designated θ_c, and, as we just found, can be expressed as:

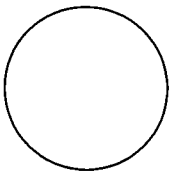
$$\theta_c = \sin^{-1}\frac{n_2}{n_1} \qquad (27\text{-}2)$$

28 Thin Lenses: Ray Tracing

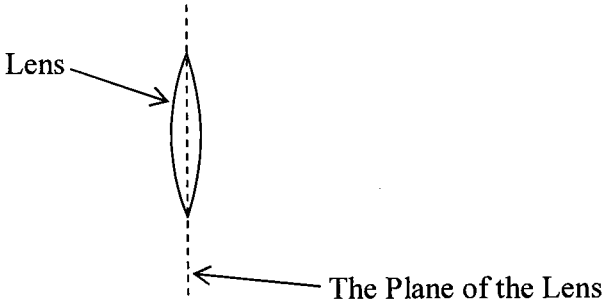
A lens is a piece of transparent material whose surfaces have been shaped so that, when the lens is in another transparent material (call it medium 0), light traveling in medium 0, upon passing through the lens, is redirected to create an image of the light source. Medium 0 is typically air, and lenses are typically made of glass or plastic. In this chapter we focus on a particular class of lenses, a class known as thin spherical lenses. Each surface of a thin spherical lens is a tiny fraction of a spherical surface. For instance, consider the two spheres:



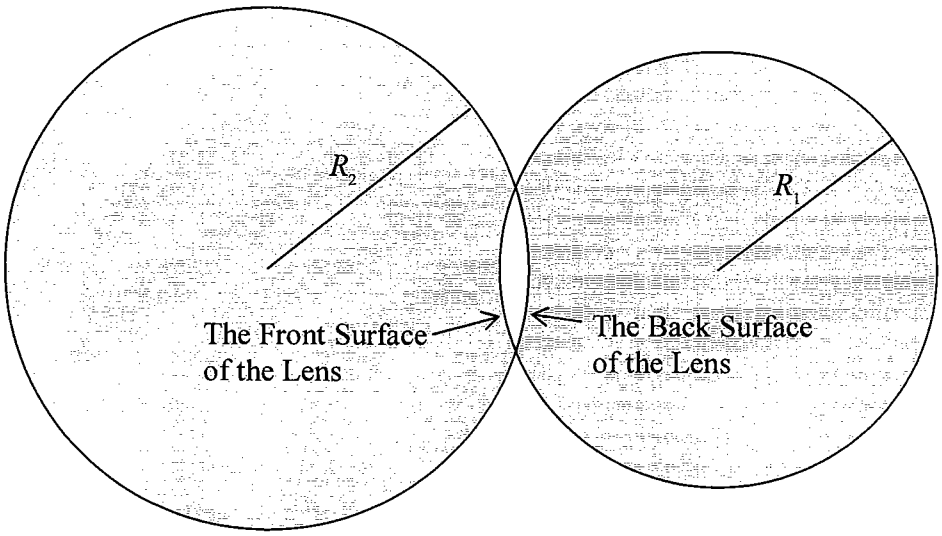
A piece of glass in the shape of the intersection of these two spherical *volumes* would be a thin spherical lens. The intersection of two spherical *surfaces* is a circle. That circle would be the rim of the lens. Viewed face on, the outline of a thin spherical lens is a circle.



The plane in which that circle lies is called the plane of the lens. Viewing the lens edge-on, the plane of the lens looks like a line.



Each surface of a thin spherical lens has a radius of curvature. The radius of curvature of a surface of a thin spherical lens is the radius of the sphere of which that surface is a part. Designating one surface of the lens as the front surface of the lens and one surface as the back surface, in the following diagram:



we can identify R_1 as the radius of curvature of the front surface of the lens and R_2 as the radius of curvature of the back surface of the lens.

The defining characteristic of a lens is a quantity called the focal length of the lens. At this point, I'm going to tell you how you can calculate a value for the focal length of a lens, based on the physical characteristics of the lens, before I even tell you what focal length means. (Don't worry, though, we'll get to the definition soon.) The lens-maker's equation gives the reciprocal of the focal length in terms of the physical characteristics of the lens (and the medium in which the lens finds itself):

The Lens-Maker's Equation:

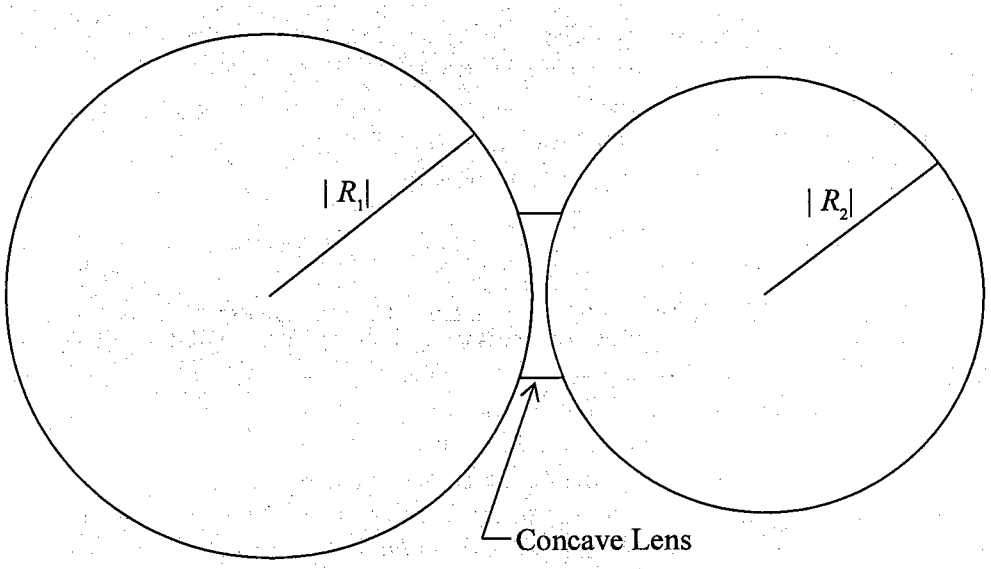
$$\frac{1}{f} = (n - n_0)\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \tag{28-1}$$

where:

- f is the focal length of the lens,
- n is the index of refraction of the material of which the lens is made,
- n_0 is the index of refraction of the medium surrounding the lens (n_0 is typically 1.00 because the medium surrounding the lens is typically air),
- R_1 is the radius of curvature of one of the surfaces of the lens, and,
- R_2 is the radius of curvature of the other surface of the lens.

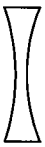
Before we move on from the lens-maker's equation, I need to tell you about an algebraic sign convention for the R values. There are two kinds of spherical lens surfaces. One is the "curved out" kind possessed by any lens that is the intersection of two spheres. (This is the kind of lens that we have been talking about.) Such a lens is referred to as a *convex lens* (a.k.a. a converging lens) and each ("curved out") surface is referred to as a *convex surface*. The radius of curvature R for a convex surface is, by convention, positive.

The other kind of lens surface is part of a sphere that does not enclose the lens itself. Such a surface is said to be "curved in" and is called a concave surface.

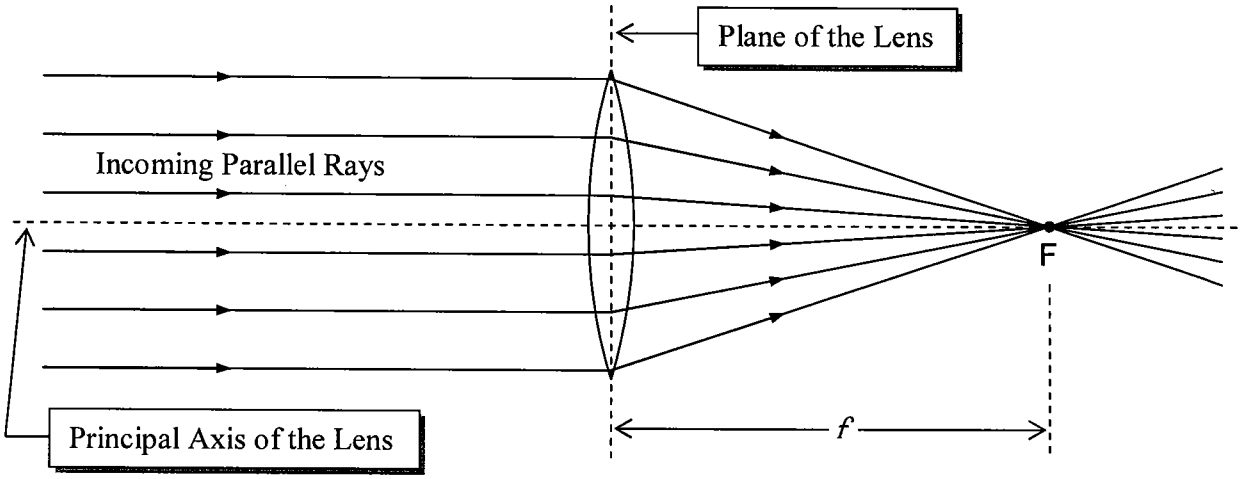


By convention, the absolute value of R for a concave surface is still the radius of the sphere whose surface coincides with that of the lens. But, the quantity R contains additional information in the form of a minus sign used to designate the fact that the surface of the lens is concave. R is still called *the radius of curvature of the surface of the lens* despite the fact that there is no such thing as a sphere whose radius is actually negative.

Summarizing, our convention for the radius of curvature of the surface of a lens is:

Surface of Lens	Algebraic Sign of Radius of Curvature R
Convex 	+
Concave 	−

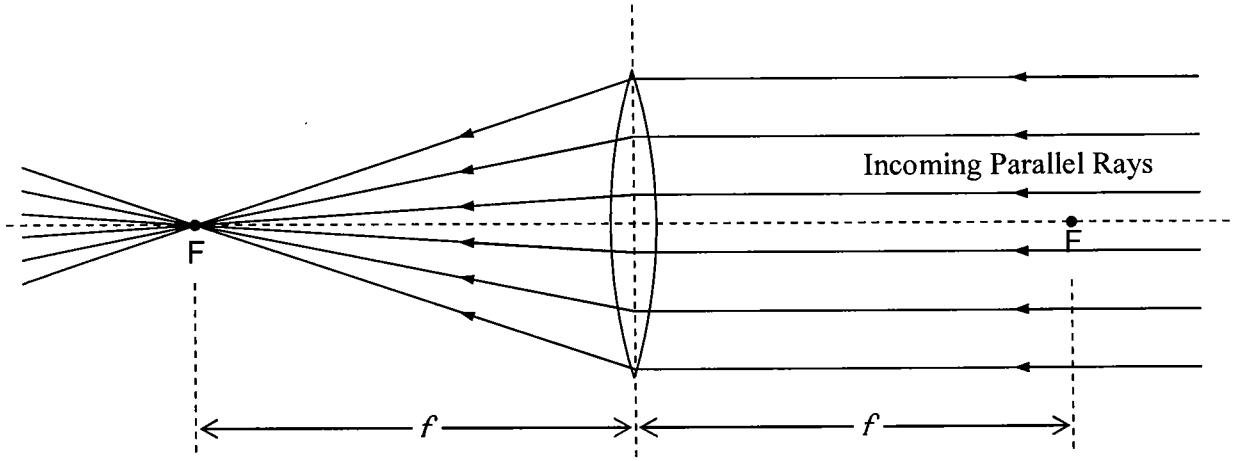
So, what does a lens do? It refracts light at both surfaces. What's special about a lens is the effect that it has on an infinite set of rays, collectively. We can characterize the operational effect of a lens in terms of the effect that it has on incoming rays that are all parallel to the principal axis of the lens. (The principal axis of a lens is an imaginary line that is perpendicular to the plane of the lens and passes through the center of the lens.) A converging lens causes all such rays to pass through a single point on the other side of the lens. That point is the *focal point* F of the lens. Its distance from the lens is called the *focal length* f of the lens.



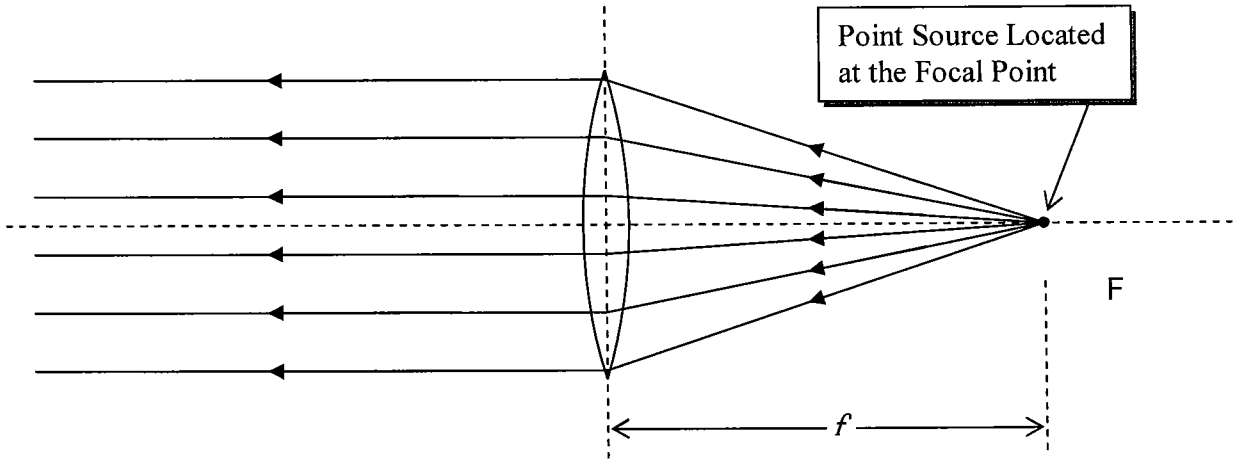
Note that in the diagram, we show the rays of light undergoing an abrupt change in direction *at the plane of the lens.* This is called the thin lens approximation and we will be using it in all our dealings with lenses. You know that the light is refracted twice in passing through a lens, once at the interface where it enters the lens medium, and again where it exits the lens medium. The two refractions together cause the incoming rays to travel in the directions in which they do travel. The thin lens approximation treats the pair of refractions as a single path change occurring at the

plane of the lens. The thin lens approximation is good as long as the thickness of the lens is small compared to the focal length, the object distance, and the image distance.

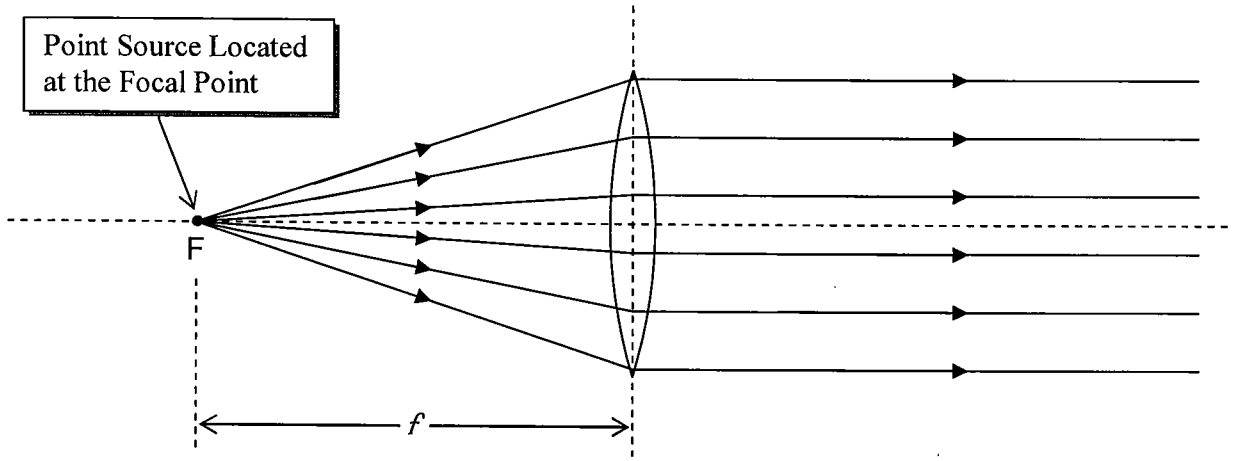
Rays parallel to the principal axis of the lens that enter the lens from the opposite direction (opposite the direction of the rays discussed above) will also be caused to converge to a focal point on the other side of the lens. The two focal points are one and the same distance f from the plane of the lens.



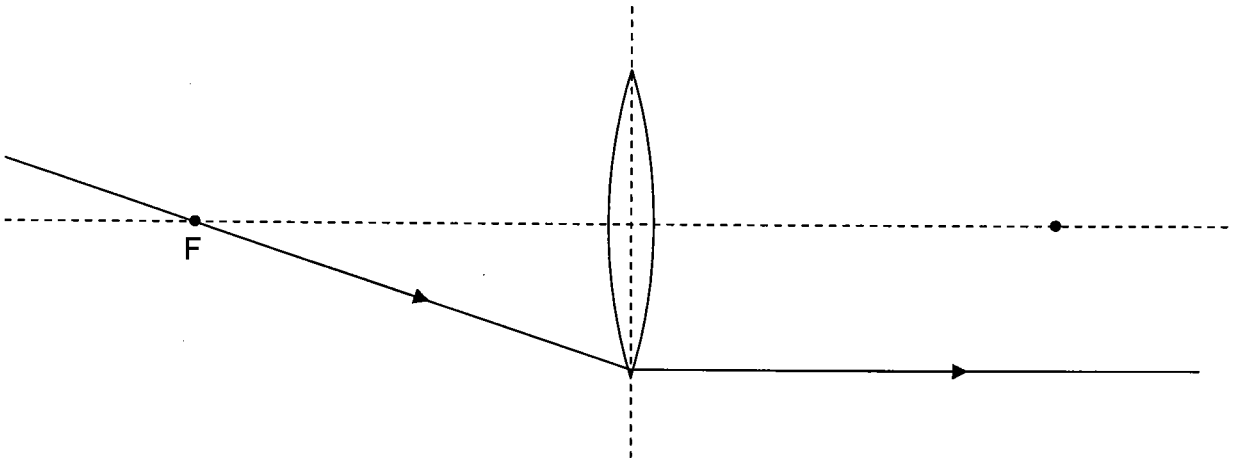
The two phenomena discussed above are reversible in the sense that rays of light coming from a point source, at either focal point, will result in parallel rays on the other side of the lens. Here we show that situation for the case of a point source at one of the focal points:



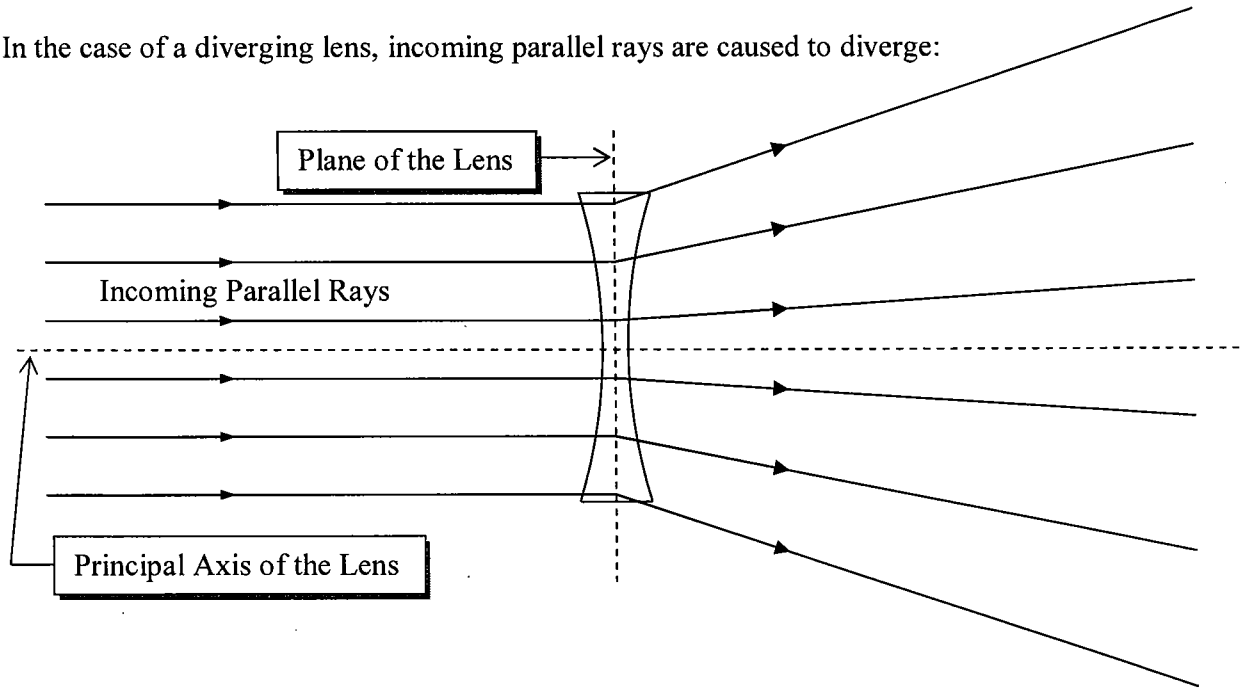
and here we show it for the case of a point source at the other focal point.



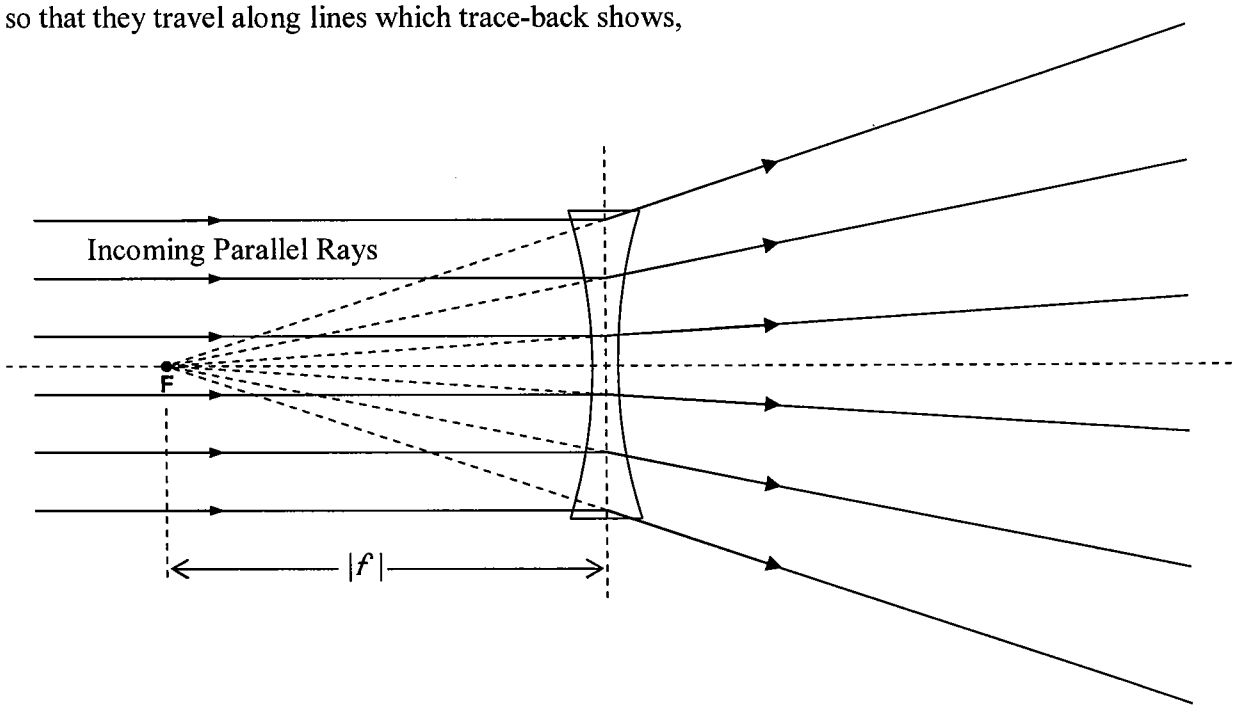
The important thing about this is that, any ray that passes through the focal point on its way to the lens is, after passing through the lens, going to be parallel to the principal axis of the lens.



In the case of a diverging lens, incoming parallel rays are caused to diverge:

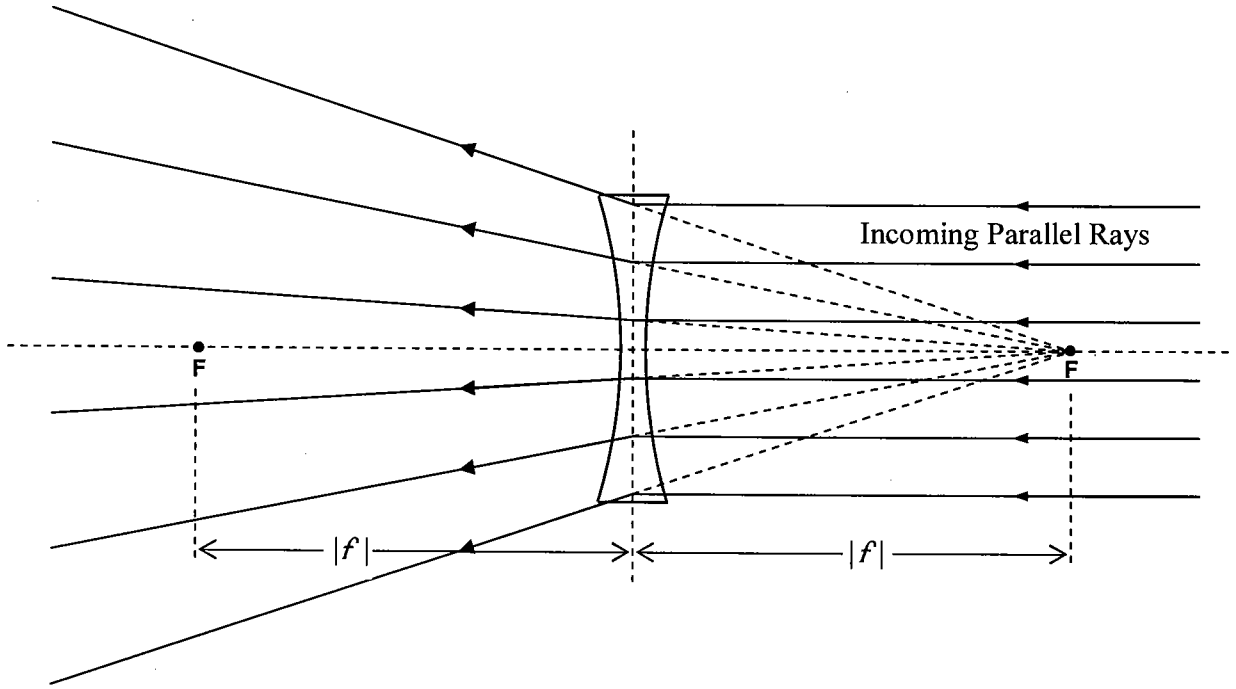


so that they travel along lines which trace-back shows,

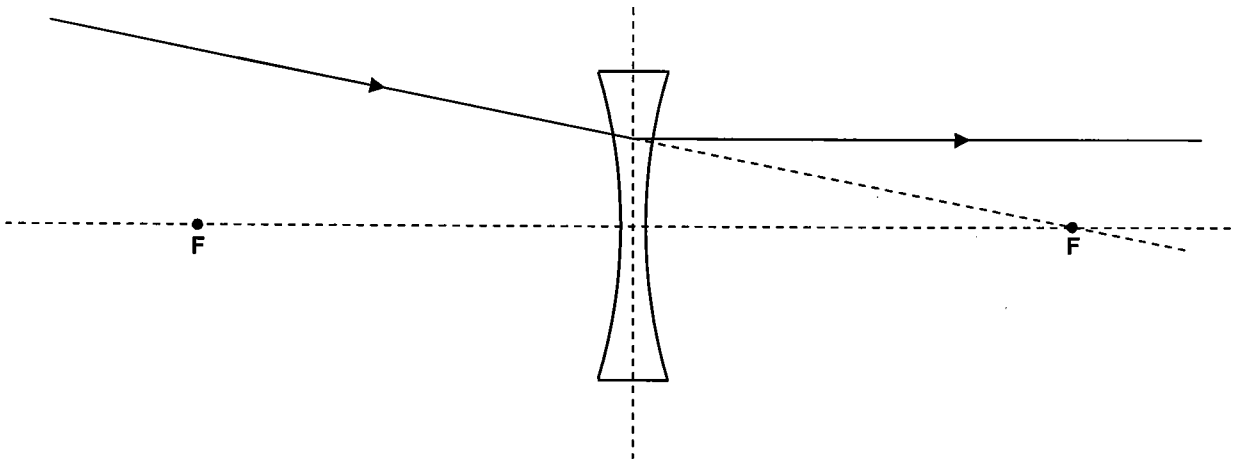


all pass through one and the same point. That is, on passing through the lens, the once-parallel rays diverge as if they originated from a point. That point is known as the *focal point* of the diverging lens. The distance from the plane of the lens to the focal point is the magnitude of the focal length of the lens. But, by convention, the focal length of a diverging lens is negative. In other words, the focal length of a *diverging lens* is the *negative* of the distance from the plane of the lens to the focal point.

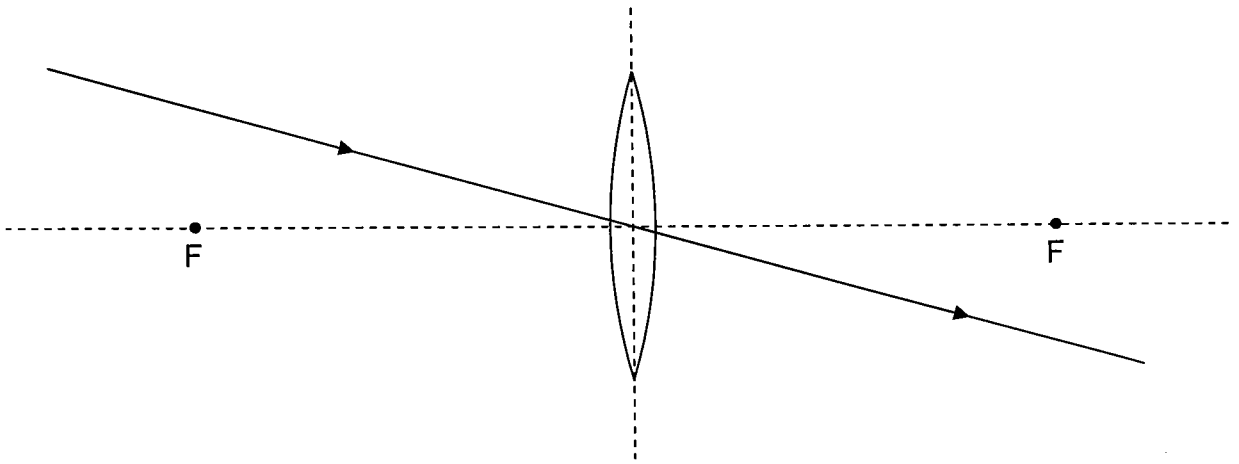
As in the case of the converging lens, there is another focal point on the other side of the lens, the same distance from the plane of the lens as the focal point discussed above:

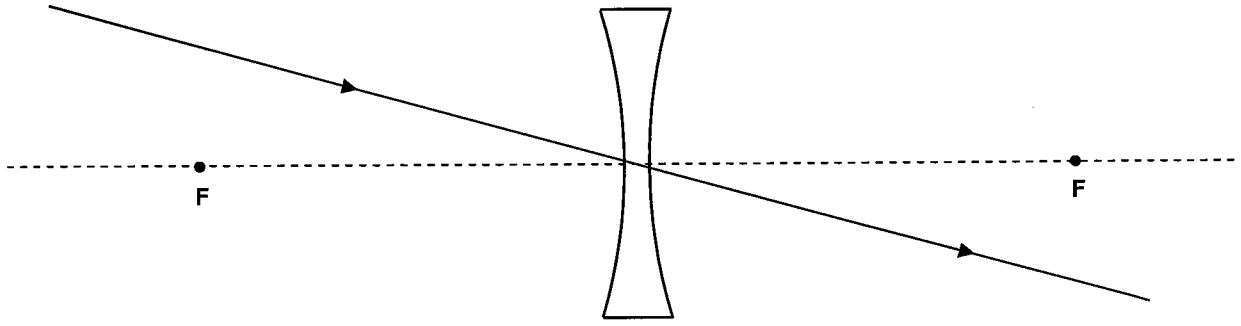


This effect is reversible in that any ray that is traveling through space on one side of the lens, and is *headed directly toward the focal point on the other side of the lens*, will, upon passing through the lens, *become parallel to the principal axis of the lens.*



Our plan here is to use the facts about what a lens does to incoming rays of light that are parallel to the principal axis of a lens or are heading directly toward or away from a focal point, to determine where a lens will form an image of an object. Before we do that, I need to tell you one more thing about both kinds of thin spherical lenses. This last fact is a reminder that our whole discussion is an approximation that hinges on the fact that the lenses we are dealing with are indeed *thin*. Here's the new fact: Any ray that is headed directly toward the center of a lens goes straight through. The justification is that at the center of the lens, the two surfaces of the lens are parallel. So, to the extent that they are parallel in a small region about the center of the lens, it is as if the light is passing through a thin piece of plate glass (or any transparent medium shaped like plate glass.) When light in air, is incident at some angle of incidence other than 0°, on plate glass, after it gets through both air/glass interfaces, the ray is parallel to the incoming ray. The amount by which the outgoing ray is shifted sideways, relative to the incoming ray, depends on how thick the plate is—the thinner the plate, the closer the outgoing ray is to being collinear with the incoming ray. In the thin lens approximation, we treat the outgoing ray as being *exactly* collinear with the incoming ray.

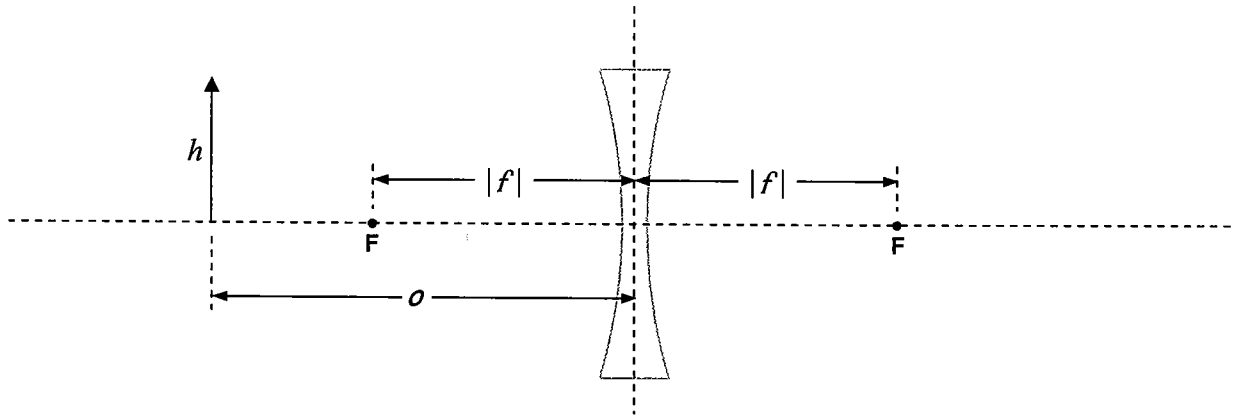




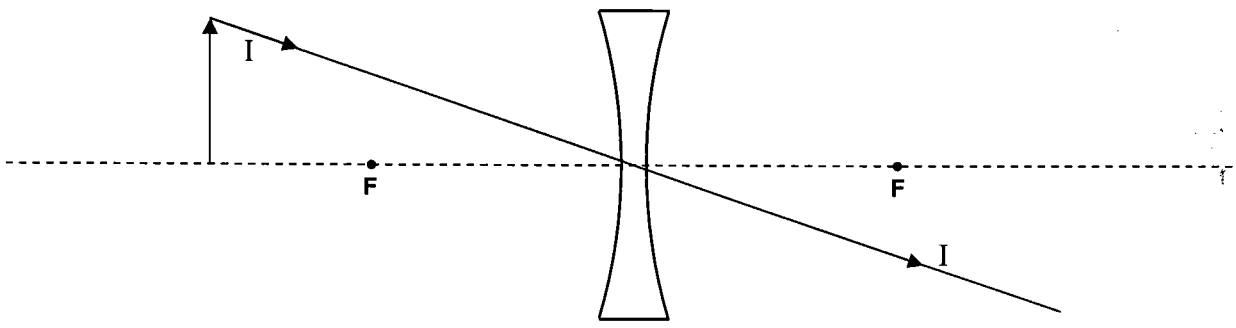
Using Ray Tracing Diagrams

Given an object of height h, the object position o, and the focal length f of the lens with respect to which the object position is given, you need to be able to diagrammatically determine: where the image of that object will be formed by the lens, how big the image is, whether the image is erect (right side up) or inverted (upside down), and whether the image is real or virtual (these terms will be defined soon). Here's how you do that for the case of a diverging lens of specified focal length for which the object distance $o > |f|$:

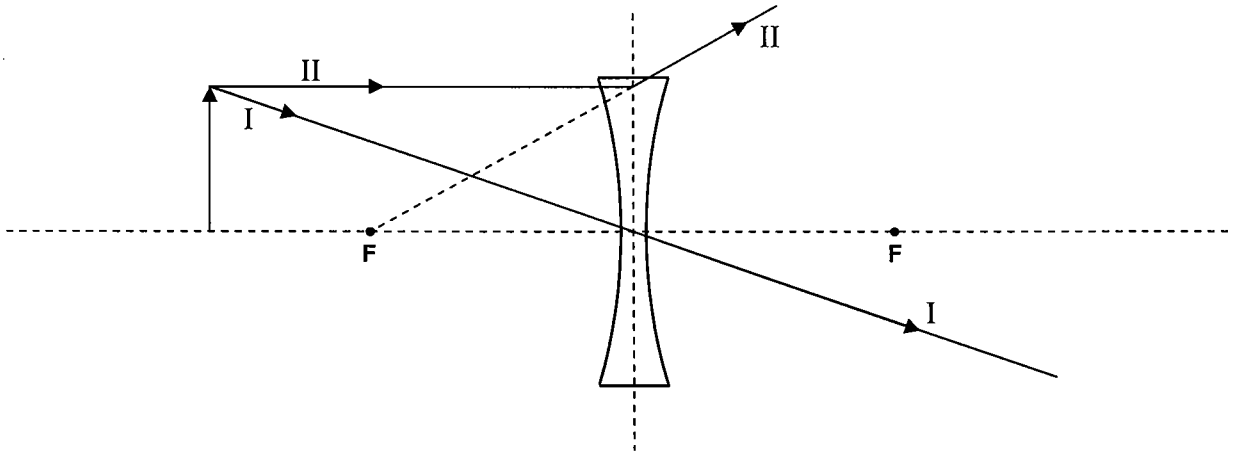
Draw the plane of the lens and the principal axis of the lens. Draw the lens, but think of it as an icon, just telling you what kind of lens you are dealing with. As you proceed with the diagram be careful not to show rays changing direction at the surface of your icon. Also, make sure you draw a diverging lens if the focal length is negative. Measure off the distance $|f|$ to both sides of the plane of the lens and draw the focal points. Measure off the object distance o from the plane of the lens, and, the height h of the object. Draw in the object.



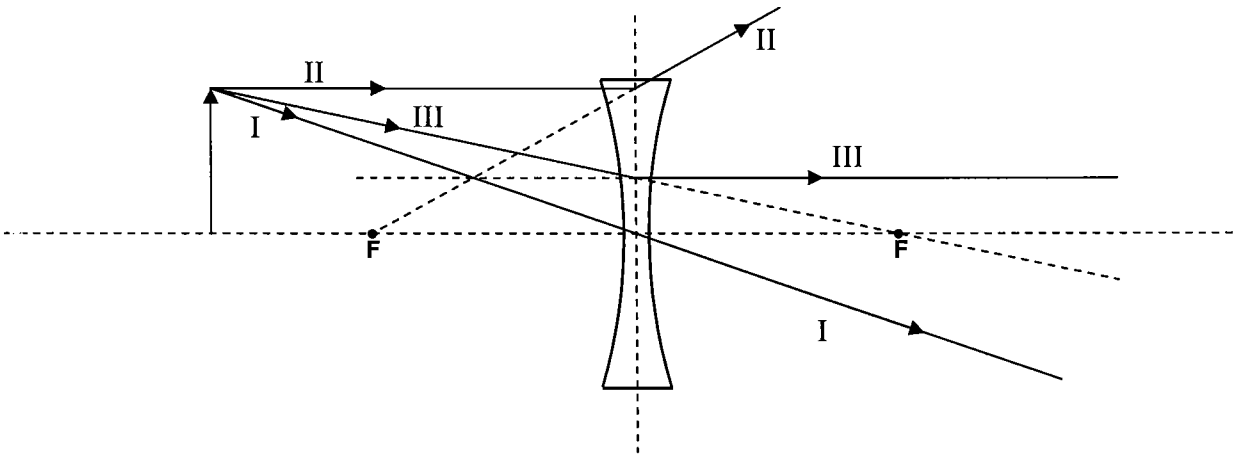
We determine the position of the image of the tip of the arrow by means of three principal rays. The three principal rays are rays on which the effect of the lens is easy to determine based on our understanding of what a lens does to incoming rays that are traveling toward the center of the lens, incoming rays that are traveling toward or away from a focal point, and incoming rays that are traveling directly toward the center of the lens. Let's start with the easy one, Principal Ray I. It leaves the tip of the arrow and heads directly toward the center of the lens. It goes straight through.



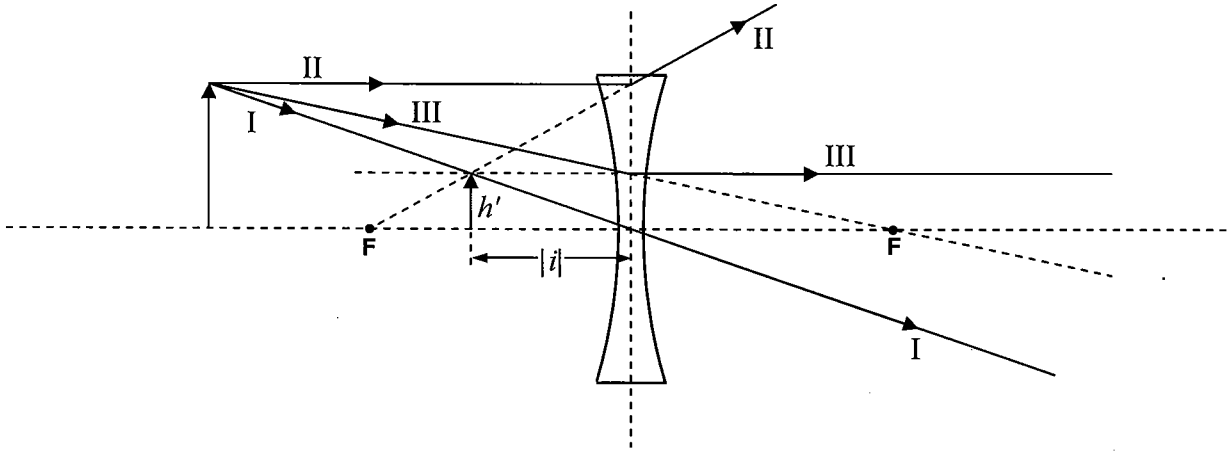
Next comes Principal Ray II. It comes in parallel to the principal axis of the lens, and, at the plane of the lens, jumps on a diverging line, which, if traced back, passes through the focal point on the same side of the lens as the object. Note the need for trace-back.



In the case of a diverging lens, Principal Ray III is the ray that, as it approaches the lens, is headed straight for the focal point on the *other side* of the lens. At the plane of the lens, Principal Ray III jumps onto a path that is parallel to the principal axis of the lens.



Note that, after passing through the lens, all three rays are diverging from each other. Trace-back yields the apparent point of origin of the rays, the image of the tip of the arrow. It is at the location where the three lines cross. (In practice, using a ruler and pencil, due to human error, the lines will cross at three different points. Consider these to be the vertices of a triangle and draw the tip of the arrow at what you judge to be the geometric center of the triangle.) Having located the image of the tip of the arrow, draw the shaft of the image of the arrow, showing that it extends from point of intersection, to the principal axis of the lens, and, that it is perpendicular to the principal axis of the lens.

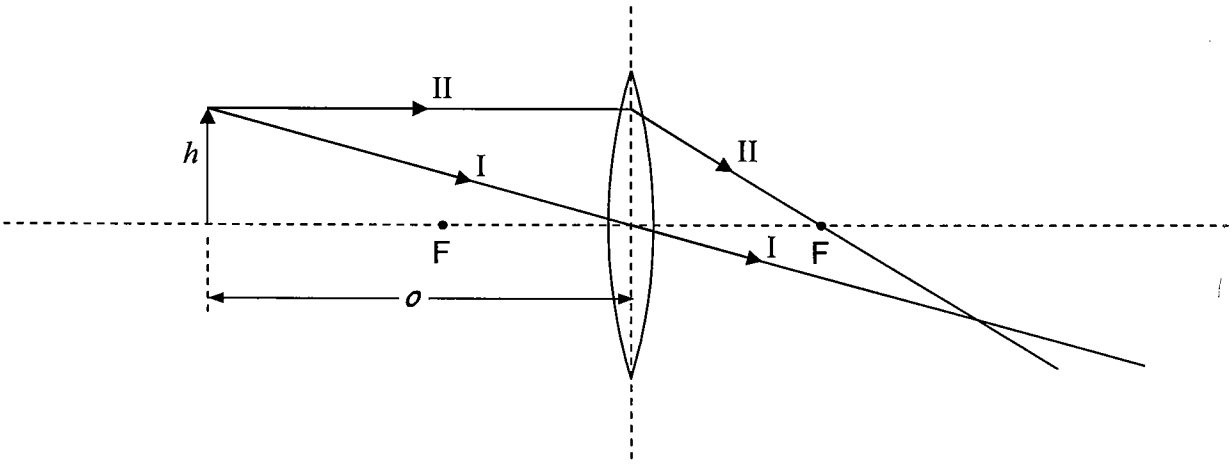


Measurements with a ruler yield the image height h' and the magnitude of the image distance $|i|$. The image is said to be a virtual image. A *virtual* image of a point, is a point from which rays appear to come, as determined by trace-back, but, through which the rays do not all, actually pass. By convention, the image distance is negative when the image is on the same side of the lens as the object. A negative image distance also signifies a virtual image. Note that the image is erect. By convention, an erect image has a positive image height h'. The magnification M is given by:

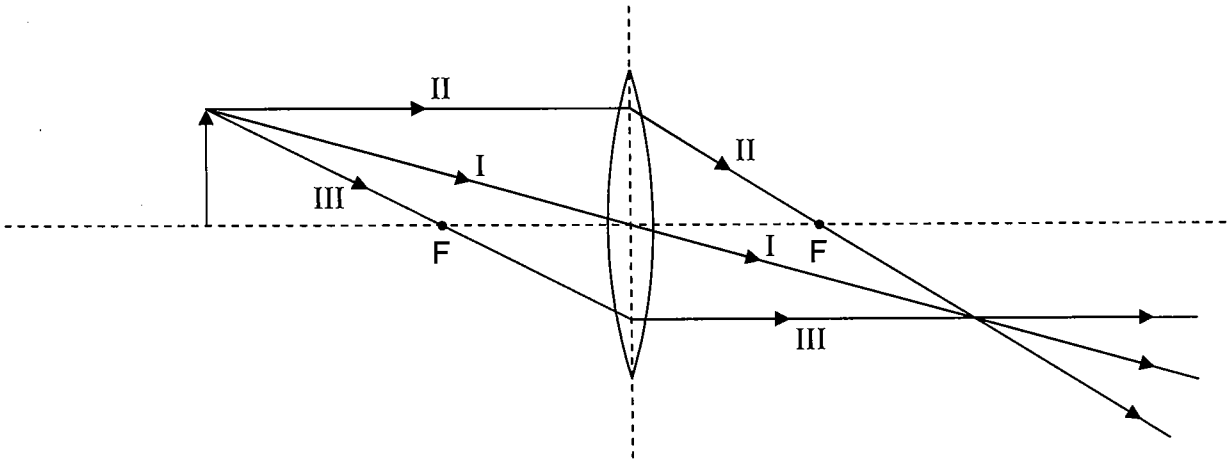
$$M = \frac{h'}{h}$$

By convention, a positive value of M means the image is erect (right side up).

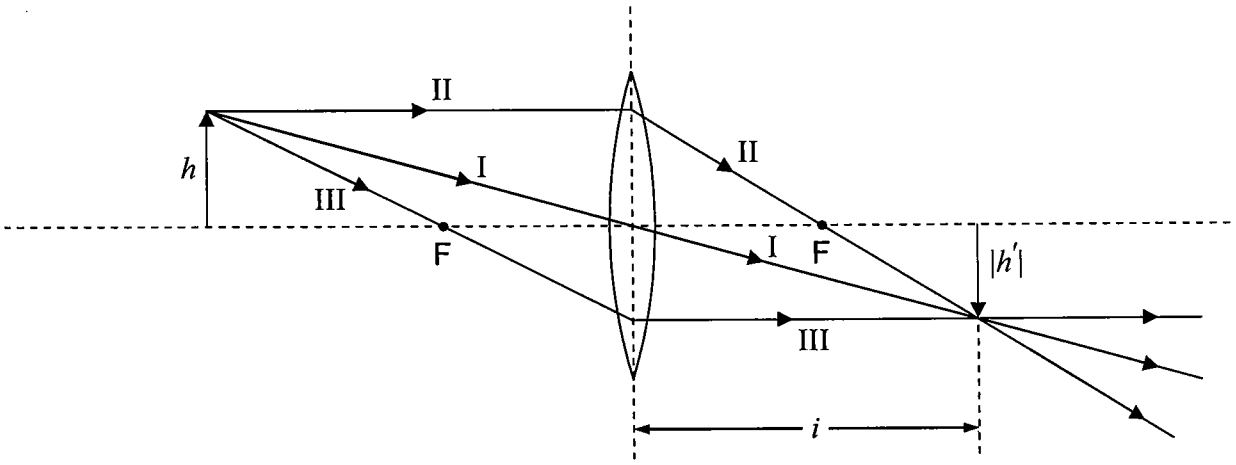
For the case of a converging lens, Principal Ray I is identical to the corresponding ray for the diverging lens. It starts out headed straight for the center of the lens, and, it goes straight through. Principal Ray II starts out the same way Principal Ray II did for the diverging lens—it comes in parallel to the principal axis of the lens—but, starting at the plane of the lens, rather than diverging, it is caused to converge to the extent that it passes through the focal point on the other side of the lens.



Principal Ray III, for a converging lens (with the object farther from the lens than the focal point is), passes through the focal point on the *same side* of the lens (the side of the lens the object is on) and then, when it gets to the plane of the lens, comes out parallel to the principal axis of the lens.



If you position yourself so that the rays, having passed through the lens, are coming at you, and, you are far enough away from the lens, you will again see the rays diverging from a point. But this time, all the rays actually go through that point. That is, the lens converges the rays to a point, and they don't start diverging again until after they pass through that point. That point is the image of the tip of the arrow. It is a real image. You can tell because if you trace back the lines the rays are traveling along, you come to a point through which all the rays actually travel. Identifying the crossing point as the tip of the arrow, we draw the shaft and head of the arrow.

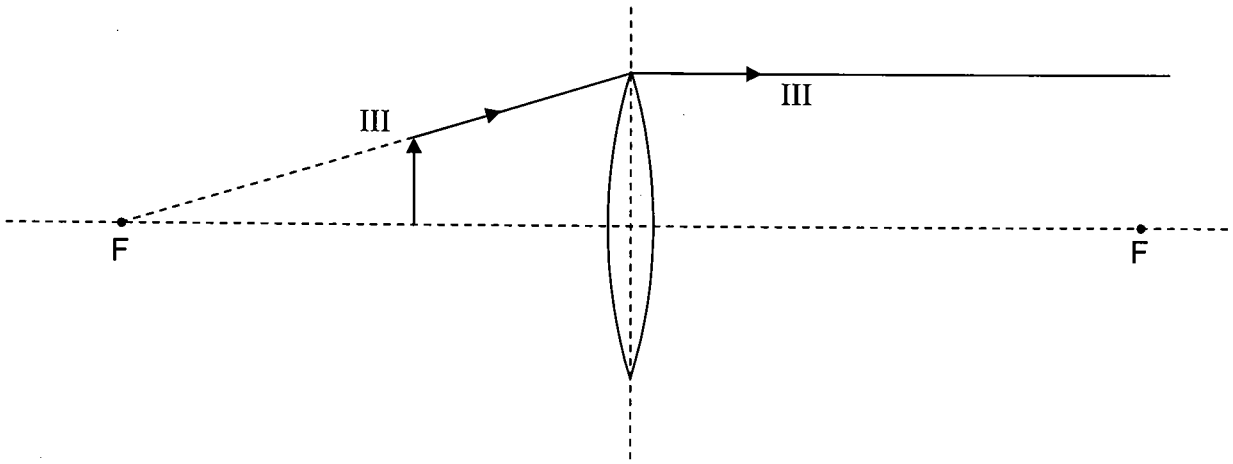


This time, the image is inverted. We can measure the length of the image and the distance of the image from the plane of the mirror. By convention, the image height is negative when the image is inverted, and, the image distance is positive when the image is on the side of the lens opposite that of the object. The magnification M is again given by

$$M = \frac{h'}{h} \tag{28-2}$$

which, with h' being negative, turns out to be negative itself. This is consistent with the convention that a negative magnification means the image is inverted.

Principal Ray III is different for the converging lens when the object is closer to the plane of the lens than the focal point is:



Principal Ray III, like every principal ray, starts at the tip of the object and travels toward the plane of the lens. In the case at hand, on its way to the plane of lens, Principal Ray III travels along a line that, if traced back, passes through the focal point on the same side of the lens as the object.

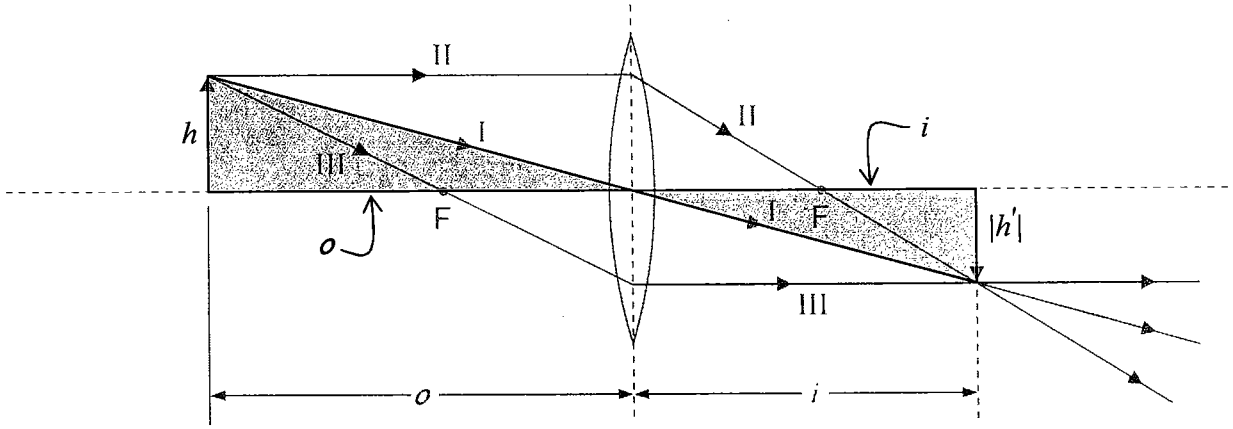
This concludes our discussion of the determination of image features and position by means of ray tracing. In closing this chapter, I summarize the algebraic sign conventions, in the form of a table:

Physical Quantity	Symbol	Sign Convention
focal length	f	+ for converging lens – for diverging lens
image distance	i	+ for real image (on opposite side of lens as object) – for virtual image (on same side of lens as object)
image height	h'	+ for erect image – for inverted image
magnification	M	+ for erect image – for inverted image

29 Thin Lenses: Lens Equation, Optical Power

From the thin lens ray-tracing methods developed in the last chapter, we can derive algebraic expressions relating quantities such as object distance, focal length, image distance, and magnification.

Consider for instance the case of a converging lens with an object more distant from the plane of the lens than the focal point is. Here's the diagram from the last chapter. In this copy, I have shaded two triangles in order to call your attention to them. Also, I have labeled the sides of those two triangles with their lengths.



By inspection, the two shaded triangles are similar to each other. As such, the ratios of corresponding sides are equal. Thus:

$$\frac{|h'|}{h} = \frac{i}{o}$$

Recall the conventions stated in the last chapter:

Physical Quantity	Symbol	Sign Convention
focal length	f	+ for converging lens – for diverging lens
image distance	i	+ for real image – for virtual image
image height	h'	+ for erect image – for inverted image
magnification	M	+ for erect image – for inverted image

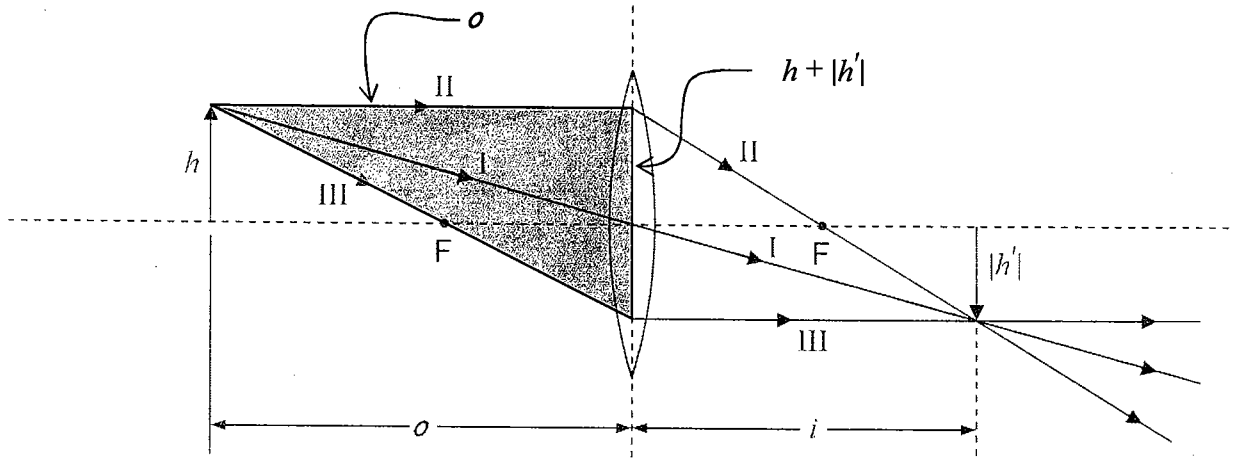
In the case at hand, we have an inverted image, so h' is negative, so $|h'| = -h'$. Thus, the equation $\frac{|h'|}{h} = \frac{i}{o}$ can be written as $\frac{-h'}{h} = \frac{i}{o}$, or, as

$$\frac{h'}{h} = -\frac{i}{o}$$

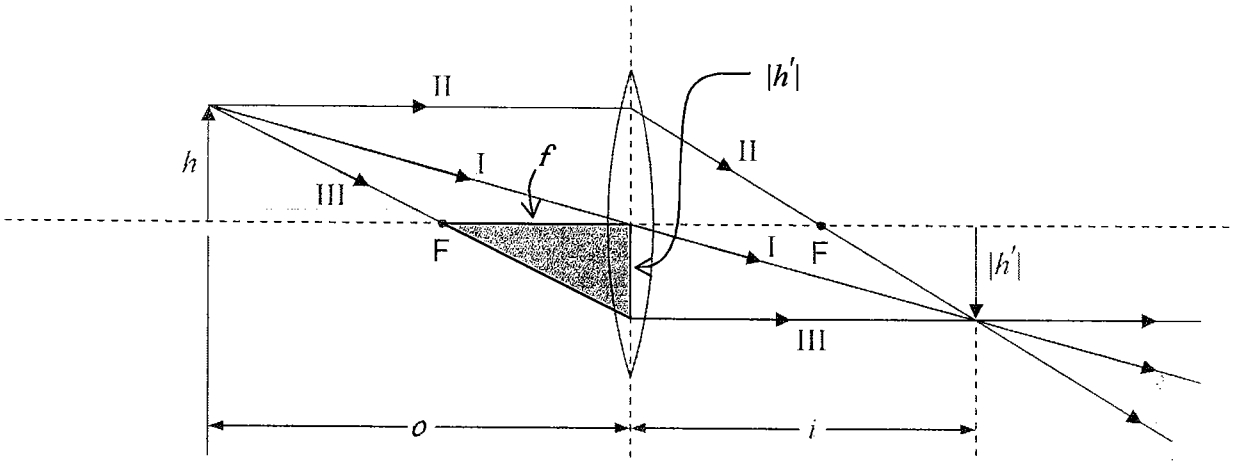
But $\frac{h'}{h}$ is, by definition, the magnification. Thus, we can write the magnification as:

$$M = -\frac{i}{o} \qquad (29\text{-}1)$$

Here's another copy of the same diagram with another triangle shaded.



By inspection, that shaded triangle is *similar* to the triangle that is shaded in the following copy of the same diagram:



Using the fact that the ratios of corresponding sides of similar triangles are equal, we set the ratio of the two top sides (one from each triangle) equal to the ratio of the two right sides:

$$\frac{o}{f} = \frac{h + |h'|}{|h'|}$$

Again, since the image is upside down, h' is negative so $|h'| = -h'$. Thus,

$$\frac{o}{f} = \frac{h - h'}{-h'}$$

$$\frac{o}{f} = 1 - \frac{h}{h'}$$

From our first pair of similar triangles we found that $\frac{h'}{h} = -\frac{i}{o}$ which can be written $\frac{h}{h'} = -\frac{o}{i}$

Substituting this into the expression $\frac{o}{f} = 1 - \frac{h}{h'}$ which we just found, we have

$$\frac{o}{f} = 1 - \left(-\frac{o}{i}\right)$$

Dividing both sides by o and simplifying yields:

$$\frac{1}{f} = \frac{1}{o} + \frac{1}{i} \qquad (29\text{-}2)$$

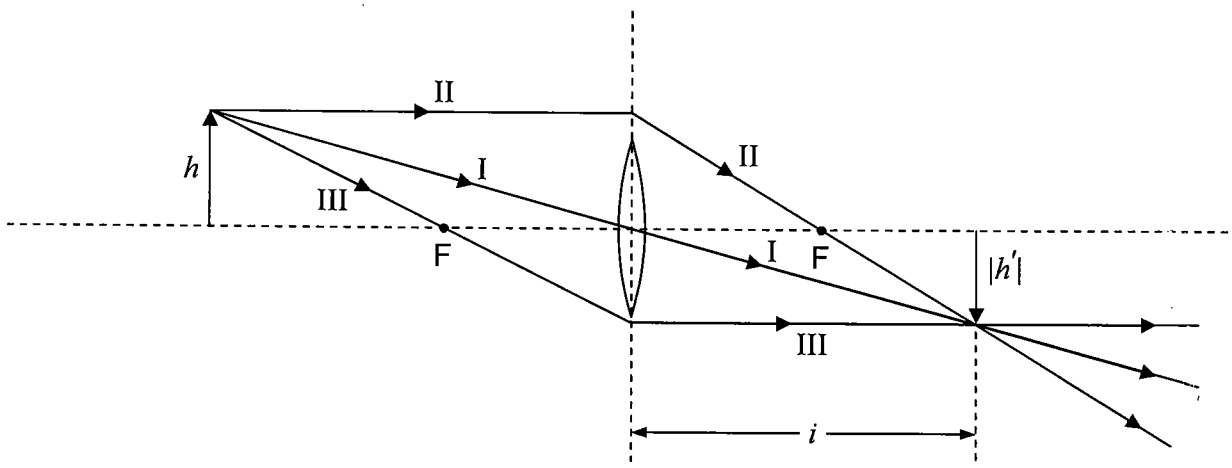
This equation is referred to as *the lens equation.* Together with our definition of the magnification $M = \frac{h'}{h}$, the expression we derived for the magnification $M = -\frac{i}{o}$, and our conventions:

Physical Quantity	Symbol	Sign Convention
focal length	f	+ for converging lens – for diverging lens
image distance	i	+ for real image – for virtual image
image height	h'	+ for erect image – for inverted image
magnification	M	+ for erect image – for inverted image

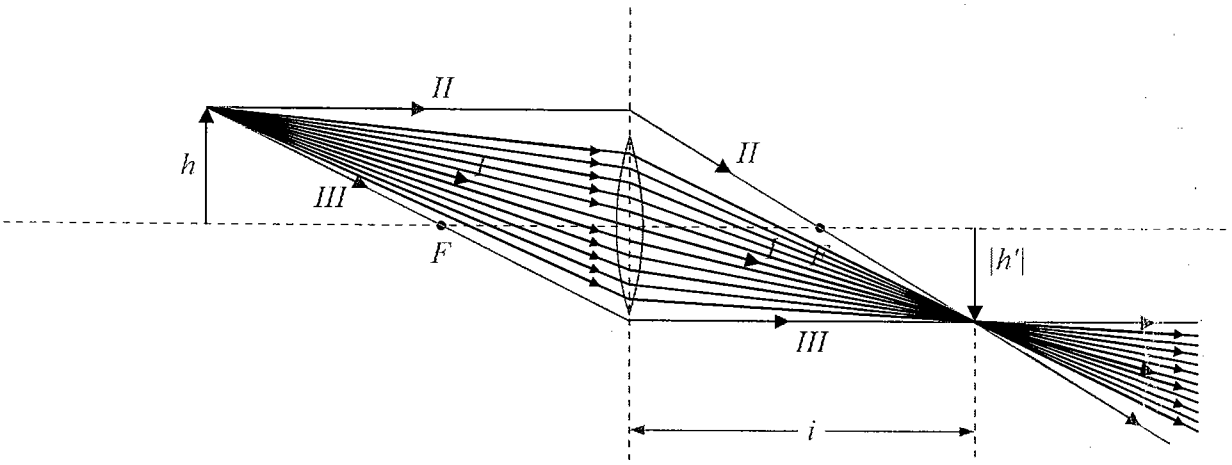
the lens equation tells us everything we need to know about the image of an object that is a known distance from the plane of a thin lens of known focal length. While we have derived it for the case of an object that is a distance greater than the focal length, from a converging lens, it works for *all* the combinations of lens and object distance for which the thin lens approximation is good. (The thin lens approximation is good as long as i, o, and f are all large compared to the thickness of the lens.) In each case, we derive the lens equation (it always turns out to be the same equation), by drawing the ray tracing diagram and analyzing the similar triangles that appear in it.

An Important Conceptual Point

(We mentioned this in the last chapter but it warrants further attention.) An infinite set of rays contributes to any given point of an image formed by a lens. Consider for instance the case of an object at a greater distance than the focal length from a thin spherical convex (converging) lens. Further, because it's easy to specify, we will consider the image of the tip of the (arrow) object. We have been using the principal rays to locate the image, as in the following diagram:



in which I have intentionally used a small lens icon to remind you that, in using the principal ray diagram to locate the image, we don't really care whether or not the principal rays actually hit the lens. Let's, for the case at hand, consider the diagram to be a life-size diagram of an actual lens. As, important as they are in helping us identify the location of the image, clearly, for the case at hand, Principal Rays II and III do not actually contribute to the image. Principal Ray I does contribute to the image. Let's draw in some more of the contributors:



The fact that every ray that comes from the tip of the object and hits the lens contributes to the image of the tip of the arrow (and the corresponding fact for each and every point on the object) explains why you can cover up a fraction of the lens (such as half the lens) and still get a complete image (albeit dimmer).

The Power of a Lens

When an ophthalmologist writes a prescription for a spherical lens, she or he will typically write either a value around −.5 or .5, or, a value around −500 or 500 without units. You might well wonder what quantity the given number is a value for, and what the units should be. The answer to the first question is that the physical quantity is the *power of the lens* being prescribed. In this context, the power is sometimes called the *optical power of the lens.* The power of a lens has nothing to do with the rate at which energy is being transformed or transferred but instead represents the assignment of a completely different meaning to the same word. In fact, the power of a lens is, by definition, the reciprocal of the focal length of the lens:

$$P = \frac{1}{f} \tag{29-3}$$

In that the SI unit of focal length is the meter (m), the unit of optical power is clearly the reciprocal meter which you can write as $\frac{1}{\text{m}}$ or m^{-1} in accord with your personal preferences. This unit has been assigned a name. It is called the diopter, abbreviated D. Thus, by definition,

$$1\text{D} = \frac{1}{\text{m}}$$

Thus, a value of −.5 on the ophthalmologist's prescription can be interpreted to mean that what is being prescribed is a lens having a power of −0.5 diopters. The minus sign means that the lens is a concave (diverging) lens. Taking the reciprocal yields:

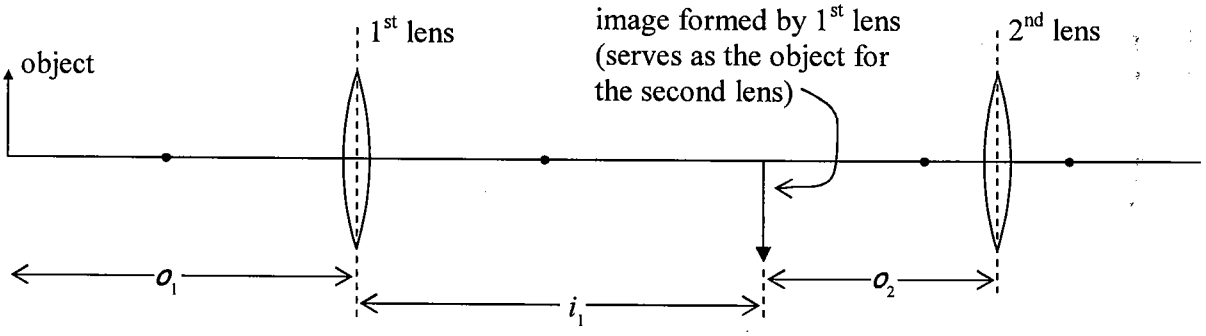
$$f = \frac{1}{P} = \frac{1}{-0.5\,\text{D}} = -2\frac{1}{\text{D}} = -2\,\text{m}$$

If you see a number around −500 or 500 on the ophthalmologist's lens prescription, you can assume that the ophthalmologist is giving the power of the lens in units of millidiopters (mD). 500 mD is, of course, equivalent to .5 D. To avoid confusion, if you are given an optical power in units of mD, convert it to units of diopters before using it to calculate the corresponding focal length.

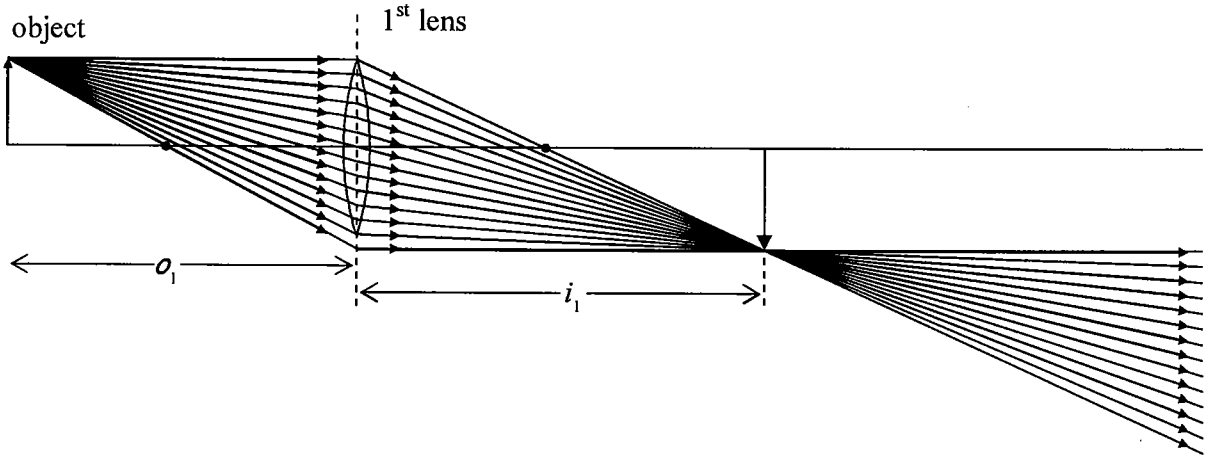
Two-Lens Systems

To calculate the image of a two-lens system, one simply calculates the position of the image for the lens that light from the object hits first, and then uses that image as the object for the second lens. In general, one has to be careful to recognize that for the first lens, the object distance and the image distance are both measured relative to the plane of the first lens. Then, for the second lens, the object distance and the image distance are measured relative to the plane of the *second* lens. That means that, in general, the object distance for the second lens is not equal in value to the image distance for the first lens.

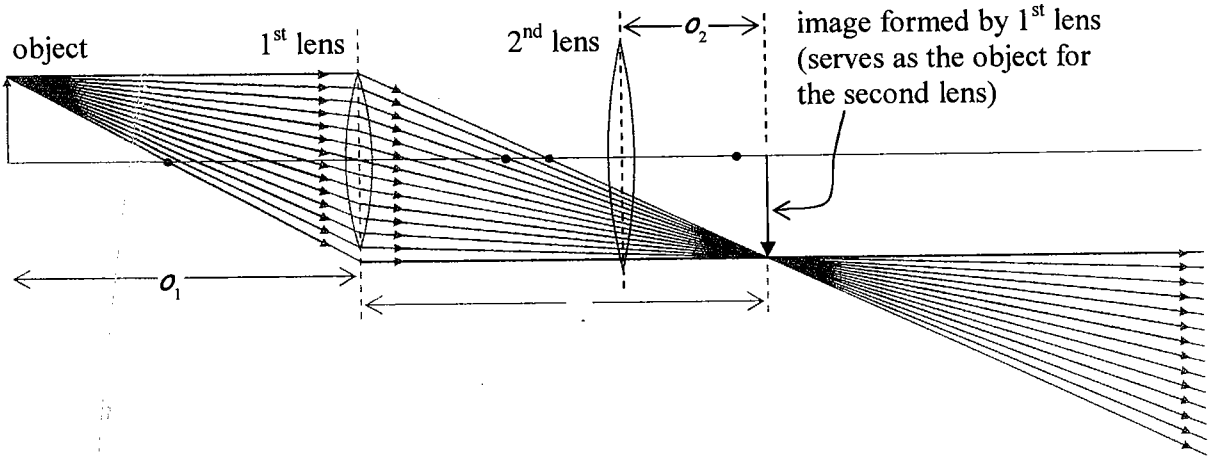
For instance, in the following diagram of two lenses separated by 12 cm, if the object is to the left of the first lens, and i_1 turns out to be 8 cm to the right of the first lens,



then o_2, the object distance for the second lens, is 4 cm. A peculiar circumstance arises when the second lens is closer to the first lens than the image formed by the first lens is. Suppose for instance, that we have the image depicted above, formed by the first lens:



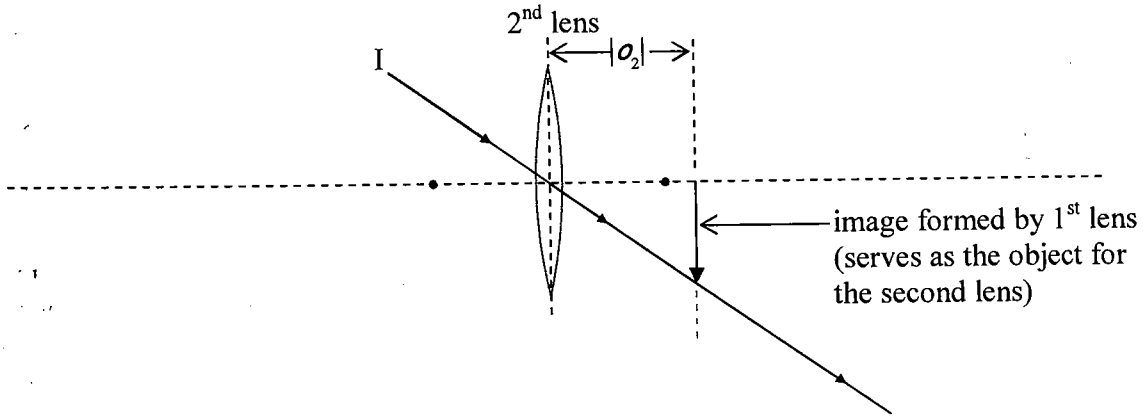
Now suppose that we put a second lens in between the 1st lens and the image.



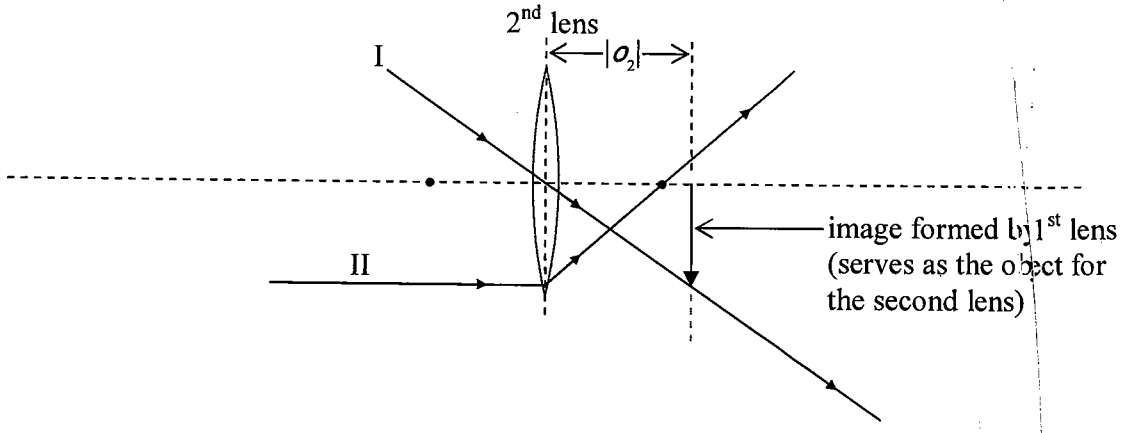
Note that, for the second lens, we have an object to the right of the lens, but, the light associated with that object approaches the object from the left! This can only happen when the object is actually an image formed by another lens. In such a case, we call the object a virtual object. More generally, when an object's light approaches a lens from the side opposite that side to which the object is, the object is considered to be a virtual object, and, the object distance, is, by convention, negative. So, we have one more convention to put in a table for you:

Physical Quantity	Symbol	Sign Convention
Object Distance	o	+ for real object (always the case for a physical object) – for virtual object (only possible if "object" is actually the image formed by another lens)

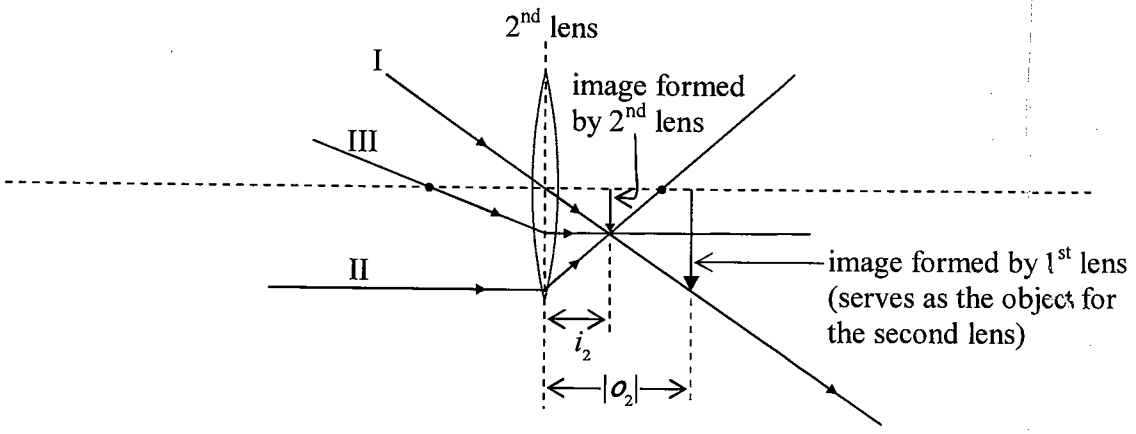
In forming the ray-tracing diagram for the case of the virtual object, we have to remember that every ray coming into the second lens is headed straight for the tip of the arrow that is the virtual object for the second lens. Thus, our Principal Ray I is one that is headed straight toward the tip of the arrow, and, is headed straight toward the center of the lens. It goes straight through.



Principal Ray II is headed straight for the head of the object along a line that is parallel to the principal axis of the lens. At the plane of the lens it jumps onto the straight line path that takes it straight through the focal point on the other side of the lens.

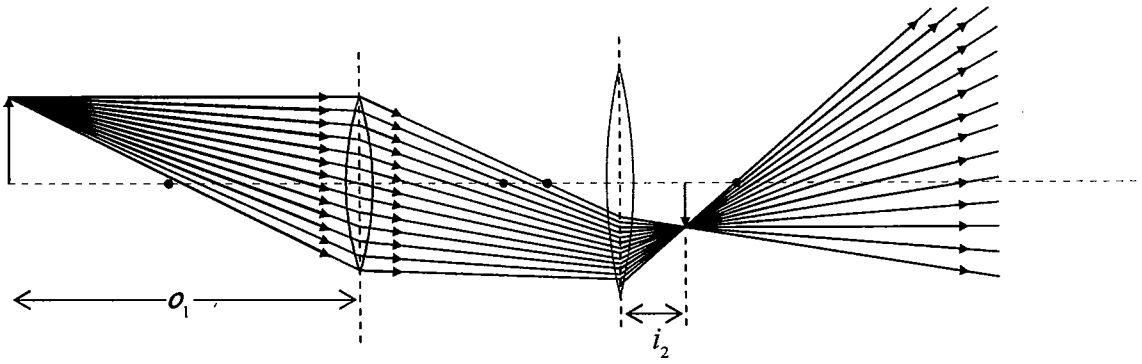


Principal Ray III, is headed straight toward the tip of the virtual object, and, on its way to the lens, it passes through the focal point on the side of the lens from which it approaches the lens. When it hits the plane of the lens, Principal Ray III adopts a path that is parallel to the principal axis of the lens.



Note that, for the case at hand, we get a real image. Relative to the virtual object, the image is not inverted. The virtual object was already upside down. The fact that we can draw a ray-tracing diagram for the case of a virtual object means that we can identify and analyze similar triangles to establish the relationship between the object distance, the image distance and the focal length of the lens. Doing so, with the convention that the object distance of a virtual object is negative, again yields the lens equation $\frac{1}{f}=\frac{1}{o}+\frac{1}{i}$.

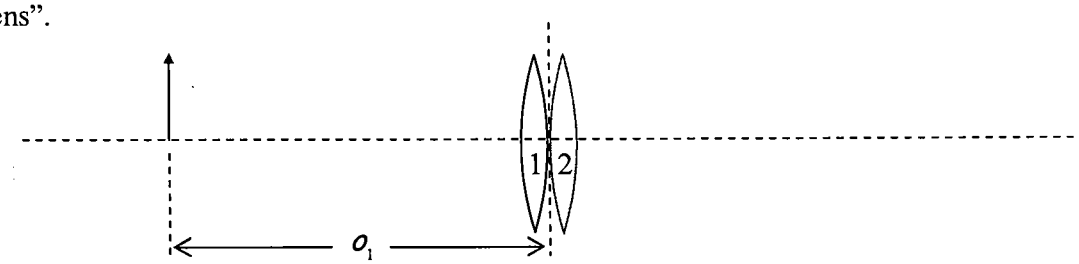
Here's a diagram of the entire two-lens system for the case at hand:



Note that the real image of lens 1 alone is never actually formed, but it was crucial in our determination of the image location, orientation, and size, in the case of the two-lens system.

Two Lenses at (Essentially) the Same Location

In the thin lens approximation (in which we consider the thickness of a lens to be negligible), two lenses placed in contact with one another are considered to have one and the same "plane of the lens".



As such, the object distance for the second lens is the negative of the image distance for the first lens. The second lens forms an image in accord with:

$$\frac{1}{f_2} = \frac{1}{o_2} + \frac{1}{i_2}$$

in which, as we just stated, the object distance o_2 is the negative of the image distance i_1 for the first lens. Substituting $o_2 = -i_1$ yields

$$\frac{1}{f_2} = \frac{1}{-i_1} + \frac{1}{i_2}$$

i_1 represents the image distance for the image that would be formed by the first lens if the second lens wasn't there. It is related to the focal length and object distance by the lens equation:

$$\frac{1}{f_1} = \frac{1}{o_1} + \frac{1}{i_1}$$

Solving this for $\frac{1}{-i_1}$ yields $\frac{1}{-i_1}=\frac{1}{o_1}-\frac{1}{f_1}$ which, when substituted into $\frac{1}{f_2}=\frac{1}{-i_1}+\frac{1}{i_2}$ from above, yields:

$$\frac{1}{f_2}=\frac{1}{o_1}-\frac{1}{f_1}+\frac{1}{i_2}$$

which can be written as:

$$\frac{1}{f_1}+\frac{1}{f_2}=\frac{1}{o_1}+\frac{1}{i_2}$$

The object distance o_1 for the first lens is the object distance for the pair of lenses. I'm going to call that o_c for the object distance for the *combination* of two lenses. In other words, $o_c \equiv o_1$. The image distance for the *second* lens is the image distance for the pair two lenses. I'm going to call that i_c for the image distance for the *combination* of two lenses. In other words, $i_c \equiv i_2$. Thus, on the right, we have the reciprocal of the object distance for the two-lens system plus the reciprocal of the image distance for the two-lens system.

$$\frac{1}{f_1}+\frac{1}{f_2}=\frac{1}{o_c}+\frac{1}{i_c}$$

On the left we have the sum of the powers of the two lenses.

$$P_1+P_2=\frac{1}{o_c}+\frac{1}{i_c}$$

For a single lens, the lens equation $\frac{1}{f}=\frac{1}{o}+\frac{1}{i}$ expressed in terms of the power, reads $P=\frac{1}{o}+\frac{1}{i}$. Thus, if, for our combination of two lenses at one and the same location, we identify P_1+P_2 as the power P_c of the combination of two lenses, we have,

$$P_c=\frac{1}{o_c}+\frac{1}{i_c}$$

Thus, a pair of lenses, each of which is at essentially one and the same location, acts as a single lens whose power P_c is the sum of the powers of the two lenses making up the combination.

$$P_c=P_1+P_2 \qquad (29\text{-}4)$$

30 The Electric Field Due to a Continuous Distribution of Charge on a Line

Every integral must include a differential (such as dx, dt, dq, etc.). An integral is an infinite sum of terms. The differential is necessary to make each term infinitesimal (vanishingly small). $\int f(x)\,dx$ *is okay,* $\int g(y)\,dy$ *is okay, and* $\int h(t)\,dt$ *is okay, but never write* $\int f(x)$*, never write* $\int g(y)$ *and never write* $\int h(t)$*.*

Here we revisit Coulomb's Law for the Electric Field. Recall that Coulomb's Law for the Electric Field gives an expression for the electric field, at an empty point in space, due to a charged particle. You have had practice at finding the electric field at an empty point in space due to a single charged particle and due to several charged particles. In the latter case, you simply calculated the contribution to the electric field at the one empty point in space due to each charged particle, and then added the individual contributions. You were careful to keep in mind that each contribution to the electric field at the empty point in space was an electric field vector, a vector rather than a scalar, hence the individual contributions had to be added like vectors.

A Review Problem for the Electric Field due to a Discrete[1] Distribution of Charge

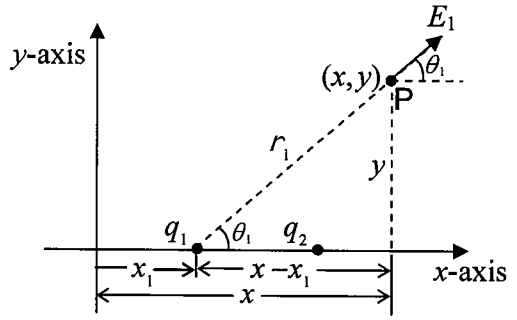
Let's kick this chapter off by doing a review problem. The following example is one of the sort that you learned how to do when you first encountered Coulomb's Law for the Electric Field. You are given a discrete distribution of source charges and asked to find the electric field (in the case at hand, just the x component of the electric field) at an empty point in space.

The example is presented on the next page. Here, a word about one piece of notation used in the solution. The symbol P is used to identify a point in space so that the writer can refer to that point, unambiguously, as "point P." The symbol P in this context does not stand for a variable or a constant. It is just an identification tag. It has no value. It cannot be assigned a value. It does not represent a distance. It just labels a point.

[1] The charge distribution under consideration here is called a discrete distribution as opposed to a continuous distribution because it consists of several individual particles that are separated from each other by some space. A continuous charge distribution is one in which some charge is "smeared out" along some line or over some region of space.

Example 30-1 (A Review Problem)

There are two charged particles on the x-axis of a Cartesian coordinate system, q_1 at $x = x_1$ and q_2 at $x = x_2$ where $x_2 > x_1$. Find the x component of the electric field, due to this pair of particles, valid for all points on the x-y plane for which $x > x_2$.

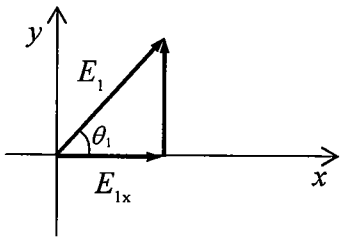


$\vec{\mathbf{E}}_1$ is the contribution to the electric field at point P (at x, y) due to charge q_1. Charge q_2 contributes $\vec{\mathbf{E}}_2$ to the electric field at P.

$$\vec{\mathbf{E}} = \vec{\mathbf{E}}_1 + \vec{\mathbf{E}}_2$$

$$E_x = E_{1x} + E_{2x}$$

First, let's get E_{1x}:



$$\frac{E_{1x}}{E_1} = \cos\theta_1$$

$$E_{1x} = E_1 \cos\theta_1$$

Looking at the diagram at the top of this column, we see that Coulomb's Law for the Electric Field yields:

$$E_1 = \frac{k q_1}{r_1^2}$$

$$E_{1x} = \frac{k q_1}{r_1^2}\cos\theta_1$$

Again, from that first diagram,

$$r_1 = \sqrt{(x - x_1)^2 + y^2}$$

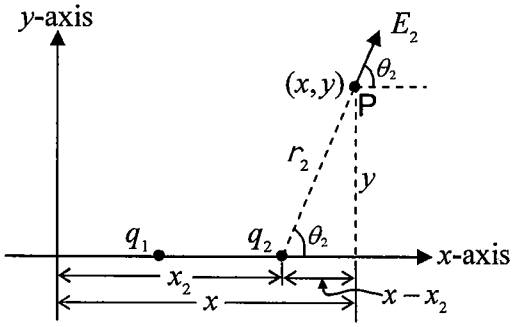
and

$$\cos\theta_1 = \frac{x - x_1}{r_1} = \frac{x - x_1}{\sqrt{(x - x_1)^2 + y^2}}$$

Substituting both of these into $E_{1x} = \frac{k q_1}{r_1^2}\cos\theta_1$ yields:

$$E_{1x} = \frac{k q_1}{\left(\sqrt{(x - x_1)^2 + y^2}\right)^2}\frac{x - x_1}{\sqrt{(x - x_1)^2 + y^2}}$$

$$E_{1x} = \frac{k q_1 (x - x_1)}{\left[(x - x_1)^2 + y^2\right]^{3/2}}$$



It is left as an exercise for the reader to show that:

$$E_{2x} = \frac{k q_2 (x - x_2)}{\left[(x - x_2)^2 + y^2\right]^{3/2}}$$

Since $E_x = E_{1x} + E_{2x}$, we have:

$$E_x = \frac{k q_1 (x - x_1)}{\left[(x - x_1)^2 + y^2\right]^{3/2}} + \frac{k q_2 (x - x_2)}{\left[(x - x_2)^2 + y^2\right]^{3/2}}$$

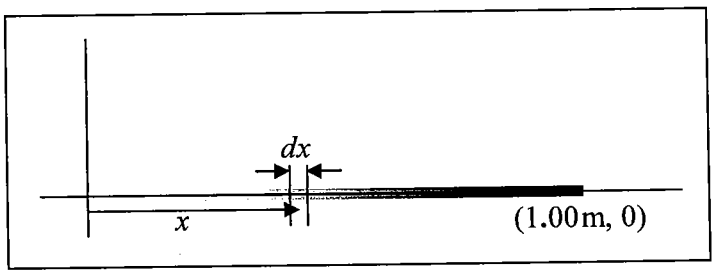
Linear Charge Density

Okay, enough review, now lets consider the case in which we have a continuous distribution of charge along some line segment. In practice, we could be talking about a charged piece of string or thread, a charged thin rod, or even a charged piece of wire. First we need to discuss how one even specifies such a situation. We do so by stating what the linear charge density, the charge-per-length, λ is. For now we'll consider the meaning of λ for a few different situations (before we get to the heart of the matter, finding the electric field due to the linear charge distribution). Suppose for instance we have a one-meter string extending from the origin to $x = 1.00$ m along the x axis, and that the linear charge density on that string is given by:

$$\lambda = 2.56\frac{\mu C}{m^2}x.$$

(Just under the equation, we have depicted the linear charge density graphically by drawing a line whose darkness represents the charge density.)

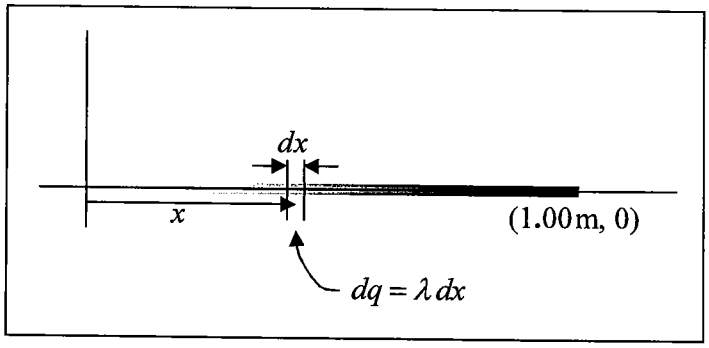
Note that if the value of x is expressed in meters, λ will have units of $\frac{\mu C}{m}$, units of charge-per-length, as it must. Further note that for small values of x, λ is small, and for larger values of x, λ is larger. That means that the charge is more densely packed near the far (relative to the origin) end of the string. To further familiarize ourselves with what λ is, let's calculate the total amount of charge on the string segment. What we'll do is to get an expression for the amount of charge on any infinitesimal length dx of the string, and add up all such amounts of charge for all of the infinitesimal lengths making up the string segment.



The infinitesimal amount of charge dq on the infinitesimal length dx of the string is just the charge per length λ times the length dx of the infinitesimal string segment.

$$dq = \lambda\, dx$$

Note that you can't take the amount of charge on a finite length (such as 15 cm) of the string to be λ times the length of the segment because λ varies over the length of the segment. In the case of an infinitesimal segment, every part of it is within an infinitesimal distance of the position specified by one and the same value of x. The linear charge density doesn't vary on an infinitesimal segment because x doesn't—the segment is simply too short.



To get the total charge we just have to add up all the dq's. Each dq is specified by its corresponding value of x. To cover all the dq's we have to take into account all the values of x from 0 to 1.00 m. Because each dq is the charge on an infinitesimal length of the line of charge, the sum is going to have an infinite number of terms. An infinite sum of infinitesimal pieces is an integral. When we integrate

$$dq = \lambda\, dx$$

we get, on the left, the sum of all the infinitesimal pieces of charge making up the whole. By definition, the sum of all the infinitesimal amounts of charge is just the total charge Q (which by the way, is what we are solving for); we don't need the tools of integral calculus to deal with the left side of the equation. Integrating both sides of the equation yields:

$$Q = \int_0^{1.00\text{m}} \lambda\, dx$$

Using the given expression $\lambda = 2.56\dfrac{\mu\text{C}}{\text{m}^2}x$ we obtain

$$Q = \int_0^{1.00\text{m}} 2.56\frac{\mu\text{C}}{\text{m}^2}x\,dx = 2.56\frac{\mu\text{C}}{\text{m}^2}\int_0^{1.00\text{m}} x\,dx = 2.56\frac{\mu\text{C}}{\text{m}^2}\frac{x^2}{2}\bigg|_0^{1.00\text{m}} = 2.56\frac{\mu\text{C}}{\text{m}^2}\left[\frac{(1.00\text{m})^2}{2} - \frac{(0)^2}{2}\right] = 1.28\,\mu\text{C}$$

A few more examples of distributions of charge follow:

For instance, consider charge distributed along the x axis, from $x = 0$ to $x = L$ for the case in which the charge density is given by

$$\lambda = \lambda_{MAX}\sin(\pi\ \text{rad}\, x / L)$$

where λ_{MAX} is a constant having units of charge-per-length, rad stands for the units radians, x is the position variable, and L is the length of the charge distribution. Such a charge distribution has a maximum charge density equal to λ_{MAX} occurring in the middle of the line segment.

Another example would be a case in which charge is distributed on a line segment of length L extending along the y axis from $y = a$ to $y = a + L$ with a being a constant and the charge density given by

$$\lambda = \frac{38\ \mu\text{C} \cdot \text{m}}{y^2}$$

In this case the charge on the line is more densely packed in the region closer to the origin. (The smaller y is, the bigger the value of λ, the charge-per-length.)

The simplest case is the one in which the charge is spread out uniformly over the line on which there is charge. In the case of a uniform linear charge distribution, the charge density is the same everywhere on the line of charge. In such a case, the linear charge density λ is simply a constant. Furthermore, in such a simple case, and only in such a simple case, the charge density λ is just the total amount of charge Q divided by the length L of the line along which that charge is uniformly distributed. For instance, suppose you are told that an amount of charge $Q = 2.45$ C is uniformly distributed along a thin rod of length $L = 0.840$ m. Then λ is given by:

$$\lambda = \frac{Q}{L}$$

$$\lambda = \frac{2.45\,\text{C}}{0.840\,\text{m}}$$

$$\lambda = 2.92\frac{\text{C}}{\text{m}}$$

The Electric Field Due to a Continuous Distribution of Charge along a Line

Okay, now we are ready to get down to the nitty-gritty. We are given a continuous distribution of charge along a straight line segment and asked to find the electric field at an empty point in space in the vicinity of the charge distribution. We will consider the case in which both the charge distribution and the empty point in space lie in the x-y plane. The values of the coordinates of the empty point in space are not necessarily specified. We can call them x and y. In solving the problem for a single point in space with unspecified coordinates (x,y), our final answer will have the symbols x and y in it, and our result will actually give the answer for an infinite set of points on the x-y plane.

The plan for solving such a problem is to find the electric field, due to an infinitesimal segment of the charge, at the one empty point in space. We do that for every infinitesimal segment of the charge, and then add up the results to get the total electric field.

Now once we chop up the charge distribution (in our mind, for calculational purposes) into infinitesimal (vanishingly small) pieces, we are going to wind up with an infinite number of

pieces and hence an infinite sum when we go to add up the contributions to the electric field at the one single empty point in space due to all the infinitesimal segments of the linear charge distribution. That is to say, the result is going to be an integral.

An important consideration that we must address is the fact that the electric field, due to each element of charge, at the one empty point in space, is a vector. Hence, what we are talking about is an infinite sum of infinitesimal vectors. In general, the vectors being added are all in different directions from each other. (Can you think of a case so special that the infinite set of infinitesimal electric field vectors are all in the same direction as each other?[2] Note that we are considering the general case, not such a special case.) We know better than to simply add the magnitudes of the vectors, infinite sum or not. Vectors that are not all in the same direction as each other, add like vectors, not like numbers. The thing is, however, the x components of all the infinitesimal electric field vectors at the one empty point in space do add like numbers. Likewise for the y components. Thus, if, for each infinitesimal element of the charge distribution, we find, not just the electric field at the empty point in space, but the x component of that electric field, then we can add up all the x components of the electric field at the empty point in space to get the x component of the electric field, due to the entire charge distribution, at the one empty point in space. The sum is still an infinite sum, but this time it is an infinite sum of scalars rather than vectors, and we have the tools for handling that. Of course, if we are asked for the total electric field, we have to repeat the entire procedure to get the y component of the electric field and then combine the two components of the electric field to get the total[3].

The easy way to do the last step is to use $\hat{\imath}$, $\hat{\jmath}$, $\hat{k}$ notation. That is, once we have E_x and E_y, we can simply write:

$$\vec{E} = E_x\hat{\imath} + E_y\hat{\jmath}$$

[2] Such a special case occurs when one is asked to find the electric field at points on the same line as the charge distribution, at points outside the charge distribution.

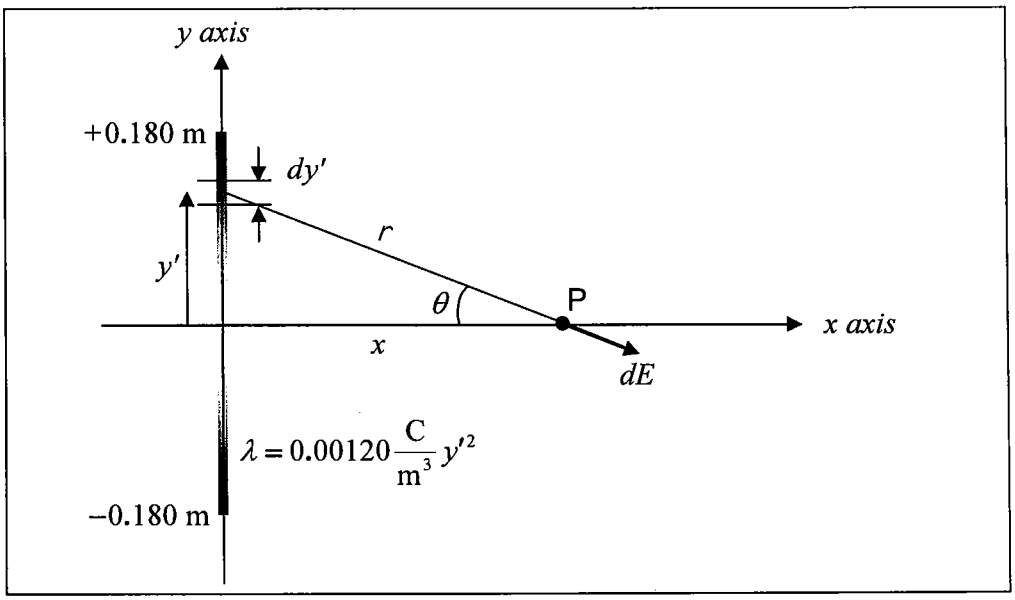
[3] Just to reduce the complexity of the problems you will be dealing with, we limit the discussion to the electric field at a point in the x-y plane due to a charge distribution in the x-y plane. Thus, $E_z = 0$ by inspection and we omit it from the discussion.

Example 30-2

Find the electric field valid for any point on the positive x axis due a 36.0 cm long line of charge, lying on the y axis and centered on the origin, for which the charge density is given by

$$\lambda = 0.00120 \frac{\text{C}}{\text{m}^3} y^2$$

As usual, we'll start our solution with a diagram:



Note that we use (and strongly recommend that you use) primed quantities (x', y') to specify a point on the charge distribution and unprimed quantities (x, y) to specify the empty point in space at which we wish to know the electric field. Thus, in the diagram, the infinitesimal segment of the charge distribution is at $(0, y')$ and point P, the point at which we are finding the electric field, is at $(x, 0)$. Also, our expression for the given linear charge density $\lambda = 0.00120 \frac{\text{C}}{\text{m}^3} y^2$ expressed in terms of y' rather than y is:

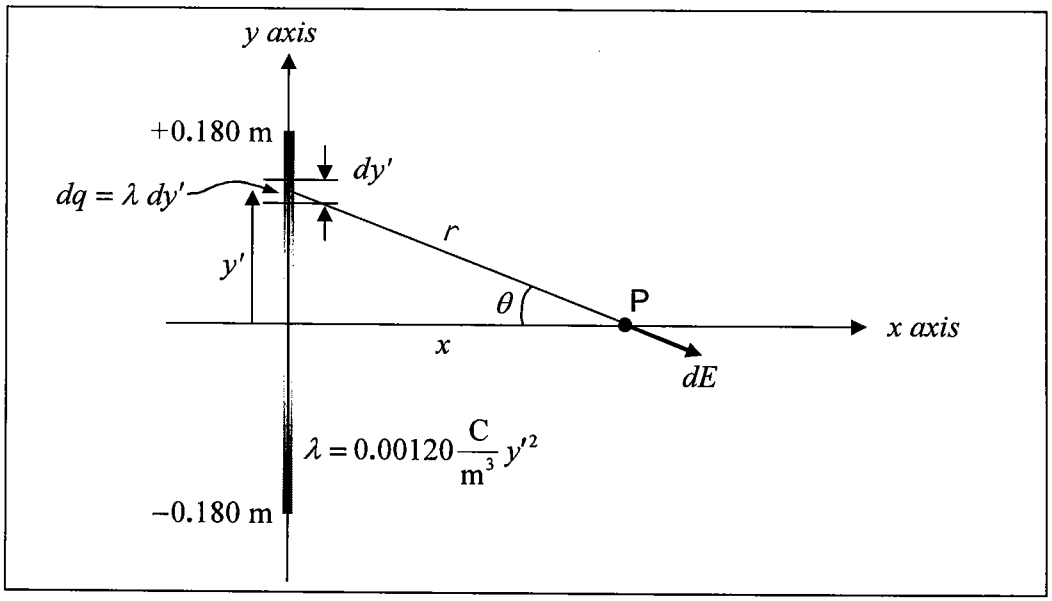
$$\lambda = 0.00120 \frac{\text{C}}{\text{m}^3} y'^2$$

The plan here is to use Coulomb's Law for the Electric Field to get the magnitude of the infinitesimal electric field vector $\mathbf{d\vec{E}}$ at point P due to the infinitesimal amount of charge dq in the infinitesimal segment of length dy'.

$$dE = \frac{k\, dq}{r^2}$$

The amount of charge dq in the infinitesimal segment dy' of the linear charge distribution is given by

$$dq = \lambda\, dy'$$



From the diagram, it clear that we can use the Pythagorean theorem to express the distance r that point P is from the infinitesimal amount of charge dq under consideration as:

$$r = \sqrt{x^2 + y'^2}$$

Substituting this and $dq = \lambda\, dy'$ into our equation for dE ($dE = \frac{k\, dq}{r^2}$) we obtain

$$dE = \frac{k\,\lambda\, dy'}{x^2 + y'^2}$$

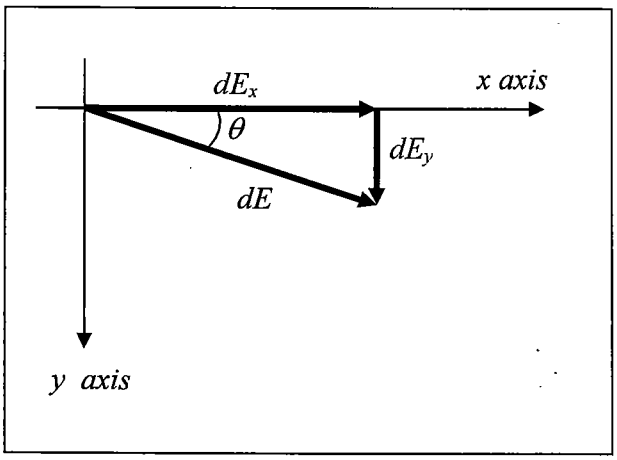
Recall that our plan is to find E_x, then E_y and then put them together using $\vec{\mathbf{E}} = E_x\hat{\mathbf{i}} + E_y\hat{\mathbf{j}}$. So for now, let's get an expression for E_x.

Based on the vector component diagram at right we have

$$dE_x = dE\cos\theta$$

The θ appearing in the diagram at right is the same θ that appears in the diagram above. Based on the plane geometry evident in that diagram (above), we have:

$$\cos\theta = \frac{x}{r} = \frac{x}{\sqrt{x^2 + y'^2}}$$



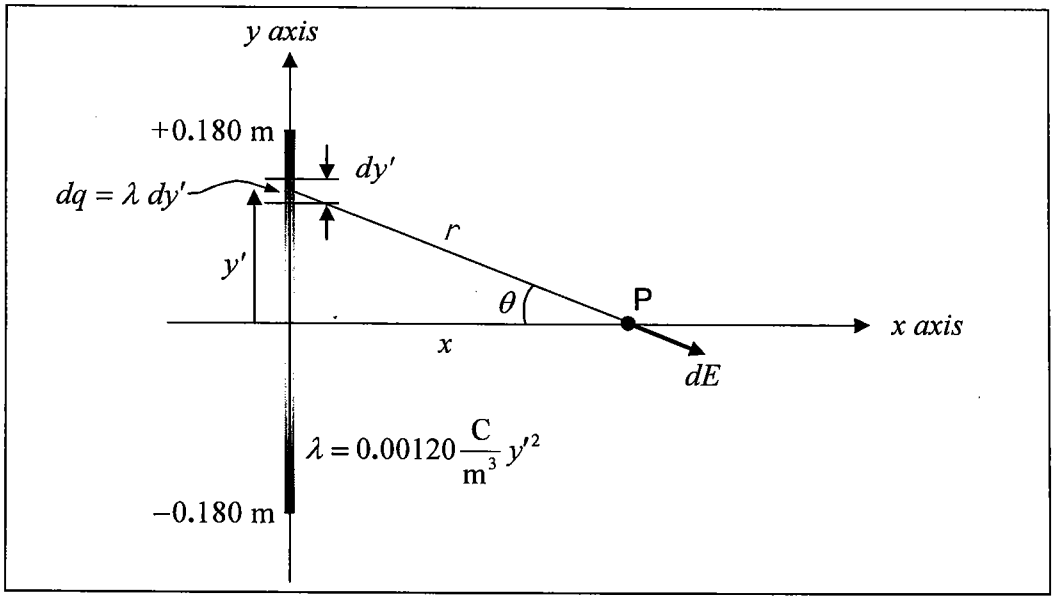
Substituting both this expression for $\cos\theta$ ($\cos\theta = \frac{x}{\sqrt{x^2 + y'^2}}$) and the expression we derived for dE above ($dE = \frac{k\lambda\, dy'}{x^2 + y'^2}$) into our expression $dE_x = dE\cos\theta$ from the vector component diagram yields:

$$dE_x = \frac{k\lambda x\, dy'}{\left(x^2 + y'^2\right)^{3/2}}$$

Also, let's go ahead and replace λ with the given expression $\lambda = 0.00120\ \frac{\text{C}}{\text{m}^3}\, y'^2$:

$$dE_x = \left(0.00120\ \frac{\text{C}}{\text{m}^3}\right)\frac{k\, y'^2\, x\, dy'}{\left(x^2 + y'^2\right)^{3/2}}$$

Now we have an expression for dE_x that includes only one quantity, namely y', that depends on which bit of the charge distribution is under consideration. Furthermore, although in the diagram



it appears that we picked out a particular infinitesimal line segment dy', in fact, the value of y' needed to establish its position is not specified. That is, we have an equation for dE_x that is good for any infinitesimal segment dy' of the given linear charge distribution. To identify a particular dy' we just have to specify the value of y'. Thus to sum up all the dE_x's we just have to add, to a running total, the dE_x for each of the possible values of y'. Thus we need to integrate the expression for dE_x for all the values of y' from $-0.180\,\text{m}$ to $+0.180\,\text{m}$.

$$\int dE_x = \int_{-0.180\,\text{m}}^{+0.180\,\text{m}} \left(0.00120\ \frac{\text{C}}{\text{m}^3}\right)\frac{k\, y'^2\, x\, dy'}{\left(x^2 + y'^2\right)^{3/2}}$$

Copying that equation here:

$$\int dE_x = \int_{-0.180\,\mathrm{m}}^{+0.180\,\mathrm{m}} \left(0.00120\,\frac{\mathrm{C}}{\mathrm{m}^3}\right) \frac{k\,y'^2\,x\,dy'}{\left(x^2+y'^2\right)^{3/2}}$$

we note that on the left is the infinite sum of all the contributions to the x component of the electric field due to all the infinitesimal elements of the line of charge. We don't need any special mathematics techniques to evaluate that. The sum of all the parts is the whole. That is, on the left, we have E_x.

The right side, we can evaluate. First, let's factor out the constants:

$$E_x = \left(0.00120\,\frac{\mathrm{C}}{\mathrm{m}^3}\right) k\,x \int_{-0.180\,\mathrm{m}}^{+0.180\,\mathrm{m}} \frac{y'^2\,dy'}{\left(x^2+y'^2\right)^{3/2}}$$

The integral is given on your formula sheet. Carrying out the integration yields:

$$E_x = \left(0.00120\,\frac{\mathrm{C}}{\mathrm{m}^3}\right) k\,x \left[\frac{y'}{\sqrt{x^2+y'^2}} + \ln\left(y'+\sqrt{x^2+y'^2}\right)\right]_{-0.180\,\mathrm{m}}^{+0.180\,\mathrm{m}}$$

$$E_x = \left(.00120\,\frac{\mathrm{C}}{\mathrm{m}^3}\right) k\,x \cdot \left\{ \left[\frac{+.180\,\mathrm{m}}{\sqrt{x^2+(+.180\,\mathrm{m})^2}} + \ln\left(+.180\,\mathrm{m}+\sqrt{x^2+(+.180\,\mathrm{m})^2}\right)\right] - \left[\frac{-.180\,\mathrm{m}}{\sqrt{x^2+(-.180\,\mathrm{m})^2}} + \ln\left(-.180\,\mathrm{m}+\sqrt{x^2+(-.180\,\mathrm{m})^2}\right)\right] \right\}$$

$$E_x = \left(.00120\,\frac{\mathrm{C}}{\mathrm{m}^3}\right) k\,x \cdot \left[\frac{.360\,\mathrm{m}}{\sqrt{x^2+(.180\,\mathrm{m})^2}} + \ln\frac{\sqrt{x^2+(.180\,\mathrm{m})^2}+.180\,\mathrm{m}}{\sqrt{x^2+(.180\,\mathrm{m})^2}-.180\,\mathrm{m}}\right]$$

Substituting the value of the Coulomb constant k from the formula sheet we obtain

$$E_x = \left(.00120\,\frac{\mathrm{C}}{\mathrm{m}^3}\right) 8.99\times10^9\,\frac{\mathrm{N}\cdot\mathrm{m}^2}{\mathrm{C}^2}\,x \cdot \left[\frac{.360\,\mathrm{m}}{\sqrt{x^2+(.180\,\mathrm{m})^2}} + \ln\frac{\sqrt{x^2+(.180\,\mathrm{m})^2}+.180\,\mathrm{m}}{\sqrt{x^2+(.180\,\mathrm{m})^2}-.180\,\mathrm{m}}\right]$$

Finally we have

$$E_x = 1.08\times10^7\,\frac{\mathrm{N}}{\mathrm{C}\cdot\mathrm{m}}\,x \cdot \left[\frac{.360\,\mathrm{m}}{\sqrt{x^2+(.180\,\mathrm{m})^2}} + \ln\frac{\sqrt{x^2+(.180\,\mathrm{m})^2}+.180\,\mathrm{m}}{\sqrt{x^2+(.180\,\mathrm{m})^2}-.180\,\mathrm{m}}\right]$$

It is interesting to note that while the position variable x (which specifies the location of the empty point in space at which the electric field is being calculated) is a constant for purposes of integration (the location of point P does not change as we include the contribution to the electric field at point P of each of the infinitesimal segments making up the charge distribution), an actual value x was never specified. Thus our final result

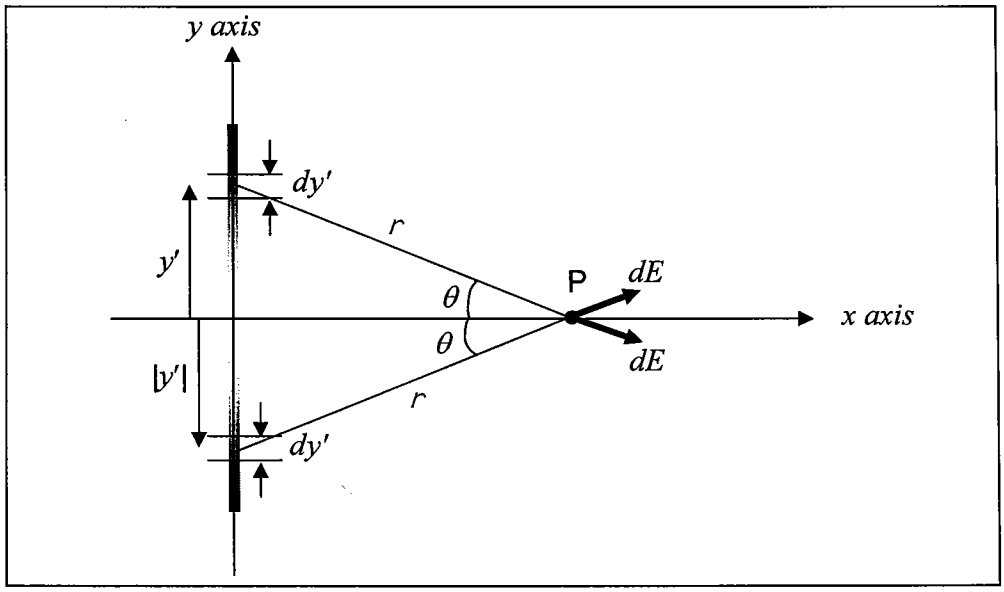
$$E_x = 1.08\times10^7 \frac{\text{N}}{\text{C}\cdot\text{m}}\, x\cdot\left[\frac{.360\,\text{m}}{\sqrt{x^2+(.180\,\text{m})^2}} + \ln\frac{\sqrt{x^2+(.180\,\text{m})^2}+.180\,\text{m}}{\sqrt{x^2+(.180\,\text{m})^2}-.180\,\text{m}}\right]$$

for E_x is a function of the position variable x.

Getting the y-component of the electric field can be done with a lot less work than it took to get E_x if we take advantage of the symmetry of the charge distribution with respect to the x axis. Recall that the charge density λ, for the case at hand, is given by:

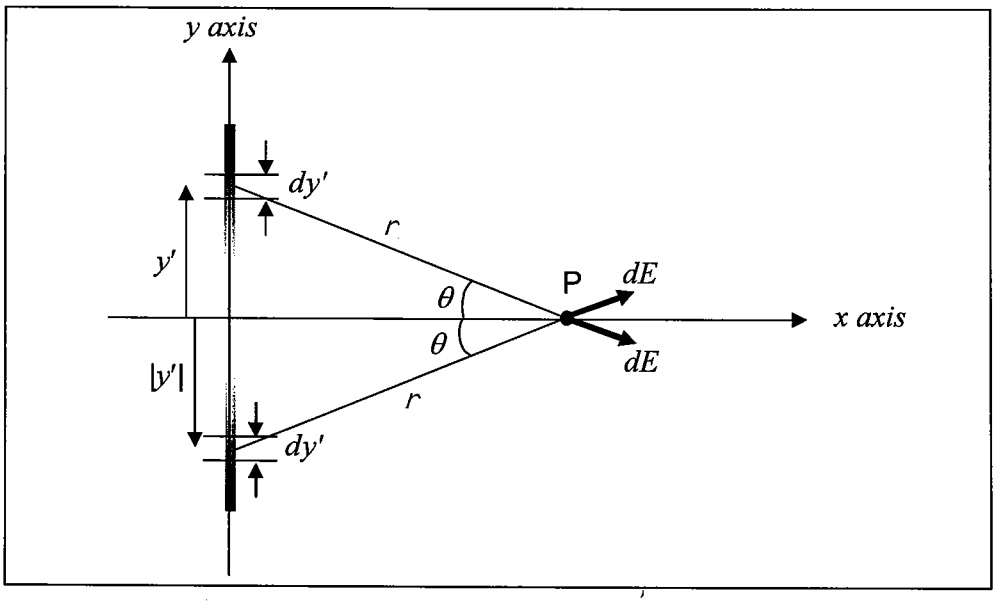
$$\lambda = 0.00120\frac{\text{C}}{\text{m}^3}\, y'^2$$

Because λ is proportional to y'^2, the value of λ is the same at the negative of a specified y' value as it is at the y' value itself. More specifically, the amount of charge in each of the two same-size infinitesimal elements dy' of the charge distribution depicted in the following diagram:



is one and the same value because one element is the same distance below the x axis as the other is above it. This position circumstance also makes the distance r that each element is from point P the same as that of the other, and, it makes the two angles (each of which is labeled θ in the diagram) have one and the same value. Thus the two $\vec{\mathbf{dE}}$ vectors have one and the same

magnitude. As a result of the latter two facts (same angle, same magnitude of $\vec{\mathbf{dE}}$), the y components of the two $\vec{\mathbf{dE}}$ vectors cancel each other out. As can be seen in the diagram under consideration:



one is in the $+y$ direction and the other in the $-y$ direction. The y components are "equal and opposite.") In fact, for each and every charge distribution element dy' that is above the x axis and is thus creating a downward contribution to the y component of the electric field at point P, there is an element dy' that is the same distance *below* the x axis that is creating an *upward* contribution to the y component of the electric field at point P, canceling the y component of the former. Thus the net sum of all the electric field y components (since they cancel pair-wise) is zero. That is to say that due to the symmetry of the charge distribution with respect to the x axis, $E_y = 0$. Thus,

$$\vec{\mathbf{E}} = E_x \hat{\mathbf{\imath}}.$$

Using the expression for E_x that we found above, we have, for our final answer:

$$\vec{\mathbf{E}} = 1.08\times10^7 \frac{\mathrm{N}}{\mathrm{C\cdot m}}\, x \left[\frac{.360\,\mathrm{m}}{\sqrt{x^2 + (.180\,\mathrm{m})^2}} + \ln \frac{\sqrt{x^2 + (.180\,\mathrm{m})^2} + .180\,\mathrm{m}}{\sqrt{x^2 + (.180\,\mathrm{m})^2} - .180\,\mathrm{m}} \right] \hat{\mathbf{\imath}}$$

31 The Electric Potential due to a Continuous Charge Distribution

We have defined *electric potential* as electric potential-energy-per-charge. Potential energy was defined as the capacity, of an object to do work, possessed by the object because of its position in space. Potential energy is one way of characterizing the effect, or the potential effect, of a force. In the case of electric potential energy, the force in question is the electrostatic force (a.k.a. the Coulomb force)—you know: the repulsive force that two like charges exert on each other, and, the attractive force that two unlike charges exert on each other. The electric potential energy of a charged particle depends on a characteristic of itself, and a characteristic of the point in space at which it finds itself. The characteristic of itself is its charge, and, the characteristic of the point in space is what this chapter is about, the electric potential-energy-per-charge, better known as the electric potential. If we can establish the electric potential-energy-per-charge for each point in space in the vicinity of some source charge, it is easy to determine what the potential energy of a victim charge would be at any such point in space. To do so, we just have to multiply the charge of the victim by the electric potential-energy-per-charge (the electric potential) applicable to the point in space at which the victim is located.

In the next chapter, we exploit the fact that if you know the electric potential throughout a region in space, you can use that knowledge to determine the electric field in that region of space.

Our purpose of *this* chapter, is to help you develop your ability to determine the electric potential, as a function of position, in the vicinity of a charge distribution—in particular, in the vicinity of a continuous charge distribution. (Recall that you can think of a continuous charge distribution as some charge that is smeared out over space, whereas a discrete charge distribution is a set of charged particles, with some space between nearest neighbors.)

It's important for you to be able to contrast the electric potential with the electric field. *The electric potential is a scalar* whereas the *electric field is a vector.* The electric potential is potential *energy*-per-charge of the would be victim whereas the electric field is a *force-per-charge* of the would be victim. Hey, that makes this chapter easy compared to the one in which we worked on calculating the electric field due to a continuous charge distribution. It is, in general, easier to calculate a scalar than it is to calculate a vector.

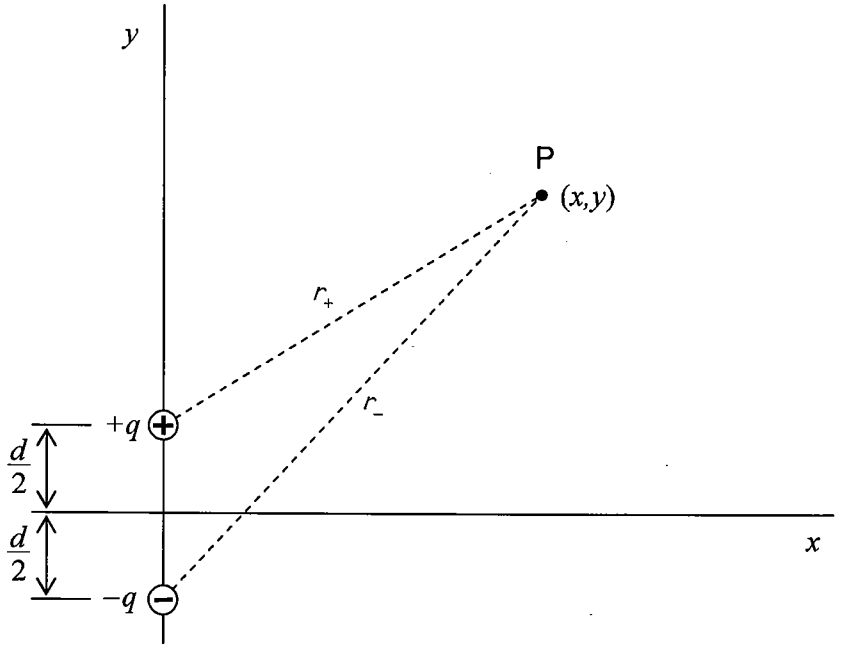
Let's kick things off by doing a review problem involving a discrete distribution of charge. Please solve the following example problem and then check your work against my solution which follows the problem statement.

Example 31-1

Find the electric potential on the x-y plane, due to a pair of charges, one of charge $+q$ at $(0, d/2)$ and the other of charge $-q$ at $(0, -d/2)$.

Solution

We define a point P to be at some unspecified position (x,y).



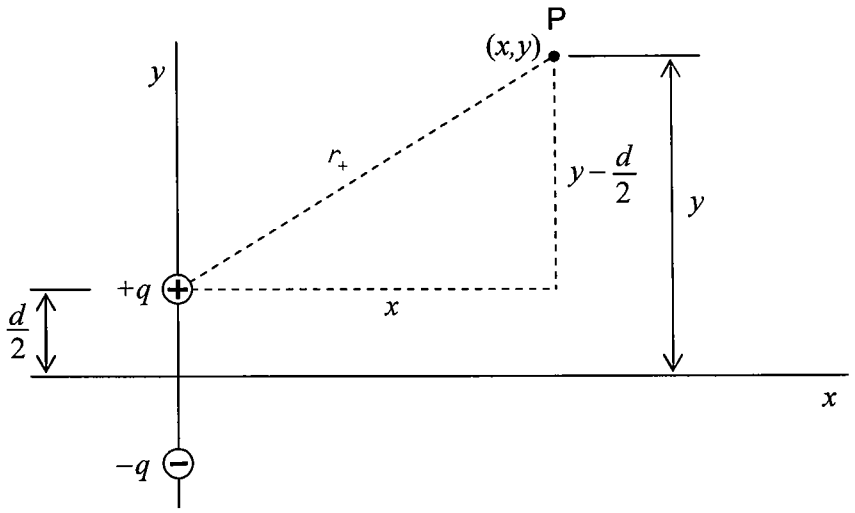
We call the distance from the positive charge to point P, r_+, and, we call the distance from the negative charge to point P, r_-. The electric potential due to a single point charge is given by $\varphi = \frac{kq}{r}$. Also, the contributions to the electric potential at one point in space due to more than one point charge simply add like numbers. So, we have:

$$\varphi = \varphi_1 + \varphi_2$$

$$\varphi = \frac{kq}{r_+} + \frac{k(-q)}{r_-}$$

$$\varphi = \frac{kq}{r_+} - \frac{kq}{r_-}$$

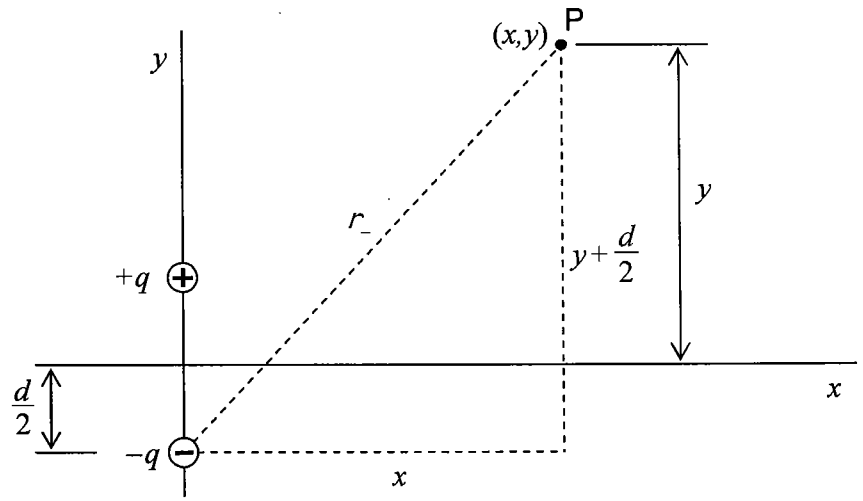
But, from the diagram:



we can determine that:

$$r_+ = \sqrt{x^2 + \left(y - \frac{d}{2}\right)^2}$$

and from the diagram:



we can see that:

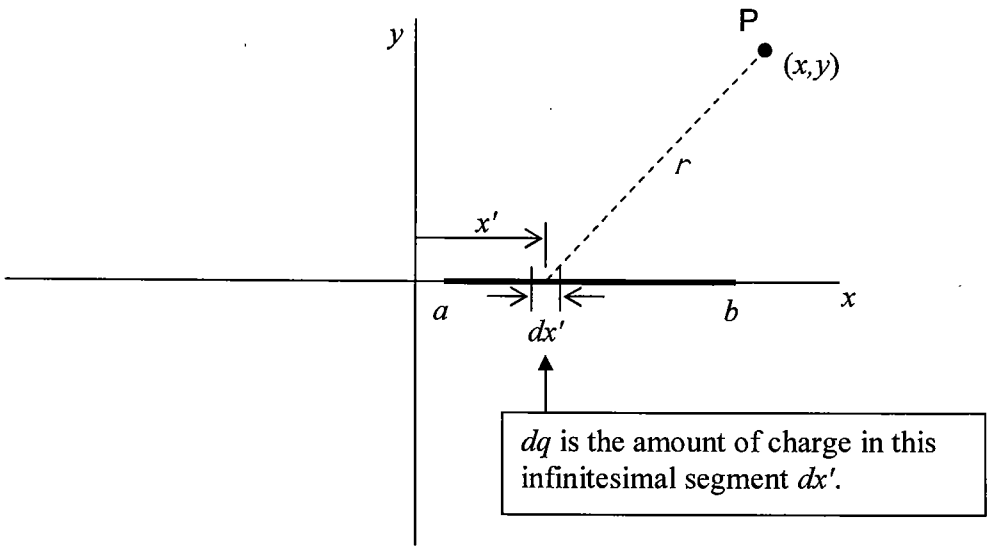
$$r_- = \sqrt{x^2 + \left(y + \frac{d}{2}\right)^2}$$

Plugging both of these results into our expression $\varphi = \frac{kq}{r_+} - \frac{kq}{r_-}$ yields:

$$\varphi = \frac{kq}{\sqrt{x^2 + \left(y - \frac{d}{2}\right)^2}} - \frac{kq}{\sqrt{x^2 + \left(y + \frac{d}{2}\right)^2}}$$

That's enough review. Please keep that $\varphi = \frac{kq}{r}$ formula in mind as we move on to the new stuff. Also keep in mind the fact that the various contributions to the electric *potential* at an empty point in space simply add (like numbers/scalars rather than like vectors).

The "new stuff" is the electric potential due to a continuous distribution of charge along a line segment. What we are dealing with is some line segment of charge. It can be anywhere, in any orientation, but for concreteness, let's consider a line segment of charge on the x axis, say from some $x = a$ to $x = b$ where $a < b$. Furthermore, let's assume the linear charge density (the charge-per-length) on the line segment to be some function $\lambda(x')$. The idea is to treat the charge distribution as an infinite set of point charges where each point charge may have a different charge value dq depending on where (at what value of x') it is along the line segment.



A particular infinitesimal segment of the line of charge, a length dx' of the line segment, will make a contribution

$$d\varphi = \frac{k\,dq}{r}$$

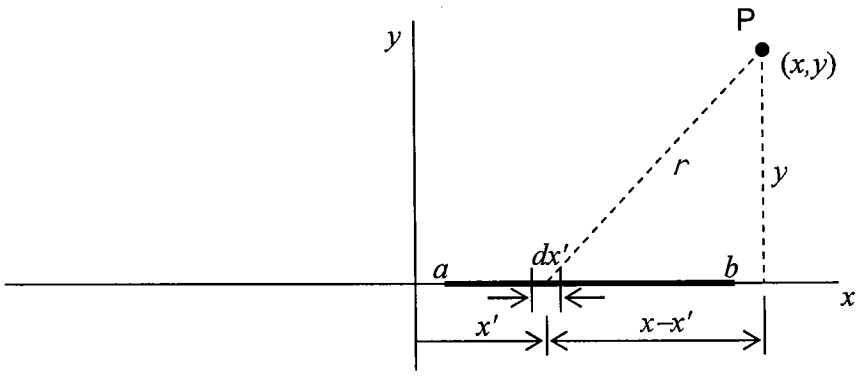
to the electric field at point P.

The amount of charge, dq, in the infinitesimal segment dx' of the line of charge is just the charge-per-length $\lambda(x')$ (the linear charge density) times the length dx' of the segment. That is to say that $dq = \lambda(x')dx'$. Substituting this into $d\varphi = \frac{k\,dq}{r}$ yields:

$$d\varphi = \frac{k\overbrace{\lambda(x')}\,dx'}{r}$$

This is "λ of x'."

Applying the Pythagorean theorem to the triangle in the diagram:



tells us that r can be written as $r = \sqrt{(x-x')^2 + y^2}$. Substituting this into our expression for dV yields:

$$d\varphi = \frac{k\lambda(x')dx'}{\sqrt{(x-x')^2 + y^2}}$$

Integrating both sides yields:

$$\int d\varphi = \int_a^b \frac{k\lambda(x')\,dx'}{\sqrt{(x-x')^2 + y^2}}$$

$$\varphi = k\int_a^b \frac{\overbrace{\lambda(x')}\,dx'}{\sqrt{(x-x')^2 + y^2}}$$

This is "λ of x'."
Do not interpret it as λ times x'.

This is the electric potential at point P due to the charged line segment on the x axis. Each bit of charge on the line segment is specified by its position variable x'. Thus, in summing the contributions to the electric potential due to each bit of charge, x' is our variable of integration. While its position coordinates have not been specified, but rather, they have been designated x and y, point P is a fixed point in space. Hence, in summing up all the contributions to the electric potential at point P; x and y are to be considered constants. After the integral is done, however, because we never specified values for x and y, the resulting expression for φ can be considered to be a function of x and y.

32 Calculating the Electric Field from the Electric Potential

The plan here is to develop a relation between the electric field and the corresponding electric potential that allows you to calculate the electric field from the electric potential.

The electric field is the force-per-charge associated with empty points in space that have a force-per-charge because they are in the vicinity of a source charge or some source charges. The electric potential is the potential energy-per-charge associated with the same empty points in space. Since the electric field is the force-per-charge, and the electric potential is the potential energy-per-charge, the relation between the electric field and its potential is essentially a special case of the relation between any force and its associated potential energy. So, I'm going to start by developing the more general relation between a force and its potential energy, and then move on to the special case in which the force is the electric field times the charge of the victim and the potential energy is the electric potential times the charge of the victim.

The idea behind potential energy was that it represented an easy way of getting the work done by a force on a particle that moves from point A to point B under the influence of the force. By definition, the work done is the force along the path times the length of the path. If the force along the path varies along the path, then we take the force along the path at a particular point on the path, times the length of an infinitesimal segment of the path at that point, and repeat, for every infinitesimal segment of the path, adding the results as we go along. The final sum is the work. The potential energy idea represents the assignment of a value of potential energy to every point in space so that, rather than do the path integral just discussed, we simply subtract the value of the potential energy at point A from the value of the potential energy at point B. This gives us the change in the potential energy experienced by the particle in moving from point A to point B. Then, the work done is the negative of the change in potential energy. For this to be the case, the assignment of values of potential energy values to points in space must be done just right. For things to work out on a macroscopic level, we must ensure that they are correct at an infinitesimal level. We can do this by setting:

Work as Change in Potential Energy = Work as Force-Along-Path times Path Length

$$-dU = \vec{\mathbf{F}} \cdot \vec{\mathbf{ds}}$$

where:

- dU is an infinitesimal change in potential energy,
- $\vec{\mathbf{F}}$ is a force, and
- $\vec{\mathbf{ds}}$ is the infinitesimal displacement-along-the-path vector.

In Cartesian unit vector notation, $\vec{\mathbf{ds}}$ can be expressed as $\vec{\mathbf{ds}} = dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}}$, and, $\vec{\mathbf{F}}$ can be expressed as $\vec{\mathbf{F}} = F_x\hat{\mathbf{i}} + F_y\hat{\mathbf{j}} + F_z\hat{\mathbf{k}}$. Substituting these two expressions into our expression $-dU = \vec{\mathbf{F}} \cdot \vec{\mathbf{ds}}$, we obtain:

$$-dU = (F_x\hat{\mathbf{i}} + F_y\hat{\mathbf{j}} + F_z\hat{\mathbf{k}}) \cdot (dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}})$$

$$-dU = F_x\,dx + F_y\,dy + F_z\,dz$$

Now check this out. If we hold y and z constant (in other words, if we consider dy and dz to be zero) then:

$$-dU = F_x\,dx \text{ (when } y \text{ and } z \text{ are held constant)}$$

Dividing both sides by dx and switching sides yields:

$$F_x = -\frac{dU}{dx} \text{ (when } y \text{ and } z \text{ are held constant)}$$

That is, if you have the potential energy as a function of x, y, and z; and; you take the negative of the derivative with respect to x while holding y and z constant, you get the x component of the force that is characterized by the potential energy function. Taking the derivative of U with respect to x while holding the other variables constant is called taking the partial derivative of U with respect to x and written

$$\frac{\partial U}{\partial x}$$

Alternatively, one writes

$$\left.\frac{\partial U}{\partial x}\right|_{y,z}$$

to be read, "the partial derivative of U with respect to x holding y and z constant." This latter expression makes it more obvious to the reader just what is being held constant. Rewriting our expression for F_x with the partial derivative notation, we have:

$$F_x = -\frac{\partial U}{\partial x}$$

Returning to our expression $-dU = F_x\,dx + F_y\,dy + F_z\,dz$, if we hold x and z constant we get:

$$F_y = -\frac{\partial U}{\partial y}$$

and, if we hold x and y constant we get,

$$F_z = -\frac{\partial U}{\partial z}$$

Substituting these last three results into the force vector expressed in unit vector notation:

$$\vec{\mathbf{F}} = F_x \hat{\mathbf{i}} + F_y \hat{\mathbf{j}} + F_z \hat{\mathbf{k}}$$

yields

$$\vec{\mathbf{F}} = -\frac{\partial U}{\partial x}\hat{\mathbf{i}} - \frac{\partial U}{\partial y}\hat{\mathbf{j}} - \frac{\partial U}{\partial z}\hat{\mathbf{k}}$$

which can be written:

$$\vec{\mathbf{F}} = -\left(\frac{\partial U}{\partial x}\hat{\mathbf{i}} + \frac{\partial U}{\partial y}\hat{\mathbf{j}} + \frac{\partial U}{\partial z}\hat{\mathbf{k}}\right)$$

Okay, now, this business of:

taking the partial derivative of U with respect to x and multiplying the result by the unit vector $\hat{\mathbf{i}}$ and then,
taking the partial derivative of U with respect to y and multiplying the result by the unit vector $\hat{\mathbf{j}}$ and then,
taking the partial derivative of U with respect to z and multiplying the result by the unit vector $\hat{\mathbf{k}}$, and then,
adding all three partial-derivative-times-unit-vector quantities up,

is called "taking the gradient of U" and is written ∇U. "Taking the gradient" is something that you do to a *scalar* function, but, the result is a *vector*. In terms of our gradient notation, we can write our expression for the force as,

$$\vec{\mathbf{F}} = -\nabla U \qquad (32\text{-}1)$$

Check this out for the gravitational potential near the surface of the earth. Define a Cartesian coordinate system with, for instance, the origin at sea level, and, with the x-y plane being horizontal and the +z direction being upward. Then, the potential energy of a particle of mass m is given as:

$$U = mgz$$

Now, suppose you knew this to be the potential but you didn't know the force. You can calculate the force using $\vec{\mathbf{F}} = -\nabla U$, which, as you know, can be written:

$$\vec{\mathbf{F}} = -\left(\frac{\partial U}{\partial x}\hat{\mathbf{i}} + \frac{\partial U}{\partial y}\hat{\mathbf{j}} + \frac{\partial U}{\partial z}\hat{\mathbf{k}}\right)$$

Substituting $U = mgz$ in for U we have

$$\vec{\mathbf{F}} = -\left(\frac{\partial}{\partial x}(mgz)\hat{\mathbf{i}} + \frac{\partial}{\partial y}(mgz)\hat{\mathbf{j}} + \frac{\partial}{\partial z}(mgz)\hat{\mathbf{k}}\right)$$

Now remember, when we take the partial derivative with respect to x we are supposed to hold y and z constant. (There is no y.) But, if we hold z constant, then the whole thing (mgz) is constant. And, the derivative of a constant, with respect to x, is 0. In other words, $\frac{\partial}{\partial x}(mgz) = 0$

Likewise, $\frac{\partial}{\partial y}(mgz) = 0$. In fact, the only non-zero partial derivative in our expression for the force is $\frac{\partial}{\partial z}(mgz) = mg$. So:

$$\vec{\mathbf{F}} = -\left(0\hat{\mathbf{i}} + 0\hat{\mathbf{j}} + mg\hat{\mathbf{k}}\right)$$

In other words:

$$\vec{\mathbf{F}} = -mg\hat{\mathbf{k}}$$

That is to say that, based on the gravitational potential $U = mgz$, the gravitational force is in the $-\hat{\mathbf{k}}$ direction (downward), and, is of magnitude mg. Of course, you knew this in advance, the gravitational force in question is just the weight force. The example was just meant to familiarize you with the gradient operator and the relation between force and potential energy.

Okay, as important as it is that you realize that we are talking about a general relationship between force and potential energy, it is now time to narrow the discussion to the case of the electric force and the electric potential energy, and, from there, to derive a relation between the electric field and electric potential (which is electric potential-energy-per-charge).

Starting with $\vec{\mathbf{F}} = -\nabla U$ written out the long way:

$$\vec{\mathbf{F}} = -\left(\frac{\partial U}{\partial x}\hat{\mathbf{i}} + \frac{\partial U}{\partial y}\hat{\mathbf{j}} + \frac{\partial U}{\partial z}\hat{\mathbf{k}}\right)$$

we apply it to the case of a particle with charge q in an electric field $\vec{\mathbf{E}}$ (caused to exist in the region of space in question by some unspecified source charge or distribution of source charge). The electric field exerts a force $\vec{\mathbf{F}} = q\vec{\mathbf{E}}$ on the particle, and, the particle has electric potential energy $U = q\varphi$ where φ is the electric potential at the point in space at which the charged particle is located. Plugging these into $\vec{\mathbf{F}} = -\left(\frac{\partial U}{\partial x}\hat{\mathbf{i}} + \frac{\partial U}{\partial y}\hat{\mathbf{j}} + \frac{\partial U}{\partial z}\hat{\mathbf{k}}\right)$ yields:

$$q\vec{\mathbf{E}} = -\left(\frac{\partial(q\varphi)}{\partial x}\hat{\mathbf{i}} + \frac{\partial(q\varphi)}{\partial y}\hat{\mathbf{j}} + \frac{\partial(q\varphi)}{\partial z}\hat{\mathbf{k}}\right)$$

which I copy here for your convenience:

$$q\vec{\mathbf{E}} = -\left(\frac{\partial(q\varphi)}{\partial x}\hat{\mathbf{i}} + \frac{\partial(q\varphi)}{\partial y}\hat{\mathbf{j}} + \frac{\partial(q\varphi)}{\partial z}\hat{\mathbf{k}}\right)$$

The q inside each of the partial derivatives is a constant so we can factor it out of each partial derivative.

$$q\vec{\mathbf{E}} = -\left(q\frac{\partial\varphi}{\partial x}\hat{\mathbf{i}} + q\frac{\partial\varphi}{\partial y}\hat{\mathbf{j}} + q\frac{\partial\varphi}{\partial z}\hat{\mathbf{k}}\right)$$

Then, since q appears in every term, we can factor it out of the sum:

$$q\vec{\mathbf{E}} = -q\left(\frac{\partial\varphi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\varphi}{\partial y}\hat{\mathbf{j}} + \frac{\partial\varphi}{\partial z}\hat{\mathbf{k}}\right)$$

Dividing both sides by the charge of the victim yields the desired relation between the electric field and the electric potential:

$$\vec{\mathbf{E}} = -\left(\frac{\partial\varphi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\varphi}{\partial y}\hat{\mathbf{j}} + \frac{\partial\varphi}{\partial z}\hat{\mathbf{k}}\right)$$

We see that the electric field $\vec{\mathbf{E}}$ is just the gradient of the electric potential φ. This result can be expressed more concisely by means of the gradient operator as:

$$\vec{\mathbf{E}} = -\nabla\varphi \tag{32-2}$$

Example 32-1

In Example **31-1**, we found that the electric potential due to a pair of particles, one of charge $+q$ at $(0, d/2)$ and the other of charge $-q$ at $(0, -d/2)$, is given by:

$$\varphi = \frac{kq}{\sqrt{x^2 + \left(y - \frac{d}{2}\right)^2}} - \frac{kq}{\sqrt{x^2 + \left(y + \frac{d}{2}\right)^2}}$$

Such a pair of charges is called an electric dipole. Find the electric field of the dipole, valid for any point on the x axis.

Solution: We can use a symmetry argument and our conceptual understanding of the electric field due to a point charge to deduce that the x component of the electric field has to be zero, and, the y component has to be negative. But, let's use the gradient method to do that, and, to get an expression for the y component of the electric field. I do argue, however that, from our conceptual understanding of the electric field due to a point charge, neither particle's electric field has a z component in the x-y plane, so we are justified in neglecting the z component altogether. As such our gradient operator expression for the electric field

$$\vec{\mathbf{E}} = -\nabla\varphi$$

becomes

$$\vec{\mathbf{E}} = -\left(\frac{\partial\varphi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\varphi}{\partial y}\hat{\mathbf{j}}\right)$$

Let's work on the $\frac{\partial\varphi}{\partial x}$ part:

$$\frac{\partial\varphi}{\partial x} = \frac{\partial}{\partial x}\left(\frac{kq}{\sqrt{x^2 + \left(y - \frac{d}{2}\right)^2}} - \frac{kq}{\sqrt{x^2 + \left(y + \frac{d}{2}\right)^2}}\right)$$

$$\frac{\partial\varphi}{\partial x} = kq\frac{\partial}{\partial x}\left(\left[x^2 + \left(y - \frac{d}{2}\right)^2\right]^{-\frac{1}{2}} - \left[x^2 + \left(y + \frac{d}{2}\right)^2\right]^{-\frac{1}{2}}\right)$$

$$\frac{\partial\varphi}{\partial x} = kq\left(-\frac{1}{2}\left[x^2 + \left(y - \frac{d}{2}\right)^2\right]^{-\frac{3}{2}}2x - -\frac{1}{2}\left[x^2 + \left(y + \frac{d}{2}\right)^2\right]^{-\frac{3}{2}}2x\right)$$

$$\frac{\partial\varphi}{\partial x} = kqx\left(\left[x^2+\left(y+\frac{d}{2}\right)^2\right]^{-\frac{3}{2}} - \left[x^2+\left(y-\frac{d}{2}\right)^2\right]^{-\frac{3}{2}}\right)$$

$$\frac{\partial\varphi}{\partial x} = \frac{kqx}{\left[x^2+\left(y+\frac{d}{2}\right)^2\right]^{\frac{3}{2}}} - \frac{kqx}{\left[x^2+\left(y-\frac{d}{2}\right)^2\right]^{\frac{3}{2}}}$$

We were asked to find the electric field on the x axis, so, we evaluate this expression at $y = 0$:

$$\frac{\partial\varphi}{\partial x} = \frac{kqx}{\left[x^2+\left(0+\frac{d}{2}\right)^2\right]^{\frac{3}{2}}} - \frac{kqx}{\left[x^2+\left(0-\frac{d}{2}\right)^2\right]^{\frac{3}{2}}}$$

$$\left.\frac{\partial\varphi}{\partial x}\right|_{y=0} = 0$$

To continue with our determination of $\vec{\mathbf{E}} = -\left(\frac{\partial\varphi}{\partial x}\hat{\imath} + \frac{\partial\varphi}{\partial y}\hat{\jmath}\right)$, we next solve for $\frac{\partial\varphi}{\partial y}$.

$$\frac{\partial\varphi}{\partial y} = \frac{\partial}{\partial y}\left(\frac{kq}{\sqrt{x^2+\left(y-\frac{d}{2}\right)^2}} - \frac{kq}{\sqrt{x^2+\left(y+\frac{d}{2}\right)^2}}\right)$$

$$\frac{\partial\varphi}{\partial y} = kq\frac{\partial}{\partial y}\left(\left[x^2+\left(y-\frac{d}{2}\right)^2\right]^{-\frac{1}{2}} - \left[x^2+\left(y+\frac{d}{2}\right)^2\right]^{-\frac{1}{2}}\right)$$

$$\frac{\partial\varphi}{\partial y} = kq\left(-\frac{1}{2}\left[x^2+\left(y-\frac{d}{2}\right)^2\right]^{-\frac{3}{2}}2\left(y-\frac{d}{2}\right) - -\frac{1}{2}\left[x^2+\left(y+\frac{d}{2}\right)^2\right]^{-\frac{3}{2}}2\left(y+\frac{d}{2}\right)\right)$$

$$\frac{\partial\varphi}{\partial y} = kq\left(\left[x^2+\left(y+\frac{d}{2}\right)^2\right]^{-\frac{3}{2}}\left(y+\frac{d}{2}\right) - \left[x^2+\left(y-\frac{d}{2}\right)^2\right]^{-\frac{3}{2}}\left(y-\frac{d}{2}\right)\right)$$

$$\frac{\partial\varphi}{\partial y}=\frac{kq\left(y+\frac{d}{2}\right)}{\left[x^2+\left(y+\frac{d}{2}\right)^2\right]^{\frac{3}{2}}}-\frac{kq\left(y-\frac{d}{2}\right)}{\left[x^2+\left(y-\frac{d}{2}\right)^2\right]^{\frac{3}{2}}}$$

Again, we were asked to find the electric field on the x axis, so, we evaluate this expression at $y=0$:

$$\left.\frac{\partial\varphi}{\partial y}\right|_{y=0}=\frac{kq\left(0+\frac{d}{2}\right)}{\left[x^2+\left(0+\frac{d}{2}\right)^2\right]^{\frac{3}{2}}}-\frac{kq\left(0-\frac{d}{2}\right)}{\left[x^2+\left(0-\frac{d}{2}\right)^2\right]^{\frac{3}{2}}}$$

$$\left.\frac{\partial\varphi}{\partial y}\right|_{y=0}=\frac{kqd}{\left[x^2+\frac{d^2}{4}\right]^{\frac{3}{2}}}$$

Plugging $\left.\frac{\partial\varphi}{\partial x}\right|_{y=0}=0$ and $\left.\frac{\partial\varphi}{\partial y}\right|_{y=0}=\frac{kqd}{\left[x^2+\frac{d^2}{4}\right]^{\frac{3}{2}}}$ into $\vec{\mathbf{E}}=-\left(\frac{\partial\varphi}{\partial x}\hat{\imath}+\frac{\partial\varphi}{\partial y}\hat{\jmath}\right)$ yields:

$$\vec{\mathbf{E}}=-\left(0\hat{\imath}+\frac{kqd}{\left[x^2+\frac{d^2}{4}\right]^{\frac{3}{2}}}\hat{\jmath}\right)$$

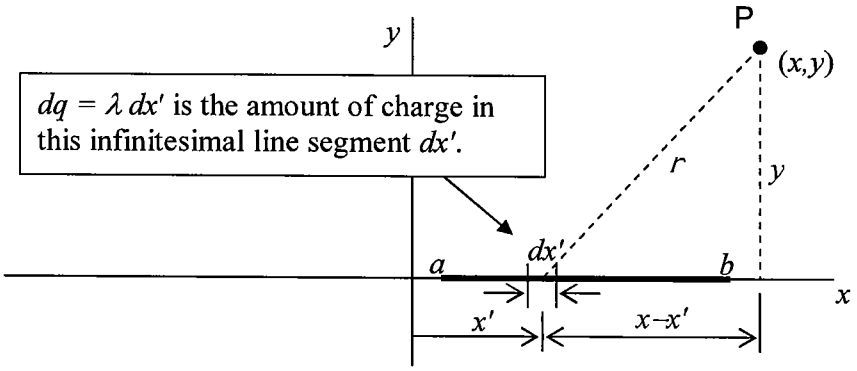
$$\vec{\mathbf{E}}=-\frac{kqd}{\left[x^2+\frac{d^2}{4}\right]^{\frac{3}{2}}}\hat{\jmath}$$

As expected, $\vec{\mathbf{E}}$ is in the –y direction. Note that to find the electric field on the x axis, you have to take the derivatives first, and *then* evaluate at $y=0$.

Example 32-2

A line of charge extends along the x axis from $x = a$ to $x = b$. On that line segment, the linear charge density λ is a constant. Find the electric potential as a function of position (x and y) due to that charge distribution on the x-y plane, and then, from the electric potential, determine the electric field on the x axis.

Solution: First, we need to use the methods of chapter 31 to get the potential for the specified charge distribution (a linear charge distribution with a constant linear charge density λ).



$$d\varphi = \frac{k\,dq}{r}$$

$$dq = \lambda dx' \quad \text{and} \quad r = \sqrt{(x-x')^2 + y^2}$$

$$d\varphi = \frac{k\lambda(x')dx'}{\sqrt{(x-x')^2 + y^2}}$$

$$\int d\varphi = \int_a^b \frac{k\lambda\,dx'}{\sqrt{(x-x')^2 + y^2}}$$

$$\varphi = k\lambda \int_a^b \frac{dx'}{\sqrt{(x-x')^2 + y^2}}$$

To carry out the integration, we use the variable substitution:

$u = x - x'$

$du = -dx' \;\Rightarrow\; dx' = -du$

Lower Integration Limit: When $x' = a$, $u = x - a$

Upper Integration Limit: When $x' = b$, $u = x - b$

Making these substitutions, we obtain:

$$\varphi = k\lambda \int_{x-a}^{x-b} \frac{-du}{\sqrt{u^2 + y^2}}$$

which I copy here for your convenience:

$$\varphi = k\lambda \int_{x-a}^{x-b} \frac{-du}{\sqrt{u^2 + y^2}}$$

Using the minus sign to interchange the limits of integration, we have:

$$\varphi = k\lambda \int_{x-b}^{x-a} \frac{du}{\sqrt{u^2 + y^2}}$$

Using the appropriate integration formula from the formula sheet we obtain:

$$\varphi = k\lambda \ln\left(u + \sqrt{u^2 + y^2}\right)\Big|_{x-b}^{x-a}$$

$$\boxed{\varphi = k\lambda \left\{ \ln\left[x - a + \sqrt{(x-a)^2 + y^2}\right] - \ln\left[x - b + \sqrt{(x-b)^2 + y^2}\right] \right\}}$$

Okay, that's the potential. Now we have to take the gradient of it and evaluate the result at $y=0$ to get the electric field on the x axis. We need to find

$$\vec{\mathbf{E}} = -\nabla\varphi$$

which, in the absence of any z dependence, can be written as:

$$\vec{\mathbf{E}} = -\left(\frac{\partial\varphi}{\partial x}\mathbf{\hat{\imath}} + \frac{\partial\varphi}{\partial y}\mathbf{\hat{\jmath}}\right)$$

We start by finding $\frac{\partial\varphi}{\partial x}$:

$$\frac{\partial\varphi}{\partial x} = \frac{\partial}{\partial x}\left(k\lambda\left\{ \ln\left[x - a + \sqrt{(x-a)^2 + y^2}\right] - \ln\left[x - b + \sqrt{(x-b)^2 + y^2}\right] \right\}\right)$$

$$\frac{\partial\varphi}{\partial x} = k\lambda\left\{ \frac{\partial}{\partial x}\ln\left[x - a + \left((x-a)^2 + y^2\right)^{\frac{1}{2}}\right] - \frac{\partial}{\partial x}\ln\left[x - b + \left((x-b)^2 + y^2\right)^{\frac{1}{2}}\right] \right\}$$

$$\frac{\partial\varphi}{\partial x} = k\lambda\left\{ \frac{1 + \frac{1}{2}\left((x-a)^2 + y^2\right)^{-\frac{1}{2}} 2(x-a)}{x - a + \left((x-a)^2 + y^2\right)^{\frac{1}{2}}} - \frac{1 + \frac{1}{2}\left((x-b)^2 + y^2\right)^{-\frac{1}{2}} 2(x-b)}{x - b + \left((x-b)^2 + y^2\right)^{\frac{1}{2}}} \right\}$$

$$\frac{\partial\varphi}{\partial x}=k\lambda\left\{\frac{1+(x-a)\left((x-a)^2+y^2\right)^{-\frac{1}{2}}}{x-a+\left((x-a)^2+y^2\right)^{\frac{1}{2}}}-\frac{1+(x-b)\left((x-b)^2+y^2\right)^{-\frac{1}{2}}}{x-b+\left((x-b)^2+y^2\right)^{\frac{1}{2}}}\right\}$$

Evaluating this at y = 0 yields:

$$\left.\frac{\partial\varphi}{\partial x}\right|_{y=0}=k\lambda\left(\frac{1}{x-a}-\frac{1}{x-b}\right)$$

Now, let's work on getting $\frac{\partial\varphi}{\partial y}$. I'll copy our result for φ from above and then take the partial derivative with respect to y (holding x constant):

$$\varphi=k\lambda\left\{\ln\left[x-a+\sqrt{(x-a)^2+y^2}\right]-\ln\left[x-b+\sqrt{(x-b)^2+y^2}\right]\right\}$$

$$\frac{\partial\varphi}{\partial y}=\frac{\partial}{\partial y}\left(k\lambda\left\{\ln\left[x-a+\sqrt{(x-a)^2+y^2}\right]-\ln\left[x-b+\sqrt{(x-b)^2+y^2}\right]\right\}\right)$$

$$\frac{\partial\varphi}{\partial y}=k\lambda\left\{\frac{\partial}{\partial y}\ln\left[x-a+\left((x-a)^2+y^2\right)^{\frac{1}{2}}\right]-\frac{\partial}{\partial x}\ln\left[x-b+\left((x-b)^2+y^2\right)^{\frac{1}{2}}\right]\right\}$$

$$\frac{\partial\varphi}{\partial y}=k\lambda\left\{\frac{\frac{1}{2}\left((x-a)^2+y^2\right)^{-\frac{1}{2}}2y}{x-a+\left((x-a)^2+y^2\right)^{\frac{1}{2}}}-\frac{\frac{1}{2}\left((x-b)^2+y^2\right)^{-\frac{1}{2}}2y}{x-b+\left((x-b)^2+y^2\right)^{\frac{1}{2}}}\right\}$$

$$\frac{\partial\varphi}{\partial y}=k\lambda\left\{\frac{y\left((x-a)^2+y^2\right)^{-\frac{1}{2}}}{x-a+\left((x-a)^2+y^2\right)^{\frac{1}{2}}}-\frac{y\left((x-b)^2+y^2\right)^{-\frac{1}{2}}}{x-b+\left((x-b)^2+y^2\right)^{\frac{1}{2}}}\right\}$$

Evaluating this at $y = 0$ yields:

$$\left.\frac{\partial\varphi}{\partial y}\right|_{y=0}=0$$

Plugging $\left.\frac{\partial\varphi}{\partial x}\right|_{y=0}=k\lambda\left(\frac{1}{x-a}-\frac{1}{x-b}\right)$ and $\left.\frac{\partial\varphi}{\partial y}\right|_{y=0}=0$ into $\vec{\mathbf{E}}=-\left(\frac{\partial\varphi}{\partial x}\mathbf{\uparrow}+\frac{\partial\varphi}{\partial y}\mathbf{\updownarrow}\right)$ yields:

$$\vec{\mathbf{E}} = -\left(k\lambda \left(\frac{1}{x-a} - \frac{1}{x-b} \right) \hat{\mathbf{i}} + 0\,\hat{\mathbf{j}} \right)$$

$$\boxed{\vec{\mathbf{E}} = k\lambda \left(\frac{1}{x-b} - \frac{1}{x-a} \right) \hat{\mathbf{i}}}$$

33 Gauss's Law

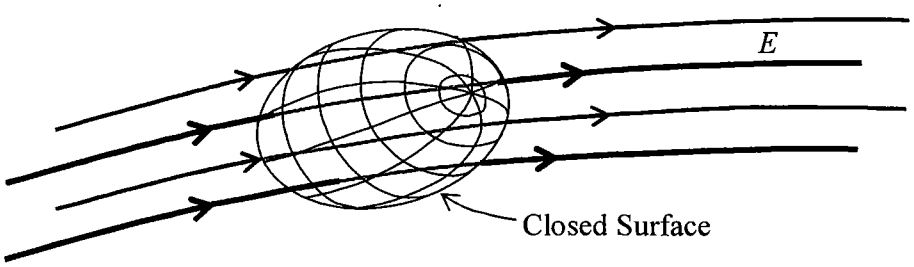
When asked to find the electric flux through a closed surface due to a specified non-trivial charge distribution, folks all too often try the immensely complicated approach of finding the electric field everywhere on the surface and doing the integral of $\vec{\mathbf{E}}$ *dot* $\vec{\mathbf{dA}}$ *over the surface instead of just dividing the total charge that the surface encloses by* ϵ_0.

Conceptually speaking, Gauss's Law states that **the number of electric field lines poking outward through an imaginary closed surface is proportional to the charge enclosed by the surface.**

A closed surface is one that divides the universe up into two parts: inside the surface, and, outside the surface. An example would be a soap bubble for which the soap film itself is of negligible thickness. I'm talking about a spheroidal soap bubble floating in air. Imagine one in the shape of a tin can, a closed jar with its lid on, or a closed box. These would also be closed surfaces. To be closed, a surface has to encompass a volume of empty space. A surface in the shape of a flat sheet of paper would not be a closed surface. In the context of Gauss's law, an imaginary closed surface is often referred to as a *Gaussian surface.*

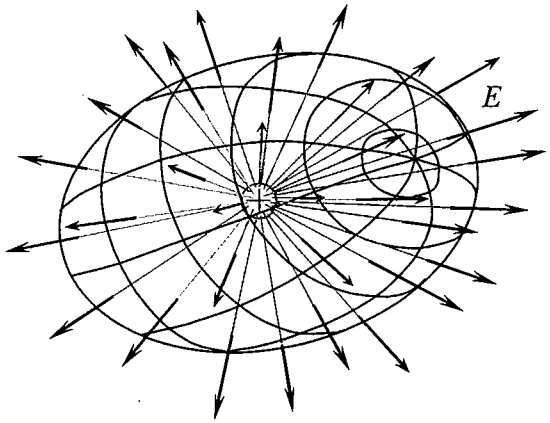
In conceptual terms, if you use Gauss's Law to determine how much charge is in some imaginary closed surface by counting the number of electric field lines poking outward through the surface, you have to consider inward-poking electric field lines as *negative* outward-poking field lines. Also, if a given electric field line pokes through the surface at more than one location, you have to count each and every penetration of the surface as another field line poking through the surface, adding +1 to the tally if it pokes outward through the surface, and –1 to the tally if it pokes inward through the surface.

So for instance, in a situation like:



we have 4 electric field lines poking inward through the surface which, together, count as –4 outward field lines, plus, we have 4 electric field lines poking *outward* through the surface which together count as +4 outward field lines for a total of 0 outward-poking electric field lines through the closed surface. By Gauss's Law, that means that the net charge inside the Gaussian surface is *zero.*

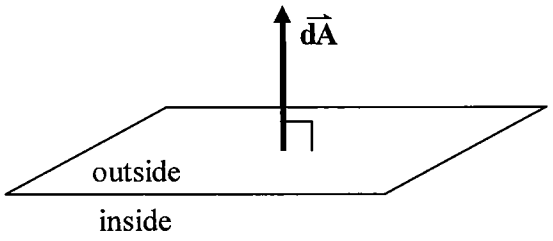
The following diagram might make our conceptual statement of Gauss's Law seem like plain old common sense to you:



The closed surface has the shape of an egg shell. There are 32 electric field lines poking outward through the Gaussian surface (and zero poking inward through it) meaning there must (according to Gauss's Law) be a net positive charge inside the closed surface. Indeed, from your understanding that electric field lines begin, either at positive charges or infinity, and end, either at negative charges or infinity, you could probably deduce our conceptual form of Gauss's Law. If the net number of electric field lines poking out through a closed surface is greater than zero, then you must have more lines *beginning* inside the surface than you have *ending* inside the surface, and, since field lines begin at positive charge, that must mean that there is more positive charge inside the surface than there is negative charge.

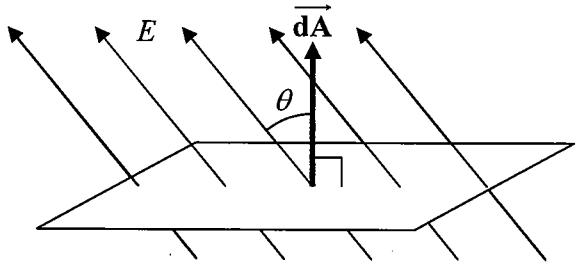
Our conceptual idea of the net number of electric field lines poking outward through a Gaussian surface corresponds to the net outward *electric flux* Φ_E through the surface.

To write an expression for the infinitesimal amount of outward flux $d\Phi_E$ through an infinitesimal area element dA, we first define an area element vector $\vec{\mathbf{dA}}$ whose magnitude is, of course, just the area dA of the element; and; whose direction is perpendicular to the area element, and, *outward*. (Recall that a closed surface separates the universe into two parts, an inside part and an outside part. Thus, at any point on the surface, that is to say at the location of any infinitesimal area element on the surface, the direction *outward*, away from the inside part, is unambiguous.)



In terms of that area element, and, the electric field $\vec{\mathbf{E}}$ at the location of the area element, we can write the infinitesimal amount of electric flux $d\Phi_E$ through the area element as:

$$d\Phi_E = \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}}$$



Recall that the dot product $\vec{\mathbf{E}} \cdot \vec{\mathbf{dA}}$ can be expressed as $E\,dA\cos\theta$. For a given E and a given amount of area, this yields a maximum value for the case of $\theta = 0°$ (when $\vec{\mathbf{E}}$ is parallel to $\vec{\mathbf{dA}}$ meaning that $\vec{\mathbf{E}}$ is perpendicular to the surface); zero when $\theta = 90°$ (when $\vec{\mathbf{E}}$ is perpendicular to $\vec{\mathbf{dA}}$ meaning that $\vec{\mathbf{E}}$ is parallel to the surface); and; a negative value when θ is greater than 90° (with 180° being the greatest value of θ possible, the angle at which $\vec{\mathbf{E}}$ is again perpendicular to the surface, but, in this case, *into* the surface.)

Now, the flux is the quantity that we can think of conceptually as the number of field lines. So, in terms of the flux, Gauss's Law states that the net outward flux through a closed surface is proportional to the amount of charge enclosed by that surface. Indeed, the constant of proportionality has been established to be $\frac{1}{\epsilon_0}$ where ϵ_0 (epsilon zero) is the universal constant known as the electric permittivity of free space. (You've seen ϵ_0 before. At the time, we stated that the Coulomb constant k is often expressed as $\frac{1}{4\pi\epsilon_0}$. Indeed, the identity $k = \frac{1}{4\pi\epsilon_0}$ appears on your formula sheet.) In equation form, Gauss's Law reads:

$$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0} \qquad (33\text{-}1)$$

The circle on the integral sign, combined with the fact that the infinitesimal in the integrand is an area element, means that the integral is over a closed surface. The quantity on the left is the sum of the product $\vec{\mathbf{E}} \cdot \vec{\mathbf{dA}}$ for each and every area element dA making up the closed surface. It is the total outward electric flux through the surface.

$$\Phi_E = \oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} \qquad (33\text{-}2)$$

Using this definition in Gauss's Law allows us to write Gauss's Law in the form:

$$\Phi_E = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0} \qquad (33\text{-}3)$$

How You Will be Using Gauss's Law

Gauss's Law is an integral equation. Such an integral equation can also be expressed as a differential equation. We won't be using the differential form, but, because of its existence, the Gauss's Law equation

$$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

is referred to as the *integral form* of Gauss's Law. The integral form of Gauss's Law can be used for several different purposes. In the course for which this book is written, you will be using it in a limited manner consistent with the mathematical prerequisites and co-requisites for the course. Here's how:

1) Gauss's Law in the form $\Phi_E = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$ makes it easy to calculate the net outward flux through a closed surface that encloses a known amount of charge Q_{ENCLOSED}. Just divide the amount of charge Q_{ENCLOSED} by ϵ_0 (given on your formula sheet as $\epsilon_0 = 8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N} \cdot \text{m}^2}$) and you have the flux through the closed surface.

2) Given the electric field at all points on a closed surface, one can use the integral form of Gauss's Law to calculate the charge inside the closed surface. This can be used as a check for a case in which the electric field due to a given distribution of charge has been calculated by a means other than Gauss's Law. You will only be expected to do this in cases in which one can treat the closed surface as being made of one or more finite (*not* vanishingly small) surface pieces on which the electric field is constant over the entire surface piece so that the flux can be calculated algebraically as EA or $EA\cos\theta$. After doing so for each of the finite surface pieces making up the closed surface, you add the results and you have the flux

$$\Phi_E = \oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}}$$

through the surface. To get the charge enclosed by the surface, you just plug that into $\Phi_E = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$ and solve for Q_{ENCLOSED}. If you are using the method as a check, you just compare your result with the amount of charge known to be enclosed by the surface.

3) In cases involving a symmetric charge distribution, Gauss's Law can be used to calculate the electric field due to the charge distribution. In such cases, the right choice of the Gaussian surface makes E a constant at all points on each of several surface pieces, and in some cases, zero on other surface pieces. In such cases the flux can be expressed as EA and one can simply solve $EA = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$ for E and use one's conceptual understanding of the electric field to get the direction of $\vec{\mathbf{E}}$. The remainder of this chapter and all of the next will be used to provide examples of the kinds of charge distributions to which you will be expected to be able to apply this method.

Using Gauss's Law to Calculate the Electric Field in the Case of a Charge Distribution Having Spherical Symmetry

A spherically-symmetric charge distribution has a well-defined center. Furthermore, if you rotate a spherically-symmetric charge distribution through any angle, about any axis that passes through the center, you wind up with the exact same charge distribution. A uniform ball of charge is an example of a spherically-symmetric charge distribution. Before we consider that one, however, let's take up the case of the simplest charge distribution of them all, *a point charge*.

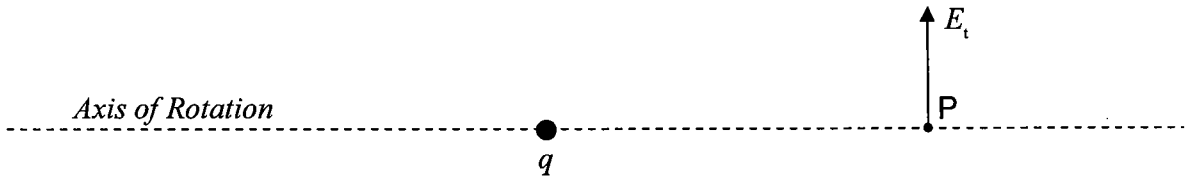
We use the symmetry of the charge distribution to find out as much as we can about the electric field and then we use Gauss's Law to do the rest. Now, when we rotate the charge distribution, we rotate the electric field with it. And, if a rotation of the charge distribution leaves you with the same exact charge distribution, then, it must also leave you with the same electric field.

We first prove that the electric field due to a point charge can have no tangential component by assuming that it does have a tangential component and showing that this leads to a contradiction.

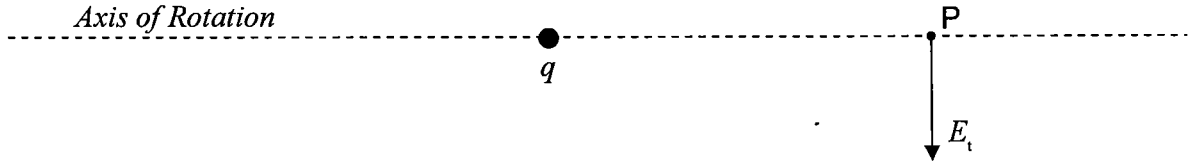
Here's our point charge q, and an assumed tangential component of the electric field at a point P which, from our perspective is to the right of the point charge.

(Note that a radial direction is any direction away from the point charge, and, a tangential direction is perpendicular to the radial direction.)

Now let's decide on a rotation axis for testing whether the electric field is symmetric with respect to rotation. Almost any will do. I choose one that passes through both the point charge, and, point P.



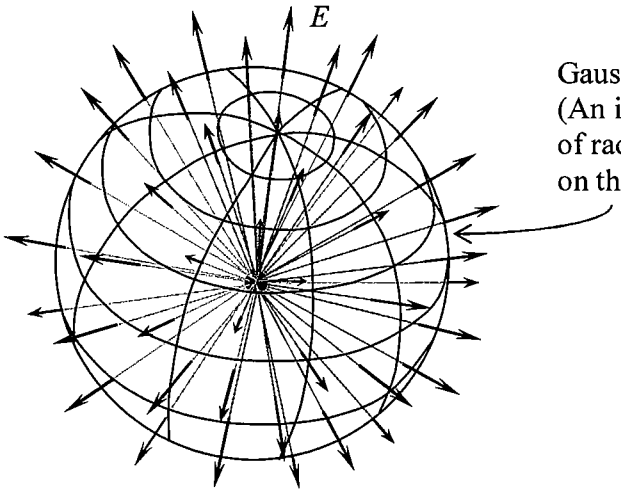
Now, if I rotate the charge, and its associated electric field, through an angle of 180° about that axis, I get:



This is different from the electric field that we started with. It is downward instead of upward. Hence the electric field cannot have the tangential component depicted at point P. Note that the argument does not depend on how far point P is from the point charge; indeed, I never specified the distance. So, no point to the right of our point charge can have an upward component to its electric field. In fact, if I assume the electric field at any point P′ in space other than the point at which the charge is, to have a tangential component, then, I can adopt a viewpoint from which point P′ appears to be to the right of the charge, and, the electric field appears to be upward. From that viewpoint, I can make the same rotation argument presented above to prove that the tangential component cannot exist. Thus, based on the spherical symmetry of the charge distribution, the electric field due to a point charge has to be strictly radial. Thus, at each point in space, the electric field must be either directly toward the point charge or directly away from it. Furthermore, again from symmetry, if the electric field is directly away from the point charge at one point in space, then it has to be directly away from the point charge at every point in space. Likewise, for the case in which it is directly toward the point charge at one point in space, the electric field has to be directly toward the point charge at every point in space.

We've boiled it down to a 50/50 choice. Let's assume that the electric field is directed *away* from the point charge at every point in space and use Gauss's Law to calculate the magnitude of the electric field. If the magnitude is positive, then the electric field is indeed directed away from the point charge. If the magnitude turns out to be negative, then the electric field is actually directed toward the point charge.

At this point we need to choose a Gaussian surface. To further exploit the symmetry of the charge distribution, we choose a Gaussian surface with spherical symmetry. More specifically, we choose a spherical shell of radius r, centered on the point charge.



At every point on the shell, the electric field, being radial, has to be perpendicular to the spherical shell. This means that for every area element, the electric field is parallel to our outward-directed area element vector $\vec{\mathbf{dA}}$. This means that the $\vec{\mathbf{E}} \cdot \vec{\mathbf{dA}}$ in Gauss's Law,

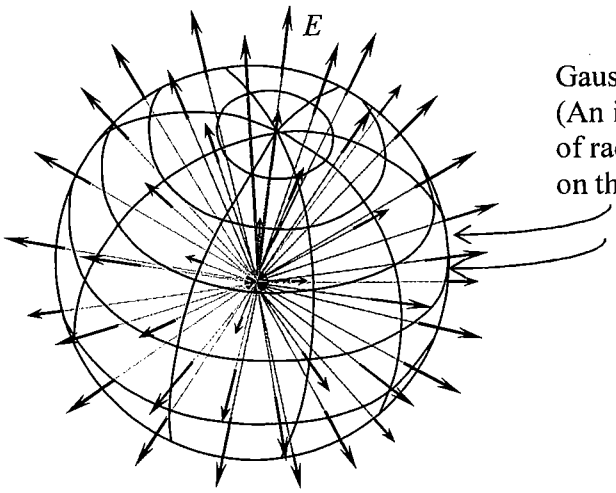
$$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

evaluates to EdA. So, for the case at hand, Gauss's Law takes on the form:

$$\oint E\, dA = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

Furthermore, the magnitude of the electric field has to have the same value at every point on the shell. If it were different at a point P′ on the spherical shell than it is at a point P on the spherical shell, then we could rotate the charge distribution about an axis through the point charge in such a manner as to bring the original electric field at point P′ to position P. But this would represent a change in the electric field at point P, due to the rotation, in violation of the fact that a point charge has spherical symmetry. Hence, the electric field at any point P′ on the Gaussian surface must have the same magnitude as the electric field at point P, which is what I set out to prove. The fact that E is a constant, in the integral, means that we can factor it out of the integral. So, for the case at hand, Gauss's Law takes on the form:

$$E \oint dA = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$



Gaussian Surface
(An imaginary sphere of radius r, centered on the point charge.)

On the preceding page we arrived at $E\oint dA = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$.

Now $\oint dA$, the integral of dA over the Gaussian surface is the sum of all the area elements making up the Gaussian surface. That means that it is just the total area of the Gaussian surface. The Gaussian surface, being a sphere of radius r, has area $4\pi r^2$. So now, Gauss's Law for the case at hand looks like:

$$E4\pi r^2 = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

Okay, we've left that right side alone for long enough. We're talking about a point charge q and our Gaussian surface is a sphere centered on that point charge q, so, the charge enclosed, Q_{ENCLOSED} is obviously q. This yields:

$$E4\pi r^2 = \frac{q}{\epsilon_0}$$

Solving for E gives us:

$$E = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}$$

This is positive when the charge q is positive, meaning that the electric field is directed outward, as per our assumption. It is negative when q is negative. So, when the charge q is negative, the electric field is directed inward, toward the charged particle. This expression is, of course, just Coulomb's Law for the electric field. It may look more familiar to you if we write it in terms of the Coulomb constant $k = \frac{1}{4\pi\epsilon_0}$ in which case our result for the outward electric field appears as:

$$E = \frac{kq}{r^2}$$

It's clear that, by means of our first example of Gauss's Law, we have derived something that you already know, the electric field due to a point charge. .

34 Gauss's Law Example

We finished off the last chapter by using Gauss's Law to find the electric field due to a point charge. It was an example of a charge distribution having spherical symmetry. In this chapter we provide another example involving spherical symmetry.

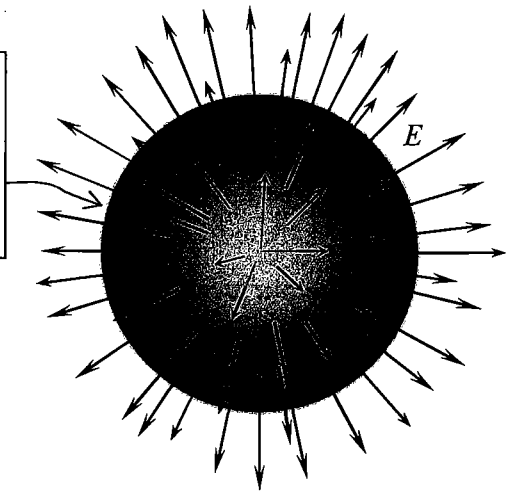
Example 34-1

Find the electric field due to a uniform ball of charge of radius R and total charge Q. Express the electric field as a function of r, the distance from the center of the ball.

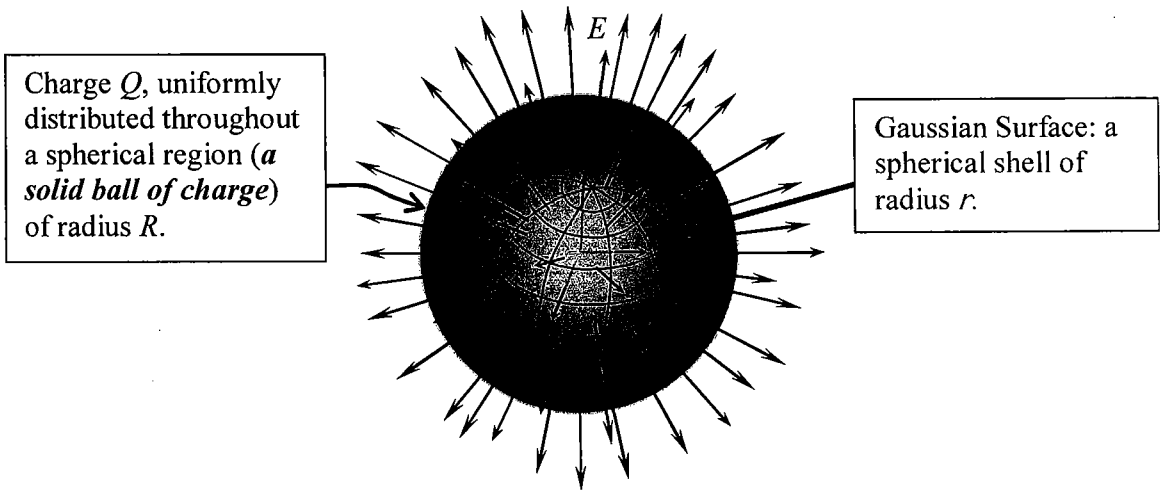
Solution

Again we have a charge distribution for which a rotation through any angle about any axis passing through the center of the charge distribution results in the exact same charge distribution. Thus, the same symmetry arguments used for the case of the point charge apply here with the result that, the electric field due to the ball of charge has to be strictly radially directed, and, the electric field has one and the same value at every point on any given spherical shell centered on the center of the ball of charge. Again, we assume the electric field to be outward-directed. If it turns out to be inward-directed, we'll simply get a negative value for the magnitude of the outward-directed electric field.

Charge Q, uniformly distributed throughout a spherical region (***a solid ball of charge***) of radius R.



The appropriate Gaussian surface for any spherical charge distribution is a spherical shell centered on the center of the charge distribution.



Okay, let's go ahead and apply Gauss's Law.

$$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

Since the electric field is radial, it is, at all points, perpendicular to the Gaussian Surface. In other words, it is parallel to the area element vector $\vec{\mathbf{dA}}$. This means that the dot product $\vec{\mathbf{E}} \cdot \vec{\mathbf{dA}}$ is equal to the product of the magnitudes, $E\,dA$. This yields:

$$\oint E\,dA = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

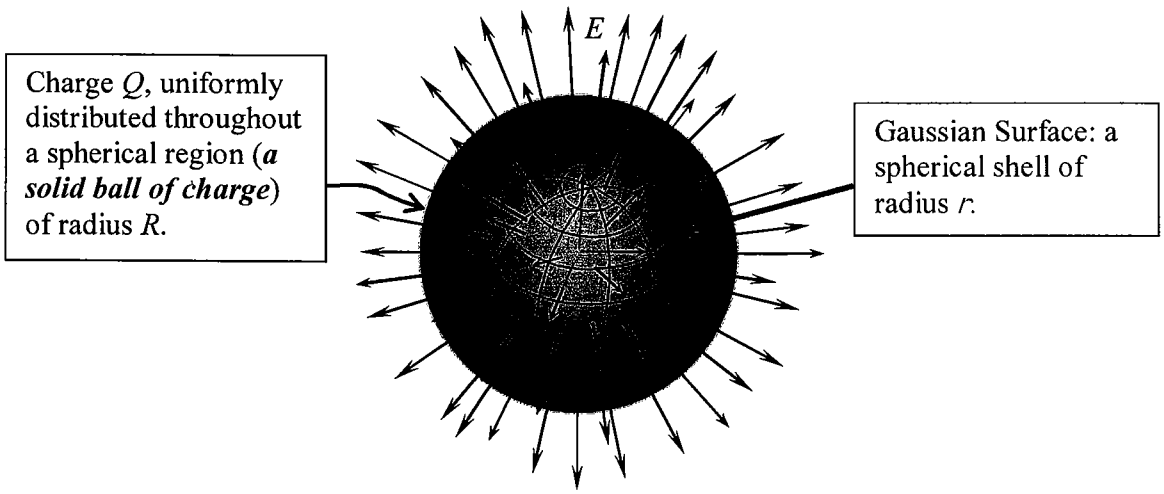
Again, since E has the same value at all points on the Gaussian surface of radius r, each dA in the infinite sum that the integral on the left is, is multiplied by the same value of E. Hence, we can factor the E out of the sum (integral). This yields

$$E\oint dA = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

The integral on the left is just the infinite sum of all the infinitesimal area elements making up the Gaussian surface, our spherical shell of radius r. The sum of all the area elements is, of course, the area of the spherical shell. The area of a sphere is $4\pi r^2$. So,

$$E4\pi r^2 = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

Now the question is, how much charge is enclosed by our Gaussian surface of radius r?



There are two ways that we can get the value of the charge enclosed. Let's try it both ways and make sure we get one and the same result.

The first way: Because the charge is *uniformly* distributed throughout the volume, the amount of charge enclosed is directly proportional to the volume enclosed. So, the ratio of the amount of charge enclosed to the total charge, is equal to the ratio of the volume enclosed *by the Gaussian surface* to the total volume of the ball of charge:

$$\frac{Q_{\text{ENCLOSED}}}{Q} = \frac{\text{Volume of Gaussian Surface}}{\text{Volume of the Entire Ball of Charge}}$$

$$\frac{Q_{\text{ENCLOSED}}}{Q} = \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi R^3}$$

$$Q_{\text{ENCLOSED}} = \frac{r^3}{R^3} Q$$

The second way: The other way we can look at it is to recognize that for a *uniform* distribution of charge, the amount of charge enclosed by the Gaussian surface is just the volume charge density, that is, the charge-per-volume ρ, times the volume enclosed.

$$Q_{\text{ENCLOSED}} = \rho(\text{Volume of the Gaussian Surface})$$

$$Q_{\text{ENCLOSED}} = \rho \tfrac{4}{3}\pi r^3$$

In this second method, we again take advantage of the fact that we are dealing with a *uniform* charge distribution. In a uniform charge distribution, the charge density is just the total charge divided by the total volume. Thus:

$$\rho = \frac{Q}{\text{Volume of Ball of Charge}}$$

$$\rho = \frac{Q}{\frac{4}{3}\pi R^3}$$

Substituting this in to our expression $Q_{\text{ENCLOSED}} = \rho 4\pi r^2$ for the charge enclosed by the Gaussian surface yields:

$$Q_{\text{ENCLOSED}} = \frac{Q}{\frac{4}{3}\pi R^3}\frac{4}{3}\pi r^3$$

$$Q_{\text{ENCLOSED}} = \frac{r^3}{R^3}Q$$

which is indeed the same expression that we arrived at in solving for the charge enclosed the first way we talked about.

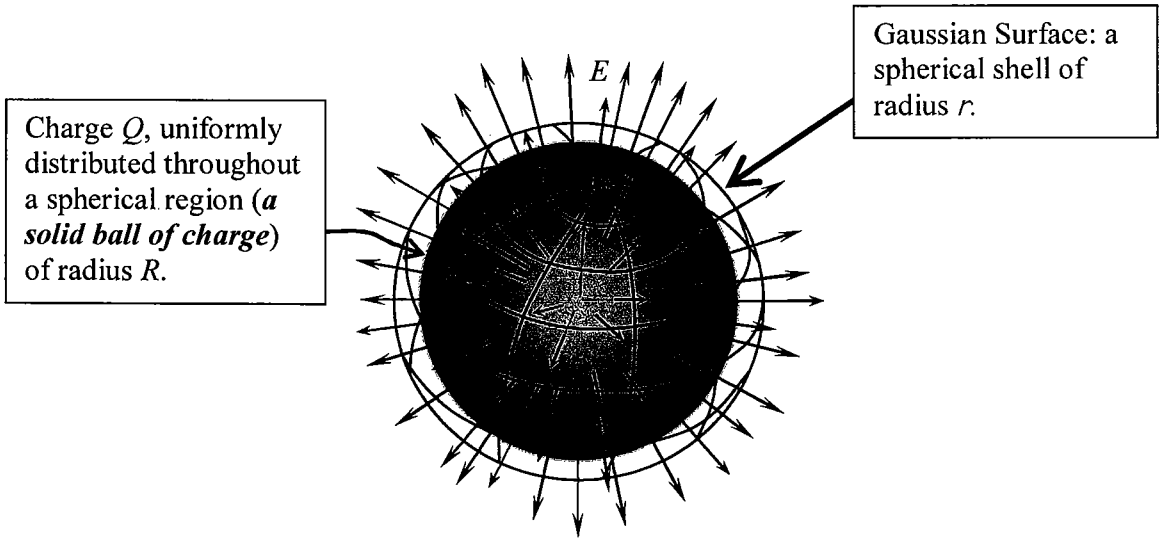
A couple of pages back we used Gauss's Law to arrive at the relation $E4\pi r^2 = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$ and now we have something to plug in for Q_{ENCLOSED}. Doing so yields:

$$E4\pi r^2 = \frac{\left(\frac{r^3}{R^3}Q\right)}{\epsilon_0}$$

$$E = \frac{Q}{4\pi\epsilon_0 R^3}r$$

This is our result for the magnitude of the electric field due to a uniform ball of charge at points inside the ball of charge ($r \le R$). E is directly proportional to the distance from the center of the charge distribution. E increases with increasing distance because, the farther a point is from the center of the charge distribution, the more charge there is inside the spherical shell that is centered on the charge distribution and upon which the point in question is situated. How about points for which $r \ge R$?

If $r \geq R$,



the analysis is identical to the preceding analysis up to and including the point where we determined that:

$$E4\pi r^2 = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

But as long as $r \geq R$, no matter by how much r exceeds R, all the charge in the spherical distribution of charge is enclosed by the Gaussian surface. "All the charge" is just Q the total amount of charge in the uniform ball of charge. So,

$$E4\pi r^2 = \frac{Q}{\epsilon_0}$$

$$E = \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}$$

The constant $\frac{1}{4\pi\epsilon_0}$ is just the Coulomb constant k so we can write our result as:

$$E = \frac{kQ}{r^2}$$

This result looks just like Coulomb's Law for a point charge. What we've proved here is that, at points outside a spherically-symmetric charge distribution, the electric field is the same as that due to a point charge at the center of the charge distribution.

35 Gauss's Law for the Magnetic Field, and, Ampere's Law Revisited

Gauss's Law for the Magnetic Field

Remember Gauss's Law for the *electric* field? It's the one that, in conceptual terms, states that the number of *electric* field lines poking outward through a closed surface is proportional to the amount of *electric* charge inside the closed surface. In equation form, we wrote it as:

$$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

We called the quantity on the left the electric flux $\Phi_E = \oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}}$.

Well, there is a Gauss's Law for the *magnetic* field as well. In one sense, it is quite similar because it involves a quantity called the magnetic flux which is expressed mathematically as $\Phi_B = \oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}}$ and represents the number of magnetic field lines poking outward through a closed surface. The big difference stems from the fact that there is no such thing as "magnetic charge." In other words, there is no such thing as a *magnetic monopole*. In Gauss's Law for the *electric* field we have electric charge (divided by ϵ_0) on the right. In Gauss's Law for the magnetic field, we have *0* on the right:

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}} = 0 \qquad (35\text{-}1)$$

As far as calculating the magnetic field, this equation is of limited usefulness. But, in conjunction with Ampere's Law in integral form (see below), it can come in handy for calculating the magnetic field in cases involving a lot of symmetry. Also, it can be used as a check for cases in which the magnetic field has been determined by some other means.

Ampere's Law

We've talked about Ampere's Law quite a bit already. It's the one that says *a current causes a magnetic field.* Note that this one says nothing about anything changing. It's just a cause and effect relation. The integral form of Ampere's Law is both broad and specific. It reads:

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_0 I_{\text{THROUGH}} \qquad (35\text{-}2)$$

where:

- the circle on the integral sign, and, $\vec{\mathbf{dl}}$, the differential length, together, tell you that the integral (the infinite sum) is around an imaginary *closed loop.*

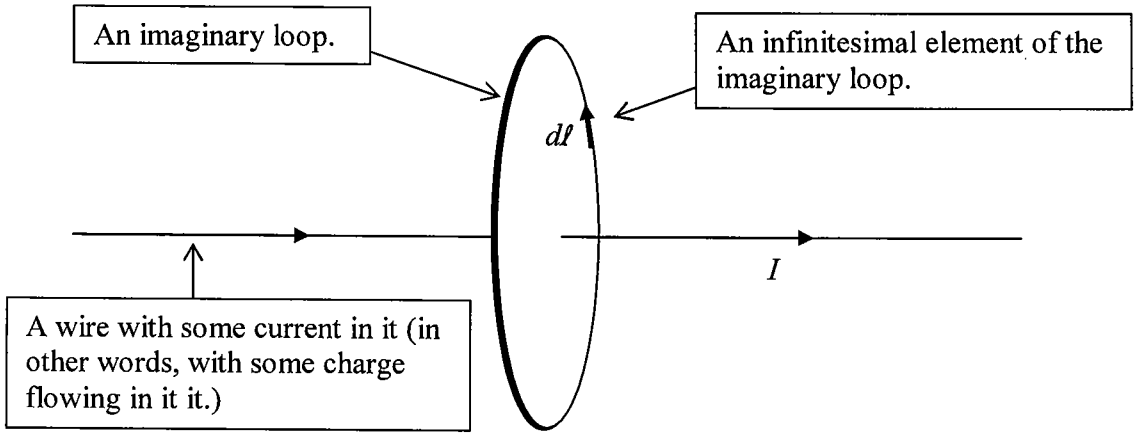
$\vec{\mathbf{B}}$ is the magnetic field,

$\vec{\mathbf{dl}}$ is an infinitesimal path element of the closed loop,

μ_0 is a universal constant called the magnetic permeability of free space, and

I_{THROUGH} is the current passing through the region enclosed by the loop.

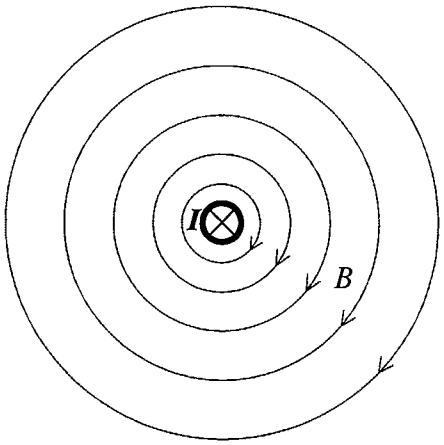
What Ampere's Law in integral form says is that, if you sum up the magnetic-field-along-a-path-segment times the length of the path segment for all the path segments making up an imaginary closed loop, you get the current through the region enclosed by the loop, times a universal constant. The integral $\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}}$ on whatever closed path upon which it is carried out, is called the *circulation* of the magnetic field on that closed path. So, another way of stating the integral form of Ampere's Law is to say that the circulation of the magnetic field on any closed path is directly proportional to the current through the region enclosed by the path. Here's the picture:



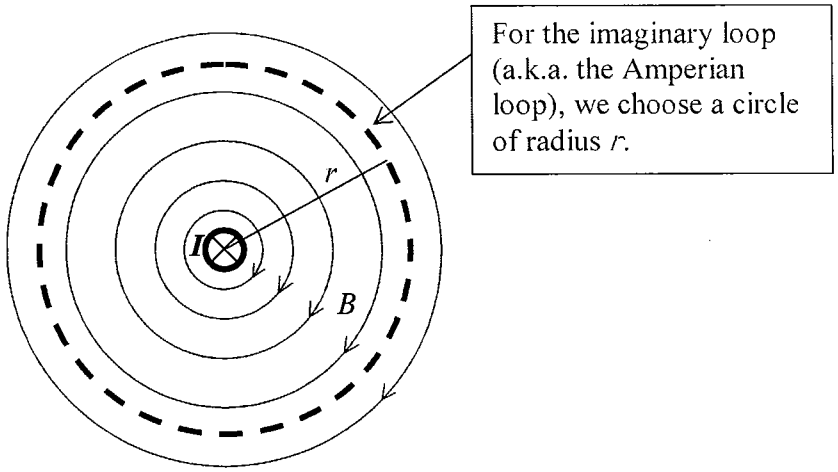
In the picture, I show everything except for the magnetic field. The idea is that, for each infinitesimal segment $\vec{\mathbf{dl}}$ of the imaginary loop, you dot the magnetic field $\vec{\mathbf{B}}$, at the position of the segment, into $\vec{\mathbf{dl}}$. Add up all such dot products. The total is equal to μ_0 times the current I through the loop.

So, what's it good for? Ampere's Law in integral form is of limited use to us. It can be used as a great check for a case in which one has calculated the magnetic field due to some set of current-carrying conductors some other way (e.g. using the Biot-Savart Law, to be introduced in the next chapter). Also, in cases involving a high degree of symmetry, we can use it to calculate the magnetic field due to some current.

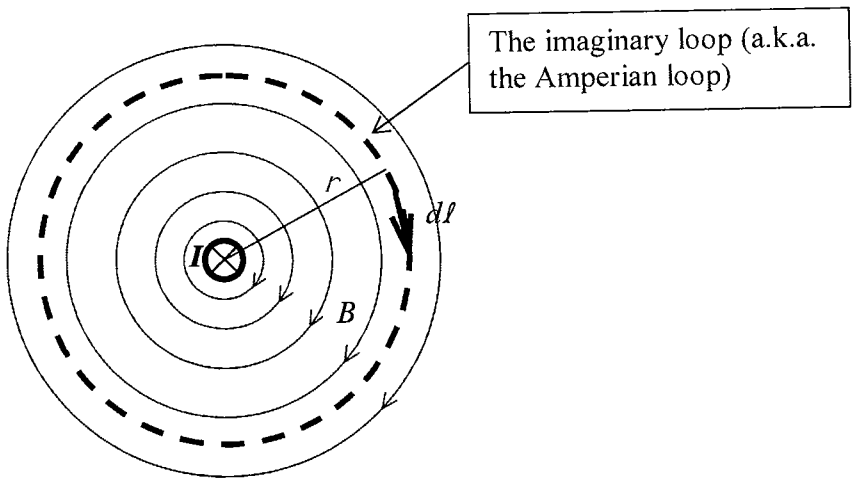
For example, we can use Ampere's Law to get a mathematical expression for the magnitude of the magnetic field due to an infinitely-long straight wire. I'm going to incorporate our understanding that, for a segment of wire with a current in it, the current creates a magnetic field which forms loops about the wire in accord with the right-hand rule for something curly something straight. In other words, we already know that for a long straight wire carrying current directly away from you, the magnetic field extends in loops about the wire, which, from your vantage point, are clockwise.



From symmetry, we can argue that the magnitude of the magnetic field is the same for a given point as it is at any other point that is the *same distance* from the wire as that given point. In implementing Ampere's Law, it is incumbent upon us to choose an imaginary loop, called an Amperian Loop in this context, that allows us to get some useful information from Ampere's Law. In this case, a circle whose plane is perpendicular to the straight wire and whose center lies on the straight wire is a smart choice.



At this point I want to share with you some directional information about the integral form of Ampere's Law. Regarding the $\vec{d\ell}$: each $\vec{d\ell}$ vector can, from a given point of view, be characterized as representing either a clockwise step along the path or a counterclockwise step along the path. And, if one is clockwise, they all have to be clockwise. If one is counterclockwise, they all have to be counterclockwise. Thus, in carrying out the integral around the closed loop, the traversal of the loop is either clockwise or counterclockwise from a specified viewpoint. Now, here's the critical direction information: Current that passes through the loop in that direction which relates to the sense (clockwise or counterclockwise) of loop traversal in accord with the right-hand rule for something curly something straight (with the loop being the something curly and the current being the something straight) is considered positive. So, for the case at hand, if I choose a clockwise loop traversal, as viewed from the vantage point that makes things look like:



then, the current I is considered positive. If you curl the fingers around the loop in the clockwise direction, your thumb points away from you. This means that current, through the loop that is directed away from you is positive. That is just the kind of current we have in the case at hand. So, when we substitute the I for the case at hand into the generic equation (Ampere's Law),

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{d\ell}} = \mu_0 I_{\text{THROUGH}}$$

for the current I_{THROUGH} it goes in with a "+" sign.

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{d\ell}} = \mu_0 I$$

Now, with the loop I chose, every $\vec{\mathbf{d\ell}}$ is exactly parallel to the magnetic field $\vec{\mathbf{B}}$ at the location of the $\vec{\mathbf{d\ell}}$, so, $\vec{\mathbf{B}} \cdot \vec{\mathbf{d\ell}}$ is simply $B\,d\ell$. That is, with our choice of Amperian loop, Ampere's Law simplifies to:

$$\oint B\,d\ell = \mu_0 I$$

Furthermore, from symmetry, with our choice of Amperian loop, the magnitude of the magnetic field B has one and the same value at every point on the loop. That means we can factor the magnetic field magnitude B out of the integral. This yields:

$$B \oint d\ell = \mu_0 I$$

Okay, now we are on easy street. The $\oint d\ell$ is just the sum of all the $d\ell$'s making up our imaginary loop (a circle) of radius r. Hey, that's just the circumference of the circle $2\pi r$. So, Ampere's Law becomes:

$$B(2\pi r) = \mu_0 I$$

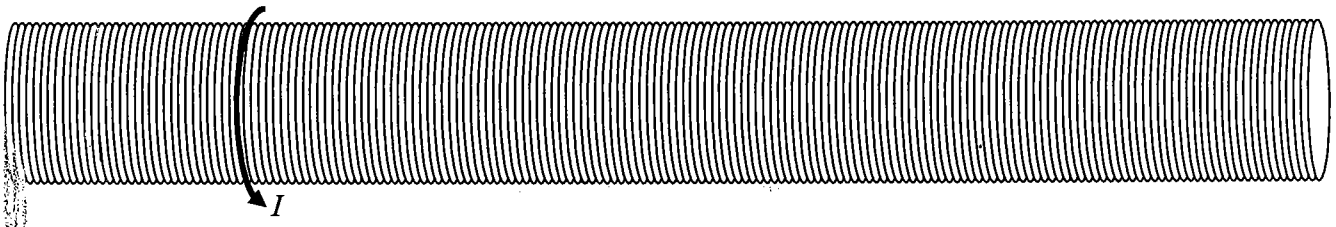
which means

$$B = \frac{\mu_0 I}{2\pi r}$$

This is our end result. The magnitude of the magnetic field due to a long straight wire is directly proportional to the current in the wire and inversely proportional to the distance from the wire.

A Long Straight Solenoid

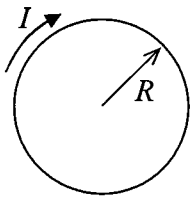
A solenoid is a coil of wire in the form of a cylindrical shell. The idealized solenoid that we consider here is infinitely long but, it has a fixed finite radius R and a constant finite current I.



It is also characterized by its number-of-turns-per-length, n, where each "turn" (a.k.a. winding) is one circular current loop. In fact, we further idealize our solenoid by thinking of it as an infinite set of circular current loops. An actual solenoid approaches this idealized solenoid, but, in one turn (in the view above), the end of the turn is displaced left or right from the start of the turn by an amount equal to the diameter of the wire. As a result, in an actual solenoid, we have (in the view above) some left-to-right or right-to-left (depending on which way the wire wraps around) current. We neglect this current and consider the current to just go "round and round."

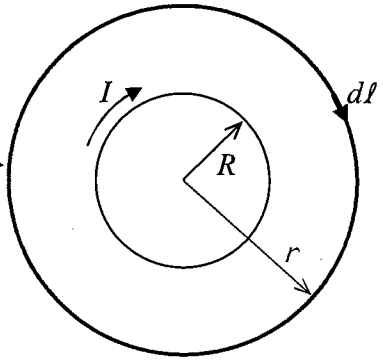
Our goal here is to find the magnetic field due to an ideal infinitely-long solenoid that has a number-of-turns-per-length n, has a radius R, and carries a current I.

We start by looking at the solenoid in cross section. Relative to the view above, we'll imagine looking at the solenoid from the left end. From that point of view, the cross section is a circle with clockwise current:



Let's try an Amperian loop in the shape of a circle, whose plane is perpendicular to the axis of symmetry of the solenoid, a circle that is centered on the axis of symmetry of the solenoid.

For the imaginary loop (a.k.a. the Amperian loop), we choose a circle of radius r.

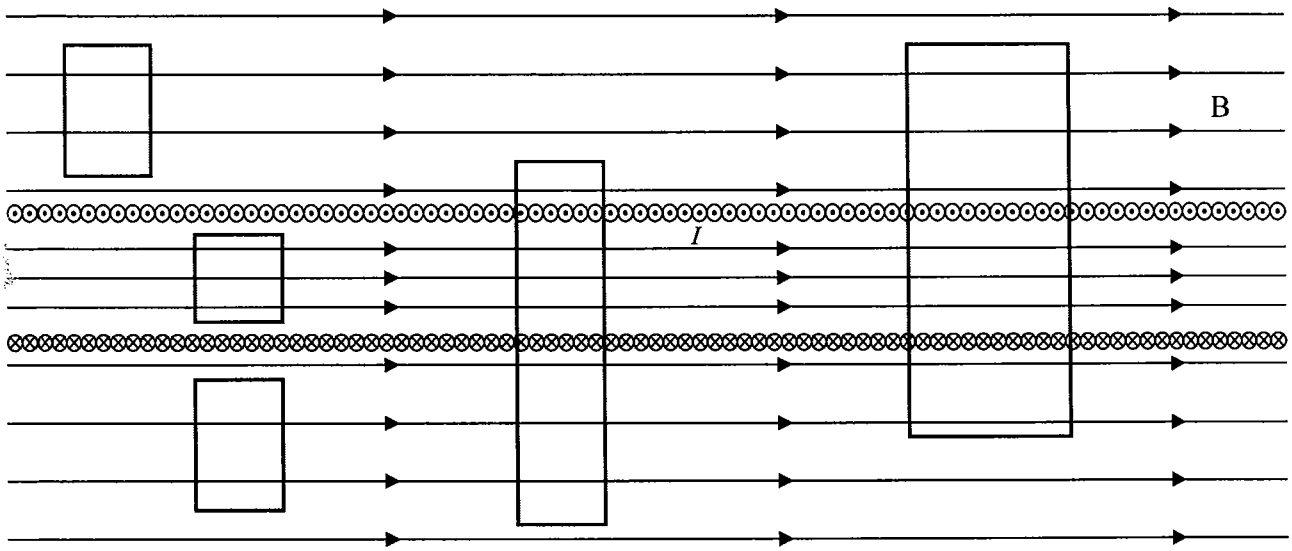


From symmetry, we can argue that if the magnetic field has a component parallel to the depicted $\vec{d\ell}$, then it must have the exact same component for every $\vec{d\ell}$ on the closed path. But this would make the circulation $\oint \vec{B} \cdot \vec{d\ell}$ non-zero in contradiction to the fact that no current passes through the region enclosed by the loop. This is true for any value of r. So, the magnetic field can have no component tangent to the circle whose plane is perpendicular to the axis of symmetry of the solenoid, a circle that is centered on the axis of symmetry of the solenoid.

Now suppose the magnetic field has a radial component. By symmetry it would have to be everywhere directed radially outward from the axis of symmetry of the solenoid, or everywhere radially inward. In either case, we could construct an imaginary cylindrical shell whose axis of symmetry coincides with that of the solenoid. The net magnetic flux through such a Gaussian surface would be non-zero in violation of Gauss's Law for the magnetic field. Hence the solenoid can have no radial magnetic field component.

The only kind of field that we haven't ruled out is one that is everywhere parallel to the axis of symmetry of the solenoid. Let's see if such a field would lead to any contradictions.

Here we view the solenoid in cross-section from the side. At the top of the coil, we see the current directed toward us, and, at the bottom, away. The possible longitudinal (parallel to the axis of symmetry of the solenoid) magnetic field is included in the diagram.



The rectangles in the diagram represent Amperian loops. The net current through any of the loops, in either direction (away from you or toward you) is zero. As such, the circulation $\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}}$ is zero. Since the magnetic field on the right and left of any one of the loops is perpendicular to the right and left sides of any one of the loops, it makes no contribution to the circulation there. By symmetry, the magnetic field at one position on the top of a loop is the same as it is at any other point on the top of the same loop. Hence, if we traverse any one of the loops counterclockwise (from our viewpoint) the contribution to the circulation is $-B_{\text{TOP}}L$ where L is the length of the top and bottom segments of whichever loop you choose to focus your attention on. The "–" comes from the fact that I have (arbitrarily) chosen to traverse the loop counterclockwise, and, in doing so, every $\vec{\mathbf{dl}}$ in the top segment is in the opposite direction to the direction of the magnetic field at the top of the loop. The contribution to the circulation by the bottom segment of the same loop is $+B_{\text{BOTTOM}}L$. What we have so far is:

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_0 I_{\text{THROUGH}}$$

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = 0$$

(where the net current through any one of the loops depicted is zero by inspection.)

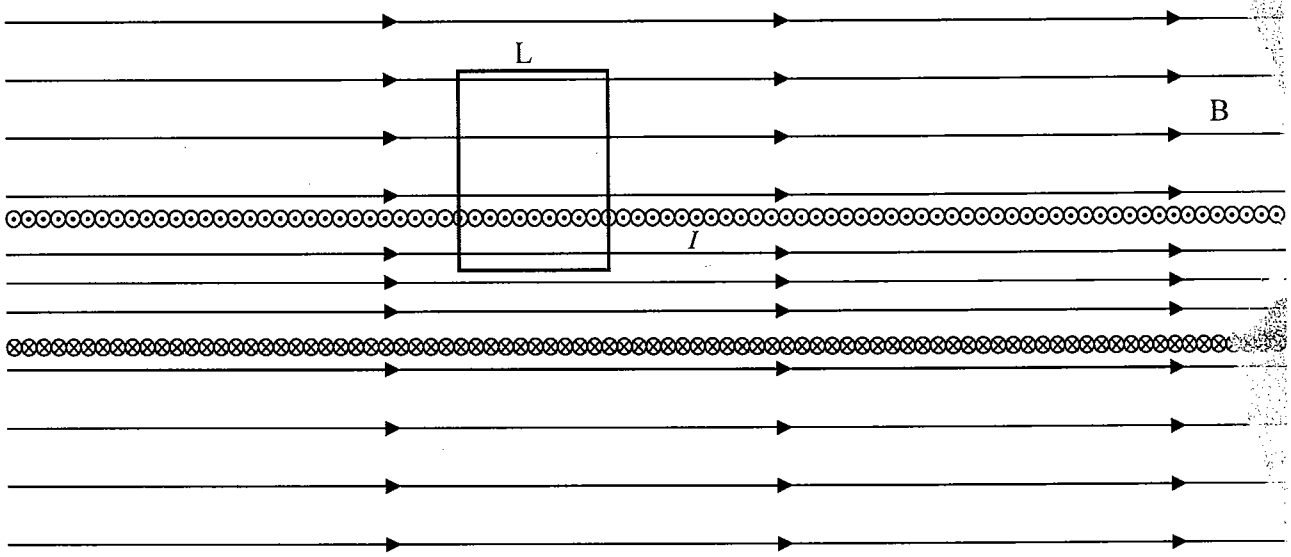
$$0 + -B_{\text{TOP}}L + 0 + B_{\text{BOTTOM}}L = 0$$

(with the two zeros on the left side of the equation being from the right and left sides of the loop where the magnetic field is perpendicular to the loop.)

Solving for B_{BOTTOM} we find that, for every loop in the diagram (and the infinite number of loops enclosing a net current of zero just like them):

$$B_{\text{BOTTOM}} = B_{\text{TOP}}$$

What this means is that the magnetic field at all points outside the solenoid has one and the same magnitude. The same can be said about all points inside the solenoid, but, the inside-the-solenoid value may be different from the outside value. In fact, let's consider a loop through which the net current is not zero:



Again, I choose to go counterclockwise around the loop (from our viewpoint). As such, by the right-hand rule for something curly something straight, current directed toward us through the loop is positive. Recalling that the number-of-turns-per-length-of-the-solenoid is n, we have, for the loop depicted above,

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_0 I_{\text{THROUGH}}$$

$$0 + -B_{\text{TOP}} L + 0 + B_{\text{BOTTOM}} L = \mu_0 n L I$$

$$B_{\text{BOTTOM}} = B_{\text{TOP}} + \mu_0 n I$$

The bottom of the loop is inside the solenoid and we have established that the magnitude of the magnetic field inside the solenoid has one and the same magnitude at all points inside the solenoid. I'm going to call that B_{INSIDE}, meaning that $B_{\text{BOTTOM}} = B_{\text{INSIDE}}$. Similarly, we have found that the magnitude of the magnetic field has one and the same (other) value at all points outside the solenoid. Let's call that B_{OUTSIDE}, meaning that $B_{\text{TOP}} = B_{\text{OUTSIDE}}$. Thus:

$$B_{\text{INSIDE}} = B_{\text{OUTSIDE}} + \mu_0 n I$$

This is as far as I can get with Gauss's Law for the magnetic field, symmetry, and Ampere's Law alone. From there I turn to experimental results with long finite solenoids. Experimentally, we find that the magnetic field outside the solenoid is vanishingly small, and that there is an appreciable magnetic field inside the solenoid. Setting

$$B_{\text{OUTSIDE}} = 0$$

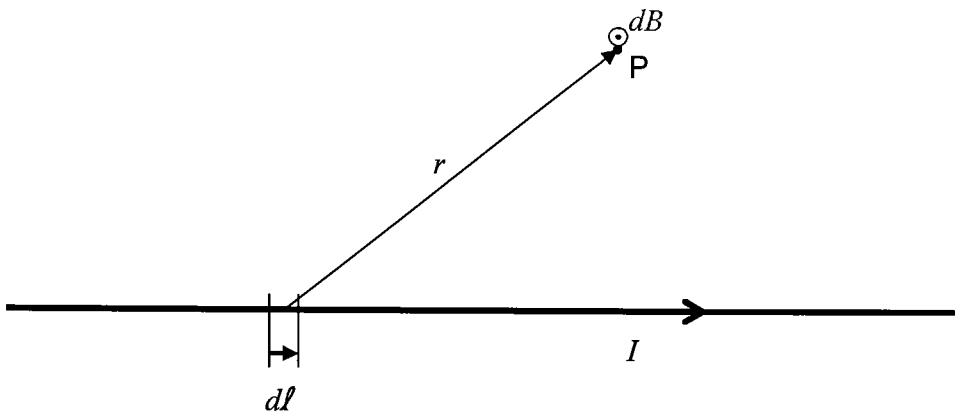
we find that the magnetic field inside a long straight solenoid is:

$$B_{\text{INSIDE}} = \mu_o n I$$

36 The Biot-Savart Law

The Biot-Savart Law provides us with a way to find the magnetic field at an empty point in space, let's call it point P, due to current in wire. The idea behind the Biot-Savart Law is that each infinitesimal element of the current-carrying wire makes an infinitesimal contribution to the magnetic field at the empty point in space. Once you find each contribution, all you have to do is add them all up. Of course, there are an infinite number of contributions to the magnetic field at point P and each one is a vector, so, we are talking about an infinite sum of vectors. This business should seem familiar to you. You did this kind of thing when you were calculating the *electric* field back in *Chapter 30 The Electric Field Due to a Continuous Distribution of Charge on a Line.* The idea is similar, but here, of course, we are talking about magnetism.

The Biot-Savart Law gives the infinitesimal contribution to the magnetic field at point P due to an infinitesimal element of the current-carrying wire. The following diagram helps to illustrate just what the Biot-Savart Law tells us.



The Biot-Savart Law states that:

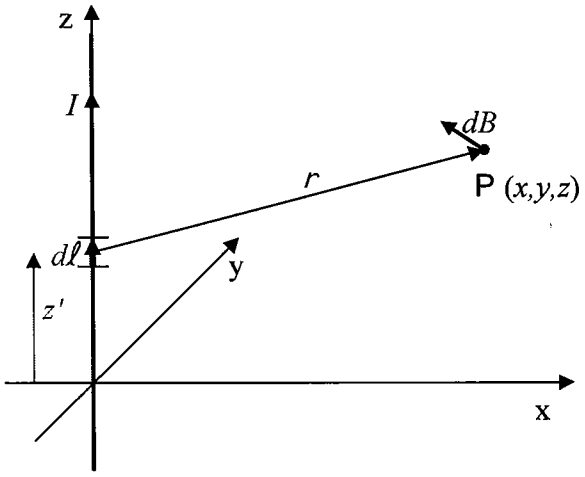
$$\vec{\mathbf{dB}} = \frac{\mu_o}{4\pi}\frac{I\,\vec{\mathbf{d\ell}}\times\vec{\mathbf{r}}}{r^3} \qquad (36\text{-}1)$$

The Biot-Savart Law represents a powerful straightforward method of calculating the magnetic field due to a current distribution.

Example 36-1

Calculate the magnetic field due to a long straight wire carrying a current I along the z axis in the positive z direction. Treat the wire as extending to infinity in both directions.

Solution



Each infinitesimal element of the current-carrying conductor makes a contribution $\mathbf{d\vec{B}}$ to the total magnetic field at point P.

The $\vec{r}$ vector extends from the infinitesimal element at $(0,0,z')$ to point P at (x, y, z).

$$\vec{r} = (x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}) - z'\hat{\mathbf{k}}$$

$$\vec{r} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + (z - z')\hat{\mathbf{k}}$$

The magnitude of $\vec{r}$ is thus:

$$r = \sqrt{x^2 + y^2 + (z - z')^2}$$

The $\mathbf{d\vec{\ell}}$ vector points in the +z direction so it can be expressed as $\mathbf{d\vec{\ell}} = dz'\hat{\mathbf{k}}$

With these expressions for $\vec{r}$, r, and $\mathbf{d\vec{\ell}}$ substituted into the Biot-Savart Law,

$$\mathbf{d\vec{B}} = \frac{\mu_0}{4\pi}\frac{I\,\mathbf{d\vec{\ell}} \times \vec{r}}{r^3}$$

we obtain:

$$\vec{\mathbf{dB}} = \frac{\mu_o I}{4\pi} \frac{dz' \hat{\mathbf{k}} \times \left(x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + (z - z')\hat{\mathbf{k}} \right)}{\left[x^2 + y^2 + (z - z')^2 \right]^{3/2}}$$

$$\vec{\mathbf{dB}} = \frac{\mu_o I}{4\pi} \frac{dz' \left(x\hat{\mathbf{k}} \times \hat{\mathbf{i}} + y\hat{\mathbf{k}} \times \hat{\mathbf{j}} + (z - z')\hat{\mathbf{k}} \times \hat{\mathbf{k}} \right)}{\left[x^2 + y^2 + (z - z')^2 \right]^{3/2}}$$

$$\vec{\mathbf{dB}} = \frac{\mu_o I}{4\pi} \frac{dz' \left(x\hat{\mathbf{j}} - y\hat{\mathbf{i}} \right)}{\left[x^2 + y^2 + (z - z')^2 \right]^{3/2}}$$

$$\vec{\mathbf{dB}} = -\frac{\mu_o I}{4\pi} y \frac{dz'}{\left[x^2 + y^2 + (z - z')^2 \right]^{3/2}} \hat{\mathbf{i}} + \frac{\mu_o I}{4\pi} x \frac{dz'}{\left[x^2 + y^2 + (z - z')^2 \right]^{3/2}} \hat{\mathbf{j}}$$

Let's work on this a component at a time. For the x component, we have:

$$dB_x = -\frac{\mu_o I}{4\pi} y \frac{dz'}{\left[x^2 + y^2 + (z - z')^2 \right]^{3/2}}$$

Integrating over z' from $-\infty$ to ∞ yields:

$$B_x = -\frac{\mu_o I}{4\pi} y \int_{-\infty}^{\infty} \frac{dz'}{\left[x^2 + y^2 + (z - z')^2 \right]^{3/2}}$$

I'm going to go with the following variable substitution:

$u = z - z'$

$du = -dz'$, so, $dz' = -du$

Upper Limit: Evaluating $u = z - z'$ at $z' = \infty$ yields $-\infty$ for the upper limit of integration.

Lower Limit: Evaluating $u = z - z'$ at $z' = -\infty$ yields ∞ for the lower limit of integration.

So, our integral becomes:

$$B_x = -\frac{\mu_o I}{4\pi} y \int_{\infty}^{-\infty} \frac{-du}{\left(x^2 + y^2 + u^2 \right)^{3/2}}$$

I choose to use one of the minus signs to interchange the limits of integration:

$$B_x = -\frac{\mu_o I}{4\pi} y \int_{-\infty}^{\infty} \frac{du}{\left(x^2 + y^2 + u^2 \right)^{3/2}}$$

Using $\int \frac{dx}{(x^2+a^2)^{3/2}} = \frac{1}{a^2}\frac{x}{\sqrt{x^2+a^2}}$ from your formula sheet; and; identifying x^2+y^2 as a^2, and, u as the x appearing on the formula sheet, we obtain:

$$B_x = -\frac{\mu_o I}{4\pi} y \frac{1}{x^2+y^2}\frac{u}{\sqrt{u^2+x^2+y^2}}\Bigg|_{-\infty}^{\infty}$$

Now, I need to take the limit of that expression as u goes to ∞ and again as u goes to $-\infty$. To facilitate that, I want to factor a u out of the square root in the denominator. But, I have to be careful. The expression $\sqrt{u^2+x^2+y^2}$, which is equivalent to $\sqrt{(z-z')^2+x^2+y^2}$ is a distance. That means it is inherently positive, whether u (or z' for that matter) is positive or negative. So, when I factor u out of the square root, I'm going to have to use absolute value signs. For the denominator: $\sqrt{u^2+x^2+y^2} = \sqrt{u^2\left(1+\frac{x^2}{u^2}+\frac{y^2}{u^2}\right)} = |u|\sqrt{1+\frac{x^2}{u^2}+\frac{y^2}{u^2}}$, so,

$$B_x = -\frac{\mu_o I}{4\pi} y \frac{1}{x^2+y^2}\frac{u}{|u|}\frac{1}{\sqrt{1+\frac{x^2}{u^2}+\frac{y^2}{u^2}}}\Bigg|_{-\infty}^{\infty}$$

$$B_x = -\frac{\mu_o I}{4\pi} y \frac{1}{x^2+y^2}\left(1\frac{1}{\sqrt{1+0+0}} - -1\frac{1}{\sqrt{1+0+0}}\right)$$

$$B_x = -\frac{\mu_o I}{4\pi} y \frac{1}{x^2+y^2}(2)$$

$$B_x = -\frac{\mu_o I}{2\pi} y \frac{1}{x^2+y^2}$$

Now for the y component. Recall that we had:

$$\vec{\mathbf{dB}} = -\frac{\mu_o I}{4\pi} y \frac{dz'}{\left[x^2+y^2+(z-z')^2\right]^{3/2}}\hat{\mathbf{i}} + \frac{\mu_o I}{4\pi} x \frac{dz'}{\left[x^2+y^2+(z-z')^2\right]^{3/2}}\hat{\mathbf{j}}$$

so,

$$dB_y = \frac{\mu_o I}{4\pi} x \frac{dz'}{\left[x^2+y^2+(z-z')^2\right]^{3/2}}$$

But, except for the replacement of $-y$ by x, this is the same expression that we had for dB_x. And those, (the $-y$ in the expression for dB_x and the x in the expression for dB_y), are, as far as the integration over z' goes, constants, out front. They don't affect the integration, they just "go along for the ride." So, we can use our B_x result for B_y if we just replace the $-y$, in our expression for B_x, with x. In other words, without having to go through the entire integration process again, we have:

$$B_y = \frac{\mu_o I}{2\pi} x \frac{1}{x^2 + y^2}$$

Since we have no z component in our expression

$$\vec{\mathbf{dB}} = -\frac{\mu_o I}{4\pi} y \frac{dz'}{\left[x^2 + y^2 + (z - z')^2\right]^{3/2}} \hat{\mathbf{\imath}} + \frac{\mu_o I}{4\pi} x \frac{dz'}{\left[x^2 + y^2 + (z - z')^2\right]^{3/2}} \hat{\mathbf{\jmath}} \ ,$$

$\vec{\mathbf{B}}$ itself must have no z component.

Substituting our results for B_x, B_y, and B_z into the $\hat{\mathbf{\imath}}$, $\hat{\mathbf{\jmath}}$, $\hat{\mathbf{k}}$ expression for $\vec{\mathbf{B}}$, (Namely, $\vec{\mathbf{B}} = B_x\hat{\mathbf{\imath}} + B_y\hat{\mathbf{\jmath}} + B_z\hat{\mathbf{k}}$), we have:

$$\vec{\mathbf{B}} = -\frac{\mu_o I}{2\pi} y \frac{1}{x^2 + y^2} \hat{\mathbf{\imath}} + \frac{\mu_o I}{2\pi} x \frac{1}{x^2 + y^2} \hat{\mathbf{\jmath}} + 0\hat{\mathbf{k}}$$

$$\vec{\mathbf{B}} = \frac{\mu_o I}{2\pi} \frac{1}{x^2 + y^2} (-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}})$$

The quantity $x^2 + y^2$ is just r^2, the square of the distance that point P is from the current-carrying wire (recall that we are finding the magnetic field due to a wire, with a current I, that extends along the z axis from $-\infty$ to $+\infty$)

$$\vec{\mathbf{B}} = \frac{\mu_o I}{2\pi} \frac{1}{r^2} (-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}})$$

Furthermore, the vector $(-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}})$ has magnitude $\sqrt{(-y)^2 + x^2} = \sqrt{x^2 + y^2} = r$. Hence, the unit vector $\hat{\mathbf{u}}_B$ in the same direction as $(-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}})$ is given by

$$\hat{\mathbf{u}}_B = \frac{-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}}}{r} = -\frac{y}{r}\hat{\mathbf{\imath}} + \frac{x}{r}\hat{\mathbf{\jmath}}$$

and, expressed as its magnitude times the unit vector in its direction, the vector $(-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}})$ can be written as:

$$(-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}}) = r\hat{\mathbf{u}}_B$$

Substituting $(-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}}) = r\hat{\mathbf{u}}_B$ into our expression $\vec{\mathbf{B}} = \frac{\mu_o I}{2\pi}\frac{1}{r^2}(-y\hat{\mathbf{\imath}} + x\hat{\mathbf{\jmath}})$ yields:

$$\vec{\mathbf{B}} = \frac{\mu_o I}{2\pi}\frac{1}{r^2} r\hat{\mathbf{u}}_B$$

$$\vec{\mathbf{B}} = \frac{\mu_o I}{2\pi}\frac{1}{r}\hat{\mathbf{u}}_B$$

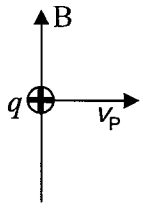
Note that the magnitude of $\vec{\mathbf{B}}$ obtained here, namely $B = \frac{\mu_o I}{2\pi}\frac{1}{r}$, is identical to the magnitude obtained using the integral form of Ampere's Law. The direction $\hat{\mathbf{u}}_B = -\frac{y}{r}\hat{\mathbf{\imath}} + \frac{x}{r}\hat{\mathbf{\jmath}}$ for the magnetic field at any point P having coordinates (x, y, z), is also the same as, "the magnetic field extends in circles about that wire, in that sense of rotation (counterclockwise or clockwise) which is consistent with the right hand rule for something curly something straight with the something straight being the current and the something curly being the magnetic field."

37 Maxwell's Equations

In this chapter, the plan is to summarize much of what we know about electricity and magnetism in a manner similar to the way in which James Clerk Maxwell summarized what was known about electricity and magnetism near the end of the nineteenth century. Maxwell not only organized and summarized what was known, but he added to the knowledge. From his work, we have a set of equations known as Maxwell's Equations. His work culminated in the discovery that light is electromagnetic waves.

In building up to a presentation of Maxwell's Equations, I first want to revisit ideas we encountered in chapter 20 and I want to start that revisit by introducing an easy way of relating the direction in which light is traveling to the directions of the electric and magnetic fields that are the light.

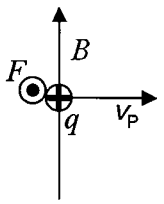
Recall the idea that a charged particle moving in a stationary magnetic field



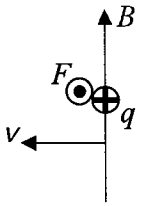
experiences a force given by

$$\vec{\mathbf{F}} = q\,\vec{\mathbf{v}}_P \times \vec{\mathbf{B}}$$

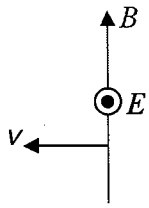
This force, by the way, is called the *Lorentz Force*. For the case depicted above, by the right-hand rule for the cross product of two vectors, this force would be directed out of the page.



Viewing the exact same situation from the reference frame in which the charged particle is at rest we see a magnetic field moving sideways (with velocity $\vec{v} = -\vec{v}_p$) through the particle. Since we have changed nothing but our viewpoint, the particle is experiencing the same force.



We introduce a "middleman" by adopting the attitude that the moving magnetic field doesn't really exert a force on the charged particle, rather it causes an electric field which does that. For the force to be accounted for by this middleman electric field, the latter must be in the direction of the force. The existence of light indicates that the electric field is caused to exist whether or not there is a charged particle for it to exert a force on.



The bottom line is that wherever you have a magnetic field vector moving sideways through space you have an electric field vector, and, the direction of the velocity of the magnetic field vector is consistent with

$$\textit{direction of } \vec{v} = \textit{direction of } \vec{\mathbf{E}} \times \vec{\mathbf{B}}.$$

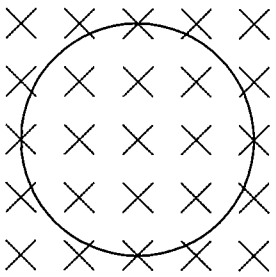
You arrive at the same result for the case of an electric field moving sideways through space. (Recall that in chapter 20, we discussed the fact that an electric field moving sideways through space causes a magnetic field.)

The purpose of this brief review of material from chapter 20 was to arrive at the result $\textit{direction of } \vec{v} = \textit{direction of } \vec{\mathbf{E}} \times \vec{\mathbf{B}}$. This direction relation will come in handy in our discussion of two of the four equations known as Maxwell's Equations.

One of Maxwell's Equations is called Faraday's Law. It brings together a couple of things we have already talked about, namely, the idea that a changing number of magnetic field lines through a loop or a coil induces a current in that loop or coil, and, the idea that a magnetic field vector that is moving sideways through a point in space causes an electric field to exist at that point in space. The former is a manifestation of the latter. For instance, suppose you have an increasing number of downward directed magnetic field lines through a horizontal loop. The idea is that for the number of magnetic field lines through the loop to be increasing, there must be magnetic field lines moving sideways through the conducting material of the loop (to get inside the perimeter of the loop). This causes an electric field in the conducting material of the

loop which in turn pushes on the charged particles of the conducting material of the loop and thus results in a current in the loop. We can discuss the production of the electric field at points in space occupied by the conducting loop even if the conducting loop is not there. If we consider an imaginary loop in its place, the magnetic field lines moving through it to the interior of the loop still produce an electric field in the loop; there are simply no charges for that field to push around the loop.

Suppose we have an increasing number of downward-directed magnetic field lines through an imaginary loop. Viewed from above the situation appears as:

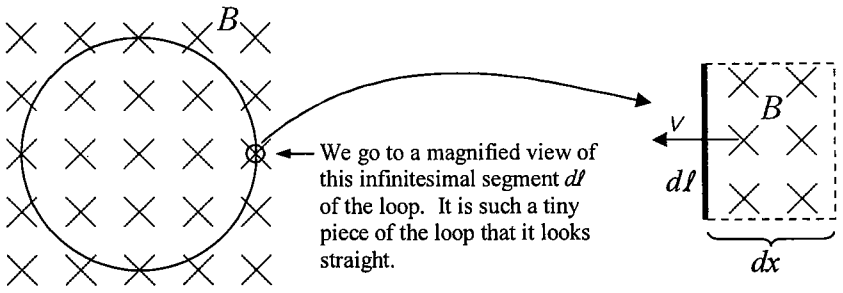


The big idea here is that you can't have an increasing number of downward-directed magnetic field lines through the region encircled by the imaginary loop without having, either, downward-directed magnetic field lines moving transversely and *inward* through the loop into the region encircled by the loop, or, upward-directed magnetic field lines moving transversely and *outward* through the loop out of the region encircled by the loop. Either way you have magnetic field lines cutting through the loop and with each magnetic field cutting through the loop there has to be an associated *electric* field with a component tangent to the loop. Our technical expression for the "number of magnetic field lines through the loop" is the magnetic flux, given, in the case of a uniform (but time-varying) magnetic field by

$$\Phi_B = \vec{\mathbf{B}} \cdot \vec{\mathbf{A}} \tag{37-3}$$

where A is the area of the region encircled by the loop.

Faraday's Law, as it appears in Maxwell's Equations, is a relation between the rate of change of the magnetic flux through the loop and the electric field (produced by this changing flux) in the loop. To arrive at it, we consider an infinitesimal segment $d\ell$ of the loop and the infinitesimal contribution to the rate of change of the magnetic flux through the loop resulting from magnetic field lines moving through that segment $d\ell$ into the region encircled by the loop.



If the magnetic field depicted above is moving sideways toward the interior of the loop with a speed $v = \frac{dx}{dt}$ then all the magnetic field lines in the region of area $A = d\ell dx$, will, in time dt, move leftward a distance dx. That is, they will all move from outside the loop to inside the loop creating a change of flux, in time dt, of

$$d\phi_B = BdA$$

$$d\phi_B = Bd\ell dx$$

Now, if I divide both sides of this equation by the time dt in which the change occurs, we have

$$\frac{d\phi_B}{dt} = Bd\ell \frac{dx}{dt}$$

which I can write as

$$\dot{\phi}_B = Bd\ell v$$

or

$$\dot{\phi}_B = vBd\ell \qquad 37\text{-}1$$

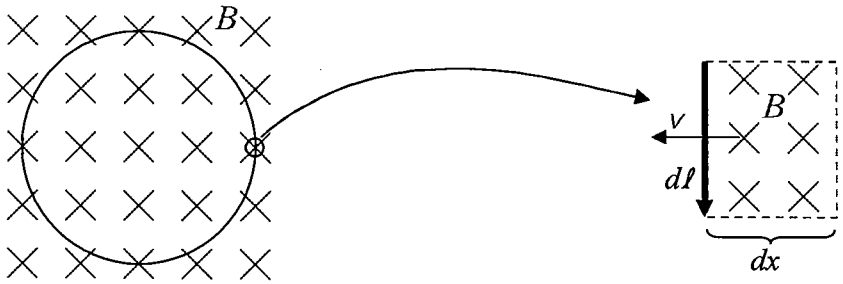
For the case at hand, looking at the diagram, we see that $\vec{\mathbf{B}}$ and $\vec{\mathbf{v}}$ are at right angles to each other so the magnitude of $\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ is just vB. In that case, since $\vec{\mathbf{E}} = -\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ (from equation 20-1 with $-\vec{\mathbf{v}}$ in place of $\vec{\mathbf{v}}_p$), we have $E = vB$. Replacing the product vB appearing on the right side of equation 37-1 ($\dot{\phi}_B = vBd\ell$) yields $\dot{\phi}_B = Ed\ell$ which I copy at the top of the following page:

$$\dot{\phi}_B = Ed\ell$$

We can generalize this to the case where the velocity vector $\vec{v}$ is not perpendicular to the infinitesimal loop segment in which case $\vec{E}$ is not along $d\vec{\ell}$. In that case the component of $\vec{E}$ that is along $d\vec{\ell}$, times the length $d\ell$ itself, is just $\vec{E}\cdot d\vec{\ell}$ and our equation becomes

$$\dot{\phi}_B = -\vec{E}\cdot d\vec{\ell}$$

In this expression, the direction of $d\vec{\ell}$ is determined once one decides on which of the two directions in which a magnetic field line can extend through the region enclosed by the loop is defined to make a positive contribution to the flux through the loop. The direction of $d\vec{\ell}$ is then the one which relates the sense in which $d\vec{\ell}$ points around the loop, to the positive direction for magnetic field lines through the loop, by the right hand rule for something curly something straight. With this convention the minus sign is needed to make the dot product have the same sign as the sign of the ongoing change in flux. Consider for instance the case depicted in the diagram:



We are looking at a horizontal loop from above. Downward is depicted as into the page. Calling downward the positive direction for flux makes clockwise, as viewed from above, the positive sense for the $d\vec{\ell}$'s in the loop, meaning the $d\vec{\ell}$ on the right side of the loop is pointing toward the bottom of the page (as depicted). For a downward-directed magnetic field moving leftward into the loop, $\vec{E}$ must be directed toward the top of the page (from *direction of* $\vec{v}$ = *direction of* $\vec{E}\times\vec{B}$). Since $\vec{E}$ is in the opposite direction to that of $d\vec{\ell}$, $\vec{E}\cdot d\vec{\ell}$ must be negative. But movement of downward-directed magnetic field lines into the region encircled by the loop, what with downward being considered the positive direction for flux, means a positive rate of change of flux. The left side of $\dot{\phi}_B = -\vec{E}\cdot d\vec{\ell}$ is thus positive. With $\vec{E}\cdot d\vec{\ell}$ being negative, we need the minus sign in front of it to make the right side positive too.

Now $\dot{\phi}_B$ is the rate of change of magnetic flux through the region encircled by the loop due to the magnetic field lines that are entering that region through the one infinitesimal $d\vec{\ell}$ that we

have been considering. There is a $\dot{\phi}_B$ for each infinitesimal $\vec{d\ell}$ making up the loop. Thus there are an infinite number of them. Call the infinite sum of all the $\dot{\phi}_B$'s $\dot{\Phi}_B$ and our equation becomes:

$$\dot{\Phi}_B = -\oint \vec{E} \cdot \vec{d\ell}$$

which is typically written

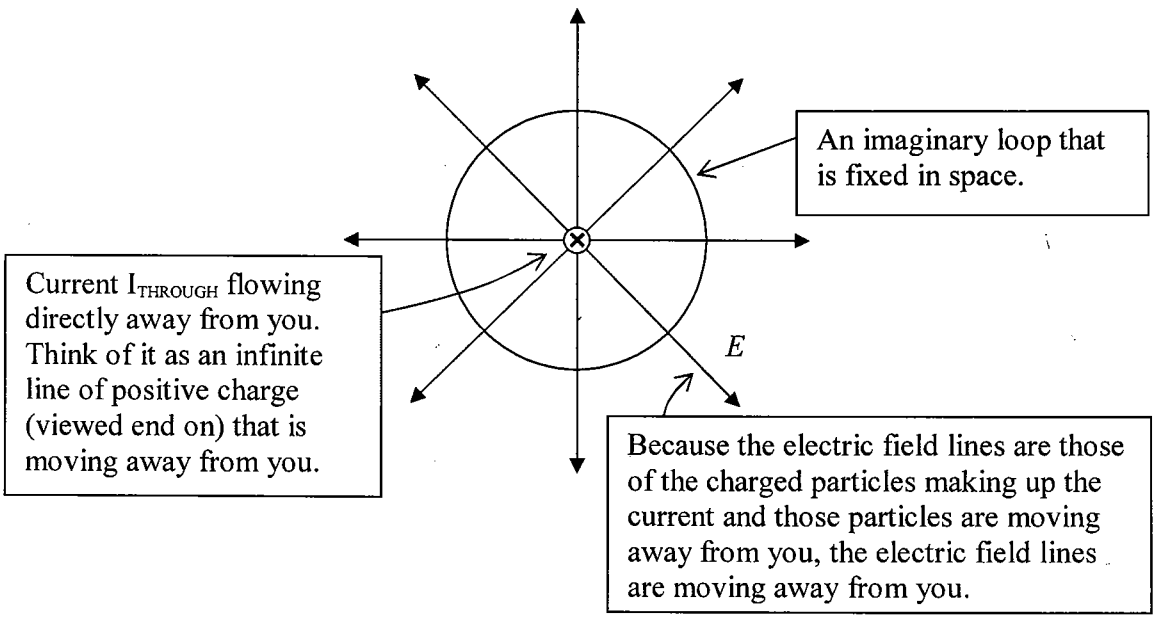
$$\oint \vec{E} \cdot \vec{d\ell} = -\dot{\Phi}_B$$

The integral is called a line integral because we integrate over a curve, namely the curve that is the loop, and the circle on the integral sign indicates that the curve in question is closed. The relation is called Faraday's Law and is one of Maxwell's equations.

The same kinds of considerations for the case of an electric field moving sideways through the perimeter of an imaginary loop leads to

$$\oint \vec{B} \cdot \vec{d\ell} = \mu_0 \epsilon_0 \dot{\Phi}_E$$

This is our loop form of the idea that an electric field moving sideways through an empty point in space causes a magnetic field to exist at that point in space. It is an incomplete version of one of Maxwell's Equations and, as it stands, is known as Maxwell's Extension to Ampere's Law. The thing is, unlike the case in which magnetic fields move sideways through the perimeter of a loop, there is a way in which electric fields can move sideways through the perimeter of a loop, without there being a change in the number of electric field lines through the region enclosed by the loop. In fact this happens whenever there is an electric current through the region enclosed by the loop. A downward current through a horizontal loop can be modeled as a vertical infinite line of positive charge (perhaps in a stationary sheath of negative charge, also infinite in length, so the overall charge of any length of the combination is zero) moving downward.



The moving charge (the current) causes electric field lines to be moving transversely through the perimeter of the loop thus causing a magnetic field in that perimeter, which, for the case depicted is clockwise as viewed from above. We typically leave out the middleman electric field when discussing this effect and say that a current through the area enclosed by a loop causes a magnetic field in that loop and call the phenomenon Ampere's Law. Careful analysis (that is not particularly difficult but I want to shorten this discussion) of the phenomenon allows us to write Ampere's Law in the form

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_0 I_{\text{THROUGH}}$$

Note that the left side is the same as the left side of what we called Maxwell's extension to Ampere's Law integral $\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_0 \epsilon_0 \dot{\Phi}_E$. Maxwell's extension covers the case in which there is a changing number of electric field lines through the region enclosed by the loop but no current through that region. Ampere's law covers the case in which there is a current through the region enclosed by the loop but no changing number of electric field lines through that region. If we have both a current and a changing number of electric field lines through a loop, then we have to add the two contributions to the magnetic field in the perimeter of the loop. This results in the equation

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_0 I_{\text{THROUGH}} + \mu_0 \epsilon_0 \dot{\Phi}_E$$

which is Ampere's Law with Maxwell's Extension. It is one of the four equations known as Maxwell's Equations.

So far in this chapter we have discussed two of the four equations known as Maxwell's Equations. The other two (both of which we have already encountered) are Gauss's Law for the electric field:

$$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$$

and Gauss's La' w for the magnetic field:

$$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}} = 0$$

That's all fou r of Maxwell's Equations. Here we list them in tabular form with the correspondin g name and conceptual statement beside each one:

Maxw ell's Equations	Name and Corresponding Conceptual Statement
$\oint \vec{\mathbf{E}} \cdot \vec{d}\vec{\mathbf{A}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$	**Gauss's Law for the Electric Field**—essentially a revised form of Coulomb's Law. It states that a charged particle or a distribution of charge causes an electric field.
$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}} = 0$	**Gauss's Law for the Magnetic Field**. It states that there is no such thing as a magnetic monopole.
$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dl}} = -\dot{\Phi}_B$	**Faraday's Law**. It states that a changing magnetic field causes an electric field.
$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_o I_{\text{THROUGH}} + \mu_0 \epsilon_0 \dot{\Phi}_E$	**Ampere's Law with Maxwell's Extension**. It states that a current causes a magnetic field, and, that a changing electric field causes a magnetic field.

The form of each of the equations given above is referred to as the integral form of the corresponding Maxwell's Equation. If for each equation, one takes the limit as the closed surface or loop becomes infinitesimal in size, one arrives at the differential form of the corresponding Maxwell's Equation. In differential form the equations are expressed in terms of the vector differential operators, the divergence $\nabla \cdot$ ("delta dot") and the curl $\nabla \times$ ("delta cross"). You aren't required to know what these operators mean or how to use them, but, you are required to be able to recognize Maxwell's Equations when you see them in differential form, to be able to associate each one with the corresponding integral form of the equation, and to be able to associate each one with the corresponding name and conceptual statement. The following two tables will help you meet these requirements:

Integral Form of Maxwell's Equations	Differential Form of Maxwell's Equations
$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dA}} = \frac{Q_{\text{ENCLOSED}}}{\epsilon_0}$	$\nabla \cdot \vec{\mathbf{E}} = \rho / \epsilon_0$
$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dA}} = 0$	$\nabla \cdot \vec{\mathbf{B}} = 0$
$\oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dl}} = -\dot{\Phi}_{\text{B}}$	$\nabla \times \vec{\mathbf{E}} = -\frac{d\vec{\mathbf{B}}}{dt}$
$\oint \vec{\mathbf{B}} \cdot \vec{\mathbf{dl}} = \mu_0 I_{\text{THROUGH}} + \mu_0 \epsilon_0 \dot{\Phi}_{\text{E}}$	$\nabla \times \vec{\mathbf{B}} = \mu_0 \vec{\mathbf{J}} + \mu_0 \epsilon_0 \frac{d\vec{\mathbf{E}}}{dt}$

The symbol ρ represents the charge density and $\vec{\mathbf{J}}$ represents the current per area.

Differential Form of Maxwell's Equations	Name and Corresponding Conceptual Statement
$\nabla \cdot \vec{\mathbf{E}} = \rho / \epsilon_0$	**Gauss's Law for the Electric Field**—essentially a revised form of Coulomb's Law. It states that a charged particle or a distribution of charge causes an electric field.
$\nabla \cdot \vec{\mathbf{B}} = 0$	**Gauss's Law for the Magnetic Field.** It states that there is no such thing as a magnetic monopole.
$\nabla \times \vec{\mathbf{E}} = -\frac{d\vec{\mathbf{B}}}{dt}$	**Faraday's Law.** It states that a changing magnetic field causes an electric field.
$\nabla \times \vec{\mathbf{B}} = \mu_0 \vec{\mathbf{J}} + \mu_0 \epsilon_0 \frac{d\vec{\mathbf{E}}}{dt}$	**Ampere's Law with Maxwell's Extension.** It states that a current causes a magnetic field, and, that a changing electric field causes a magnetic field.

EPCOS

Electronic Parts and Components

Dated: 28.08.2008

EPCOS India Private Limited
Plot No. E 22-25, MIDC, Satpur
Nashik-422 007 (INDIA)
Tel. : +91 (253) 2353756 TO 60
Fax : +91 (253) 2353761

Registered Office :
Kulia Kanchrapara Road
P.O. Netaji Subhas Sanatorium
Kalyani, Dist. Nadia
West Bengal-741251, (INDIA)