General Physics 2
PHY 2049

Yaouen Fily

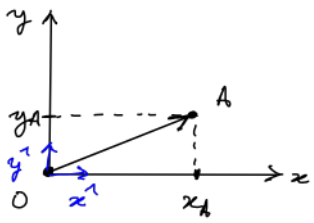
August 10, 2021 3:40am

Contents

Chapter 1: Mathematical Tools

1.1 Vector Algebra

1.1.1 Position vectors



The position vector of point A, often written $\vec{r}_A$, is the vector going from the origin O to point A.

$$\vec{r}_A \equiv \overrightarrow{OA}$$

The position vector is related to the point's cartesian coordinates by

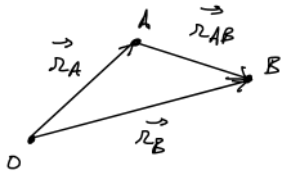
$$\vec{r}_A = x_A\hat{x} + y_A\hat{y}$$

where $\hat{x}$ and $\hat{y}$ are the unit vectors along the x and y axes respectively. Graphically, to get from O to A you need to move by x_A along $\hat{x}$ and by y_A along $\hat{y}$.

Vectors can also be written with one coordinate above the other between either square brackets or parentheses:

$$\vec{r}_A = \begin{bmatrix} x_A \\ y_A \end{bmatrix} = \begin{pmatrix} x_A \\ y_A \end{pmatrix}$$

1.1.2 Relative position



The position vector of a point B relative to another point A is the vector going from A to B. One can go from A to B by going from A to O then from O to B, therefore

$$\vec{r}_{AB} = \overrightarrow{A0} + \overrightarrow{OB} = \overrightarrow{OB} - \overrightarrow{OA} = \vec{r}_B - \vec{r}_A$$

In terms of coordinates:

$$\vec{r}_{AB} = \begin{bmatrix} x_B \\ y_B \end{bmatrix} - \begin{bmatrix} x_A \\ y_A \end{bmatrix} = \begin{bmatrix} x_B - x_A \\ y_B - y_A \end{bmatrix} = (x_B - x_A)\hat{x} + (y_B - y_A)\hat{y}$$

When the coordinates of the relative position show up multiple times in a calculation it's common to define $x_{AB} = x_B - x_A$ and $y_{AB} = y_B - y_A$ such that $\vec{r}_{AB} = x_{AB}\hat{x} + y_{AB}\hat{y}$.

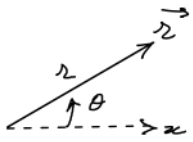
1.1.3 Distance

The distance between two points A and B, often noted AB or r_{AB}, is the length of the vector joining them.

$$r_{AB} = \|\vec{r}_{AB}\| = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

It doesn't matter which point comes first: $(x_A - x_B)^2 = (x_B - x_A)^2$ and $(y_A - y_B)^2 = (y_B - y_A)^2$ therefore $r_{AB} = r_{BA}$.

1.1.4 From distance and angle to coordinates

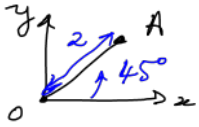


The coordinates of a vector $\vec{r}$ with length r making an angle θ with the (oriented) x axis are

$$\vec{r} = r\cos\theta\ \hat{x} + r\sin\theta\ \hat{y}$$

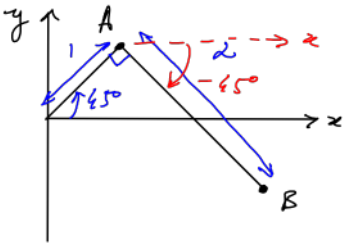
Examples:

- Compute the coordinates of A:



$$\vec{r}_A = 2\left[\cos\left(\frac{\pi}{4}\right)\hat{x} + \sin\left(\frac{\pi}{4}\right)\hat{y}\right] \approx 1.41\hat{x} + 1.41\hat{y}$$

- Compute the coordinates of B:



$$\vec{r}_A = \cos\left(\frac{\pi}{4}\right)\hat{x} + \sin\left(\frac{\pi}{4}\right)\hat{y} \approx 0.71\hat{x} + 0.71\hat{y}$$
$$\vec{r}_{AB} = 2\left[\cos\left(-\frac{\pi}{4}\right)\hat{x} + \sin\left(-\frac{\pi}{4}\right)\hat{y}\right] \approx 1.41\hat{x} - 1.41\hat{y}$$
$$\vec{r}_B = \vec{r}_A + \vec{r}_{AB} \approx 2.12\hat{x} - 0.71\hat{y}$$

1.1.5 Dot product

Let $\vec{a} = a_x\hat{x} + a_y\hat{y}$ and $\vec{b} = b_x\hat{x} + b_y\hat{y}$. The dot product of $\vec{a}$ and $\vec{b}$ is

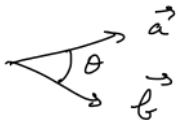
$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y$$

It's convenient to do it in column notation:

$$\vec{a} = \begin{bmatrix} a_x \\ a_y \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} b_x \\ b_y \end{bmatrix}, \quad \vec{a} \cdot \vec{b} = \begin{bmatrix} a_x \\ a_y \end{bmatrix} \cdot \begin{bmatrix} b_x \\ b_y \end{bmatrix} = a_x b_x + a_y b_y$$

The dot product is also related to the angle θ between $\vec{a}$ and $\vec{b}$:

$$\vec{a} \cdot \vec{b} = |\vec{a}|\,|\vec{b}|\cos\theta$$



Chapter 2: Electric Force

2.1 Motivation

2.1.1 Fundamental forces

A fundamental idea in physics is that there is a fairly compact set of laws of nature that all other laws can be derived from. There is a small set of fundamental building bricks – fundamental particles – and a small number of ways those bricks can interact with each other. However, there are many different ways those bricks can assemble to form different materials, and many different ways those interactions can add up to create the forces we observe at our scale.

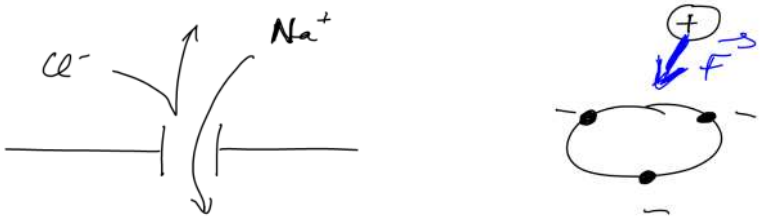
We've already encountered one fundamental force: gravity. It's perhaps the most obvious one, because it's always does the same thing: attract things together. The electric force, though, is by far the most ubiquitous. It binds the electrons to the nucleus in atoms. It is the dominant force at play in chemistry. It is behind virtually every force we discussed in physics 1 (other than gravity). Later in the class we'll talk about another fundamental force, the magnetic force. There are two more but they are largely irrelevant outside of nuclear reactions and we won't talk about them.

2.1.2 (Bio)chemistry

We're obviously not going to explain the entirety of chemistry in half a semester of talking about the electric force, but here are some of the problems we'll try to understand a little better by the end of this chapter.

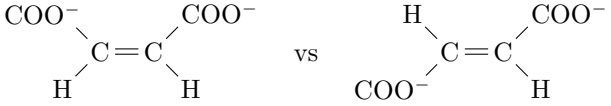
Ion channels

A ring of charge around a hole in the cell membrane can selectively allow anions or cations through. The laws of electrostatics allow us to estimate the force, or the energy, required for various ions to go through.



Shape of molecules and polymers

By comparing the electric energy of charged isomers we can partly explain their relative stability.

For example:  vs

Electric energy minimization also plays a key role in dictating the shape of molecules. In the VSEPR theory the repulsion between the molecule's valence electrons makes them spread out as much as possible.

Similar ideas apply to the shape of polymers, including how easy it is to bend them.

Electricity

Eventually we'll explain electricity in terms of electric forces on the charges moving through an electric conductor.

2.2 Electric field

2.2.1 Electric charge

Charge is a lot like mass. Every object has a mass. The mass of an object is the sum of the masses of its parts. This goes all the way down to particles: everything is made of elementary particles, every particle has a mass, and the mass of an object is the sum of the masses of its constituent particles (leaving aside relativistic effects).

Similarly, every particle has an electric charge and the charge of an object is the sum of the charges of its constituent particles. One big difference is that whereas masses are always positive, electric charges can be positive or negative. The most common symbol for an electric charge is q. Its unit is the Coulomb (symbol C).

Matter is made of atoms, and atoms are made of protons, neutrons, and electrons. The charge of a proton is $1.6 \times 10^{-19}\,\mathrm{C}$, often written e (unrelated to the e of exponentials). The charge of a neutron is $0\,\mathrm{C}$. The charge of an electron is $-1.6 \times 10^{-19}\,\mathrm{C} = -e$.

2.2.2 Electric force

Just like objects with mass exert gravitational forces on each other, electrically charged objects exert electric forces on each other. Consider two point-like charged objects, one with charge q_A located at point A and one with charge q_B located at point B. The force exerted by object A on object B is

$$\vec{F}_{A\to B} = k_e q_A q_B \frac{\overrightarrow{AB}}{AB^3}$$

where $k_e = 8.99 \times 10^9\,\mathrm{Nm^2C^{-2}}$ is Coulomb's constant and $AB = \|\overrightarrow{AB}\|$ is the distance between A and B. This is known as *Coulomb's law*.

Compare with the gravitational force exerted by A on B:

$$\vec{F}^{(G)}_{A\to B} = G m_A m_B \frac{\overrightarrow{AB}}{AB^3}$$

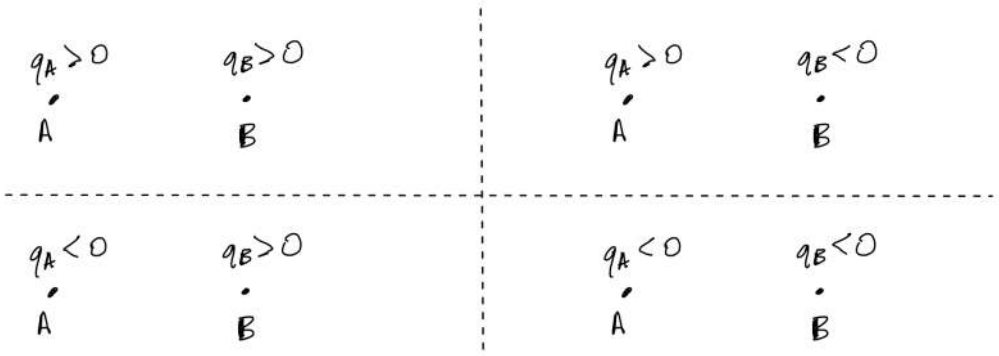
It's the same structure. k_e plays the same role as the gravitational constant G. The charges q_A and q_B play the same role as the masses m_A and m_B.

Whether it's the electric force or the graviational force, it's important to realize that A and B are just placeholders here. A is the location of the particle exerting the force. q_A is its charge. B is the location of the particle feeling the force. q_B is its charge. In applications those points may be called something else. A might even be called B, and B might be called A. To get it right every time, the formula should be learned this way:

$$\vec{F} = k_e \times \begin{pmatrix}\text{charge of the}\\ \text{object exerting}\\ \text{the force}\end{pmatrix} \times \begin{pmatrix}\text{charge of the}\\ \text{object feeling}\\ \text{the force}\end{pmatrix} \times \frac{\begin{pmatrix}\text{vector joining the object}\\ \text{exerting the force to the object}\\ \text{feeling the force}\end{pmatrix}}{\begin{pmatrix}\text{distance between the object}\\ \text{exerting the force and the}\\ \text{object feeling the force}\end{pmatrix}^3}$$

Problem 1: Direction of the electric force.

The rule is that charges with the same sign repel whereas charges with opposite signs attract. The sketch below shows the four possible cases. In each case, use the electric force formula to analyze the direction of the force on each particle and show that it does indeed follow the rule of thumb "like charges repel, opposite charges attract".



Problem 2: Two point charges.

Consider two point charges, one at point A with coordinates $(1\,\text{cm}, 1\,\text{cm})$ and charge $q_A = 1\,\mu\text{C}$, one at point B with coordinates $(2\,\text{cm}, 3\,\text{cm})$ and charge $q_B = 1\,\mu\text{C}$.

1. Sketch the system.
2. Compute $\overrightarrow{AB}$. Compute AB. Compute the force $\vec{F}_{A\to B}$ exerted by A on B.
3. Repeat question 2 for the force $\vec{F}_{B\to A}$ exerted by B on A.

Problem 3: Basic properties.

1. Compute the magnitude of the electric force between charge q_A at point A and charge q_B at point B as a function of k_e, q_A, q_B, and AB.
2. Show that the electric forces exerted by A and B on each other obey Newton's third law.

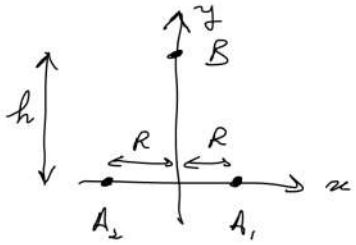
2.2.3 Superposition principle

The force exerted on a charge Q by a set of charges q_1,q_2,... is the sum of the forces exerted by each of q_1,q_2,... on Q. Let $\vec{r}_i$ be the vector going from charge q_i to charge Q, then the force on Q is

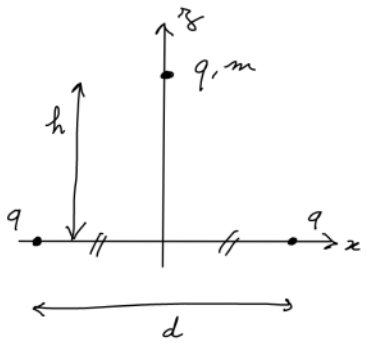
$$\vec{F} = \sum_i k_e q_i Q \frac{\vec{r}_i}{|\vec{r}_i|^3}$$

Problem 4: 2D charged "ring".

Compute the electric force $\vec{F}$ created by two charges q located at A_1 and A_2 on a charge Q located at B. Discuss the x component of $\vec{F}$.



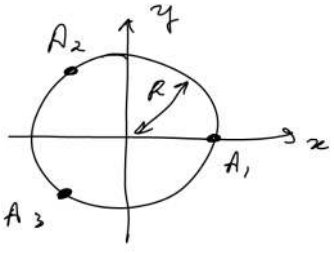
Problem 5: 2D levitation.

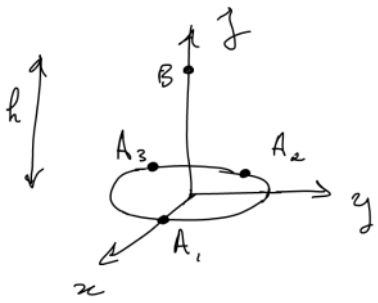


The bottom charges are fixed. The upper charge can move along the z axis. It is subject to its own weight (mass m, acceleration of gravity g) and the electric forces due to the bottom charges. All three charges have the same charge q. The upper charge is in mechanical equilibrium, i.e., levitating. What relationship must q, m, d, h, k_e, and g obey?

Problem 6: 3D charged ring.

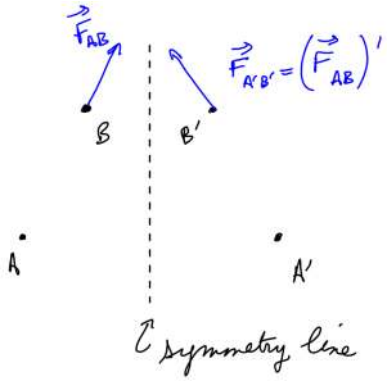
1. The three points A_1, A_2, A_3 are evenly spaced on a circle with radius R, center on the origin, in the xy plane. Compute the 3D coordinates of A_1, A_2, A_3.





2. B is located on the z axis at a distance h from the center of the ring. Compute $\overrightarrow{A_iB}$ for $i = 1, 2, 3$.
3. There's a charge q at each of A_1, A_2, A_3 and a charge Q at B. Compute the electric force on the latter.
4. Assume $q > 0$ and $Q > 0$. Is the general direction of the force on B consistent with "opposites attract/likes repel"? Why or why not? Same questions for $q > 0$ and $Q < 0$.

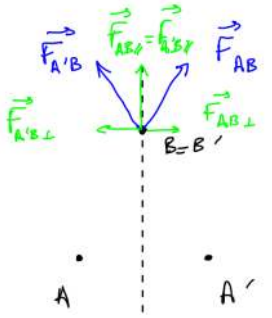
2.2.4 Mirror symmetry and the electric force



2D Mirror symmetry

A charge at A exerts a force $\vec{F}_{A\to B}$ on a second charge at B. If A' and B' are the mirror images of A and B with respect to the symmetry line, A' has the same charge as A, and B' has the same charge as B, then the force $\vec{F}_{A'\to B'}$ exerted by A' on B' is the mirror image of $\vec{F}_{A\to B}$ with respect to the same symmetry line.

The most common application of this is when B is on the symmetry line, i.e., B and B' are the same charge subject to a force from A and a force from A'. The symmetry argument above then tells us that the force $\vec{F}_{A\to B}$ exerted by A on B and the force $\vec{F}_{A'\to B}$ exerted by A' on B are mirror images of each other with respect to the symmetry line.



To understand the implications, we need to decompose each force into a part parallel to the symmetry line (subscript $||$) and a part perpendicular to the symmetry line (subscript $\perp$):

$$\vec{F}_{A\to B} = \vec{F}_{A\to B||} + \vec{F}_{A\to B\perp} \ , \quad \vec{F}_{A'\to B} = \vec{F}_{A'\to B||} + \vec{F}_{A'\to B\perp}$$

With those notations, $\vec{F}_{A\to B}$ and $\vec{F}_{A'\to B}$ being mirror images of each other means that $\vec{F}_{A\to B||} = \vec{F}_{A'\to B||}$ and $\vec{F}_{A\to B\perp} = -\vec{F}_{A'\to B\perp}$, i.e., they have the same parallel part but opposite perpendicular parts. It follows that the total force on B is

$$\vec{F}_B = \vec{F}_{A\to B} + \vec{F}_{A'\to B} = 2\vec{F}_{A\to B||}$$

It is parallel to the symmetry line, and it is equal to twice the parallel part of the force exerted by either A or A'.

This result greatly simplifies the computation of the force on B. Instead of having to compute every component of every force (parallel and perpendicular components of $\vec{F}_{A\to B}$ and $\vec{F}_{A'\to B}$), we only need to compute one, say $\vec{F}_{A\to B||}$.

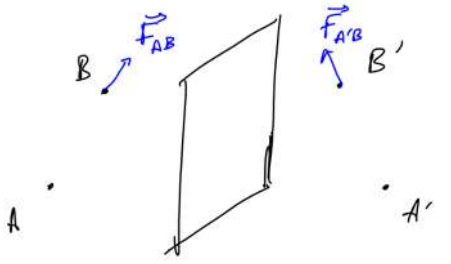
Furthermore, the result generalizes to any distribution of charges that is symmetric with respect to a line. If there is a line such that for every charge there is an equal charge mirroring it on the other side of the line, then the electric force on a charge located on the line is along the line.

Problem 7: Symmetric 2D charged "ring".

Redo problem 4 using a symmetry argument.

3D Mirror symmetry

The rule is the same except the symmetry line is now a symmetry plane. A' and B' are the mirror images of (and have the same charge as) A and B respectively, therefore $\vec{F}_{A'\to B'}$ is the mirror image of $\vec{F}_{A\to B}$ with respect to the symmetry line.



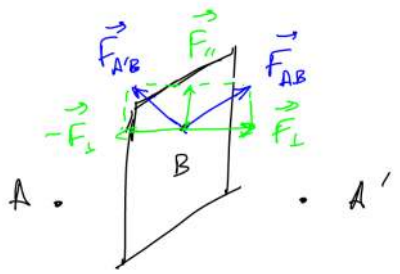
Again, if $B = B'$, i.e., if we're computing the force created by a distribution of charges that is symmetric with respect to the plane of symmetry on a charge located on the plane of symmetry, the electric force is along the plane.

Again, this comes from decomposing each force into a part parallel to the plane and one perpendicular to the plane:

$$\vec{F}_B = \vec{F}_{A\to B} + \vec{F}_{A'\to B} = \left(\vec{F}_{A\to B||} + \vec{F}_{A\to B\perp}\right) + \left(\vec{F}_{A'\to B||} + \vec{F}_{A'\to B\perp}\right)$$

The symmetry implies $\vec{F}_{A'\to B||} = \vec{F}_{A\to B||}$ and $\vec{F}_{A'\to B\perp} = -\vec{F}_{A\to B\perp}$, thus

$$\vec{F}_B = \left(\vec{F}_{A\to B||} + \vec{F}_{A\to B\perp}\right) + \left(\vec{F}_{A\to B||} - \vec{F}_{A\to B\perp}\right) = 2\vec{F}_{A\to B||}$$



Problem 8: Symmetric 3D charged ring.

Redo problem 6 using a symmetry argument.

2.2.5 Electric field

When computing the electric force exerted by a series of charges q_1,q_2,... located at A_1,A_2,... on another charge Q located at B, Q can be factored out:

$$\vec{F}_B = \sum_i k_e q_i Q \frac{\overrightarrow{A_iB}}{A_iB^3} = Q\left(\sum_i k_e q_i \frac{\overrightarrow{A_iB}}{A_iB^3}\right) \equiv Q\vec{E}_B \text{ where } \vec{E}_B = \sum_i k_e q_i \frac{\overrightarrow{A_iB}}{A_iB^3}$$

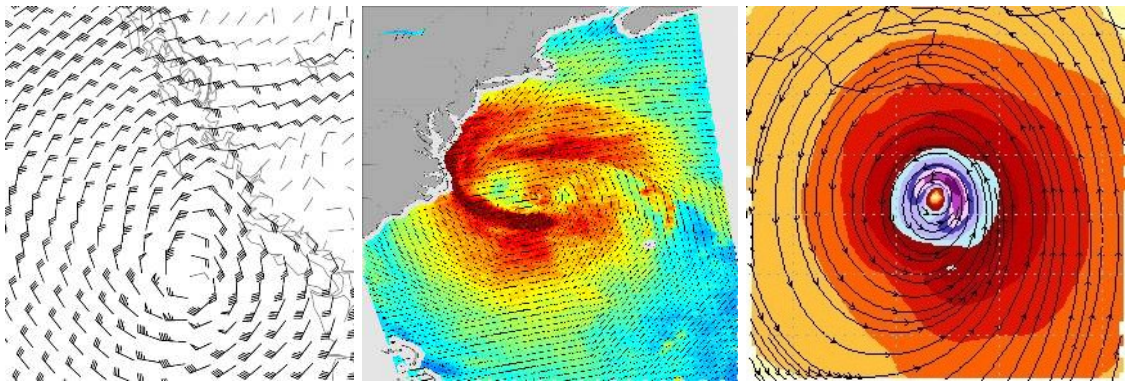
This defines the electric field $\vec{E}_B$ created at point B by the charges at A_1,A_2,... . It depends on the location of B, but not on Q, so it doesn't actually matter whether there is a charge Q at B or not. In fact, the electric field created by the charges at A_1,A_2,... can be computed anywhere and everywhere in the universe. Think of it as every point in the universe having a little arrow representing the electric field created there by the charges at A_1,A_2,... . That electric field in turn tells us about the electric force a hypothetical charge would feel if it was there.

Problem 9: Direction of the field vs direction of the force.

Redo problem 1, but this time also draw the direction of the electric field created at B by the charge at A and the direction of the electric field created at A by the charge at B.

2.2.6 Representing the electric field

The electric field is an example of vector field, i.e., a vector quantity that has a (potentially different) value at every point in the universe. The figure below shows a few ways to represent a more familiar vector field: the wind field.



The left picture is from a weather map and uses barbed wind arrows whose tail "feathers" code for the wind speed. The middle picture uses regular arrows for wind direction and color for wind speed while color code. The right picture shows stream lines, each of which represents the trajectory of something moving along with the wind. At every point the stream lines are tangent to the local wind's direction.

Both the arrow representation and the stream line representation are used for the electric field as well. The stream lines of the electric field are called electric field lines. They correspond to the trajectory of a particle that would always be following the direction of the electric field. Actual particles don't really follow the electric field (assuming there's no other force, the particle's acceleration, not its velocity, would be proportional to the electric field), but it can still help grasp field lines to think of them as a sort of trajectory. A more technical definition is that electric field lines are everywhere *tangent to the local electric field*. Also note that field lines have arrows showing which way the local electric field points. In technical terms, they are *oriented* curves.

To help you draw electric field lines, here are some rules they always follow:

1. They can only start at a positive charge or at infinity.

2. They can only end at a negative charge or at infinity.

3. They can only cross in one of three ways: (1) multiple lines ending at the same negative charge, (2) multiple lines starting at the same positive charge, (3) two lines going in and two lines coming out of a point where the electric field is zero.

4. The number of lines coming out of a positive charge or going into a negative charge should be proportional to the value of the charge. For example, two equal positive charges should have the same number of lines coming out of them, two opposite charges should have the same number of lines (coming out for the positive one, going in for the negative one), and a charge twice as strong as another should have twice as many lines coming out of going into it.

It may also help to think about electric field lines as the stream lines of an incompressible fluid, like water. In that case, positive charges correspond to places where fluid is being added (hence why the stream lines "flow out") and negative charges correspond to places where fluid is being removed (hence why the stream lines "flow in").

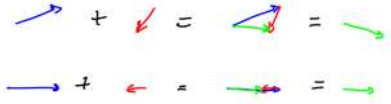
Problem 10: Electric field of a point charge.

Sketch the electric field around a single point charge q, first for $q > 0$, then for $q < 0$. Use a vector representation.

Problem 11: Electric field of a pair of point charges.

Sketch the electric field around a pair of point charges, first equal charges (q and q), then opposite charges with the same magnitude (q and $-q$).
Method: Choose a set of points at which to draw the field. At each point, sketch the field of each of the two charges, then use graphical vector addition to obtain the total field at that point. It helps to use three colors: one for the field of the first charge, one for the field of the second charge, and one for the total field.

Examples of graphical vector addition (in each example, the green arrow is the sum of the red arrow and the blue arrow):



Problem 12: Simple field lines.

Based on the vector representations you constructed in the previous problems and the idea that electric fields lines are everywhere tangent to the local electric field, sketch the field lines of each of those systems:

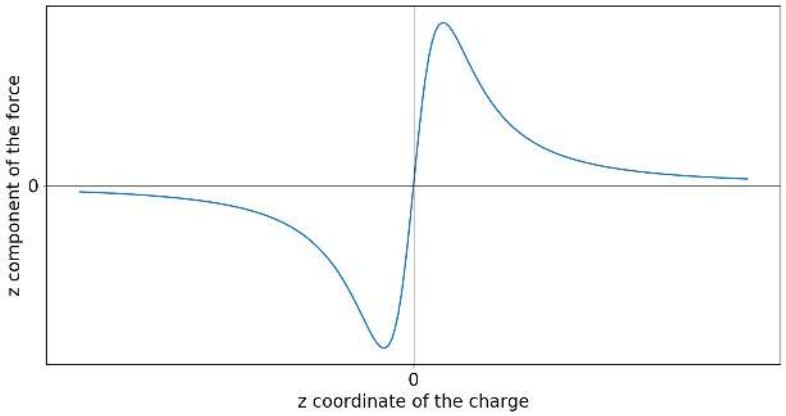
1. A positive point charge.
2. A negative point charge.
3. A pair of identical positive point charges.
4. A pair of opposite point charges.

2.2.7 Pushing through vs throwing through

Problem 13: Pushing a charge through a charged ring.

The system under study is the 3D ring of problem 6 in the case $Q = q$. Specifically, we are interested in the z component of the force exerted by the charges at A_1,A_2,A_3 on the charge at B. Let's call it F.

1. Write F as a function of the z coordinate of point B, which we'll call z.
2. The plot below shows that z component of the force as a function of the z coordinate of point B. Use the formula for F and/or symmetry arguments to explain the following features of the curve: (a) it goes through the origin, (b) it is negative on the left and positive on the right, and (c) it goes to zero far from the origin. Try to justify each feature in plain English.



3. Imagine the charge at B is pushed slowly from a position on the z axis above the ring ($z > 0$) all the way through the ring using a constant downward force F_0. Where does the ring offer the most resistance? What minimum magnitude must the downward force have to overcome the ring's force at that maximum resistance point?

Problem 14: Throwing a charge through a charged ring.

The charge at B starts at some initial distance h_0 above the ring with an initial downward speed v_0. How would you go about finding out whether it goes through? Write the equation that needs to be solved to predict the trajectory of the charge at B.

Towards potential energy

The differential equation that comes out of problem 14 is tricky to integrate, however it's possible to integrate it partially to prove the conservation of energy in this system, which turns out to be sufficient to

find out whether the charge goes through the ring. We're going to skip the integration and discuss how to apply the energy conservation method to this and a variety of other problems.

2.3 Electric potential energy

The electric force is conservative. That means that there is a electric potential energy U such that, when the only force at play is the electric force, the sum of the kinetic energy and the electric potential energy is conserved throughout whatever motion(s) the system go through.

2.3.1 Electric potential energy

The electric potential energy of two point charges q_A and q_B is

$$U_{AB} = \frac{k_e q_A q_B}{AB}$$

where AB is the distance between the two charges. When those two charges move under the sole influence of the electric forces they exert on each other, energy conservation reads:

$$\frac{1}{2}m_A v_A^2 + \frac{1}{2}m_B v_B^2 + \frac{k_e q_A q_B}{AB} = \text{constant}$$

where m_A is the mass of particle A, m_B is the mass of particle B, and v_B is the speed of particle B.

When there are more than two charges, the potential energy of the system is obtained by adding the potential energy of every pair of charges in the system. For example, if there are three charges at points A,B,C the electric potential energy is

$$U = U_{AB} + U_{AC} + U_{BC} = \frac{k_e q_A q_B}{AB} + \frac{k_e q_A q_C}{AC} + \frac{k_e q_B q_C}{BC}$$

Energy conservation then reads:

$$\frac{1}{2}m_A v_A^2 + \frac{1}{2}m_B v_B^2 + \frac{1}{2}m_C v_C^2 + \frac{k_e q_A q_B}{AB} + \frac{k_e q_A q_C}{AC} + \frac{k_e q_B q_C}{BC} = \text{constant}$$

The general formula is:

$$U = \sum_{\substack{\text{pairs of}\\ \text{charged}\\ \text{objects}}} \frac{k_e \begin{pmatrix} \text{charge of the} \\ \text{first object in} \\ \text{the pair} \end{pmatrix} \begin{pmatrix} \text{charge of the} \\ \text{second object} \\ \text{in the pair} \end{pmatrix}}{\begin{pmatrix} \text{distance} \\ \text{between the} \\ \text{two objects in} \\ \text{the pair} \end{pmatrix}}$$

If there are many charged points, a good way to list every pair is to:

1. Label the points from 1 to N (where N is the number of objects).
2. List the pairs whose first point is point 1: [1,2],[1,3],... until [1,N].
3. List the pairs whose first point is point 2 and whose second object is point 3 or more: [2,3],[2,4],... until [2,N]. Starting at 3 avoids the pair [2,1], which was already counted as [1,2] (the same two points in a different order).
4. Repeat with a new first point number all the way to N. Always make sure the number of the second point is larger than that of the first point so as to avoid recounting a pair that has already been counted.

This leads to the following formula for the total electric potential energy:

$$U = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{k_e q_i q_j}{A_i A_j}$$

where the points are called A_1,A_2,...A_N, their charges are called q_1,q_2,...q_N, and A_iA_j is the distance between A_i and A_j. Conservation of energy then reads:

$$\left(\sum_{i=1}^{N} \frac{1}{2} m_i v_i^2 \right) + \left(\sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{k_e q_i q_j}{A_i A_j} \right) = \text{constant}$$

where m_i and v_i are the mass and speed of point i.

In addition to the electric potential energy of a system, discussed above, we'll also talk about the electric potential energy of a charge. By that we mean that in the sum over the pairs of charges, only those pairs that involve the charge of interest should be considered. We'll see why that often makes sense in problem 16.

If there are other forces at play, but they are all conservative, then the total energy is conserved. It's given by the sum of the kinetic energy and the potential energies of all the conservative forces involved, including but not limited to the electric potential energy.

If there are additional nonconservative forces that do not work, then the total energy above is still conserved.

If there are additional nonconservative forces that do work, then the total energy above is not conserved and energy conservation arguments are unusable, or at least tricky to use correctly.

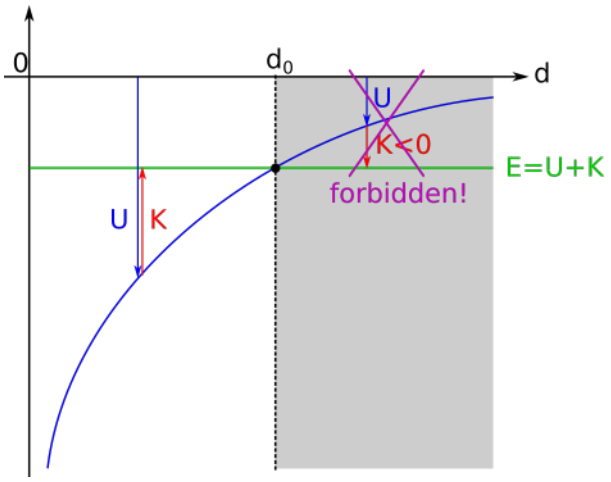
Problem 15: Collision speed.

Two point charges q and $-q$ with the same mass m start off at rest (initial speed zero) at a distance d_0 from each other.

1. Sketch the system.
2. Explain why the two charges are bound to collide.
3. By symmetry, the velocity vectors of the two charges are opposite of each other throughout their motion ($\vec{v}_2 = -\vec{v}_1$). In particular their speeds (the magnitude of their velocities) are the same: $v_2 = ||\vec{v}_2|| = ||-\vec{v}_1|| = ||\vec{v}_1|| = v_1$. Let's call that common speed v. Use the conservation of energy to write v as a function of the distance d between the charges as that distance changes.
4. What happens to v as the points near collision? Explain why this is consistent with what we know about the electric force between two point charges.
5. Instead of points, the charges are now spheres with the same radius R. The distance d is now measured from the center of one sphere to the center of the other. With this convention, the potential energy keeps the exact same form. The only difference is that the collision happens earlier, when the center-to-center distance d becomes small enough for the surfaces of the spheres to touch. What is v as they collide (or rather right before they collide; once they start actually colliding other forces come into play that we haven't accounted for)?
6. Rank those three scenarios from lowest collision speed to highest collision speed: (a) two spheres with radius R initially at rest at distance d_0, (b) two spheres with radius $R' = R/2$ initially at rest at distance d_0, (c) two spheres with radius R initially at rest at distance $d_0' = 2d_0$.

2.3.2 Interpretation of potential energy curves

Many consequences of energy conservation can be visualized on a plot of the potential energy. The one below corresponds to problem 15. The horizontal axis is the distance d between the two charges. The blue curve is the potential energy $U = -k_e q^2/d$. The kinetic energy is $K = mv^2$ ($\frac{1}{2}mv^2$ per charge). Energy being conserved means that the sum of the two is constant. Graphically, the sum of the blue arrow representing the potential energy (going from the horizontal axis to the blue curve; downwards means negative U) and the red arrow representing the kinetic energy always lands at the same value represented by the green horizontal curve. The height of the green curve is the total energy $E = U + K$, which is set by the initial condition. Here $d(0) = d_0$ and $v(0) = 0$, therefore $E = 0 + U(d_0) = -k_e q^2/d_0$, i.e., the green line intersects the blue line at $d = d_0$.

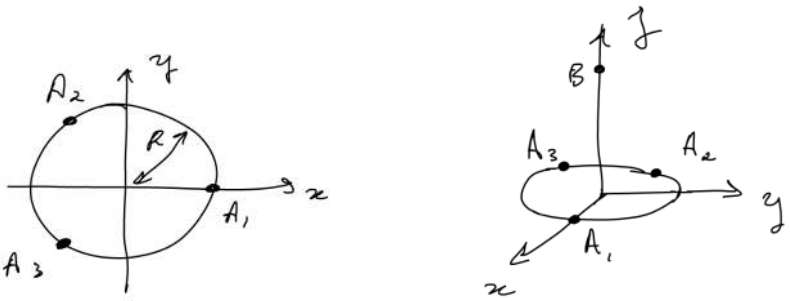


Since the kinetic energy is the gap between the blue curve and the green line, an intersection between the blue curve and the green line corresponds to $K = 0$, which means $v = 0$, i.e., (temporarily) stopped charges. When the speed is nonzero, the kinetic energy mv^2 is always positive, therefore $E = U + K > U$. Graphically, only the regions where the blue curve is below the green line are accessible. For the system to be in the gray region where $U > E$, the kinetic energy $K = E - U$ would need to be negative, which is not possible.

In the problem, the charges start at a distance d_0 from each other with no velocity ($K = 0$, blue/green intersection). As they get closer (d decreases), the blue curve goes down, so K increases so they still add up to the constant E (the blue arrow goes further down so the red arrow has to get bigger so their sum still ends at the green line), meaning that $v = \sqrt{K/m}$ increases, i.e., the charges pick up speed. The blue curve going to $-\infty$ as $d \to 0$ tells us that K, thus v, goes to infinity ∞, as shown in question 4 of the problem.

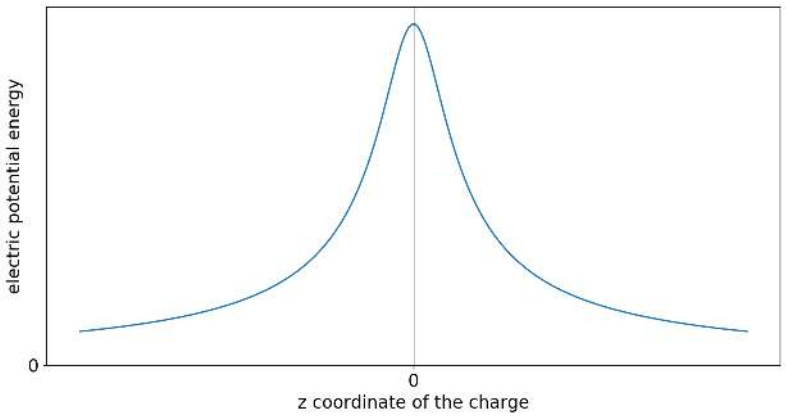
Problem 16: Throwing a charge through a charged ring 2.

We now return to the 3D charged ring, sketched below. As in problem 13, z denotes the z coordinate of B (we no longer use the h defined in problem 6) and we assume $Q = q$ (all four charges are equal).



As in problem 14, we want to know how fast the charge at B needs to be to pass through the ring given that it's being repelled by it (because $q = Q$), except this time we're going to use the conservation of energy, which will make the problem manageable. The three ring charges (A_1, A_2, A_3) remain fixed while B moves. The mass of B is m. At time $t = 0$, the moving charge is at $\vec{r}(0) = z_0\hat{z}$ with $z_0 > 0$ and its velocity is $\vec{v}(0) = -v_0\hat{z}$ with $v_0 > 0$.

1. Write the electric potential energy of the system as a function of k_e, q, R, and z.
2. What does it mean that z_0 and v_0 are positive?
3. Use the conservation of energy to write an equation relating the initial position and speed (z_0 and v_0) and the position and speed (z and v) some time later. Show that the energy terms corresponding to interactions between the charges of the ring are irrelevant. Show that dismissing said terms yields the potential energy of the moving charge as defined in section 2.3.1.
4. The curve below shows the electric potential energy of the moving charge. What condition must the energy of the charge satisfy in order to go through the ring (rather than being turn away by the repulsive electric force)?



5. Write the speed of the moving charge as a function of the other parameters. Show that the speed decreases as the charge approaches the ring.
6. Write the crossing condition as a condition for v_0. We will call this value the critical value of v_0.
7. When v_0 is smaller than the critical value from question 6, what happens to v at $z = 0$? What is the physical meaning of this mathematical result?
8. When the charge does cross the ring, what happens to its speed on the other side? What happens to it at $z = -z_0$?

2.3.3 Relationship between potential energy, force, and torque

The potential energy curve also tells us about the electric force felt by a charge. In problem 16, we plotted the potential energy as a function of the z coordinate of the position of the moving charge. The slope of that curve is the derivative $U'(z)$, also written $\frac{dU}{dz}$. The z component of the force exerted by the ring on the moving charge is equal to minus that derivative: $F_z = -\frac{dU}{dz}$. If we plotted the potential energy of the moving charge U as a function of its x coordinate, minus its derivative would give us the x component of the force on the charge: $F_x = -\frac{dU}{dx}$. If we plotted U as a function of the charge's y coordinate, minus its derivative would give us the y component of the force: $F_y = -\frac{dU}{dy}$. It's the same U every time, what change is which coordinate we treat as a variable when plotting and deriving.

In section 2.3.2, we plotted the potential energy of a pair of charges as a function of the distance d between them. Minus the derivative of that curve gives us the component of the force in the direction in which d is measured, i.e., along the line joining the two charges.

Since the force is minus the slope of U, it is positive when U decreases, meaning that it pushes the system towards higher values of the coordinate, thus lower values of U. If U increases, then the force is negative and pushes the charge towards lower values of the coordinate, thus lower values of U. In all cases, the force pushes the system towards lower values of the potential energy U.

This also work for rotation. Just like the force exerted on an object quantifies the action other objects have on its motion, the torque exerted on an object quantifies the action other objects have on its rotation. If rotating all or part of a system decreases its potential energy, then there is a torque pushing the system to rotate in the direction that decreases the potential energy. In other words, if rotating clockwise decreases the electric potential energy, then the electric force(s) tend to make the system rotate clockwise. Conversely, if rotating counterclockwise decrease the electric potential energy, then the electric force(s) tend to make the system rotate counterclockwise. (I say "tend to" because a clockwise torque does not guarantee clockwise rotation any more than a downward force guarantees downward motion; if there is an initial counterclockwise rotation, a clockwise torque will first slow it down, then effect a clockwise rotation.)

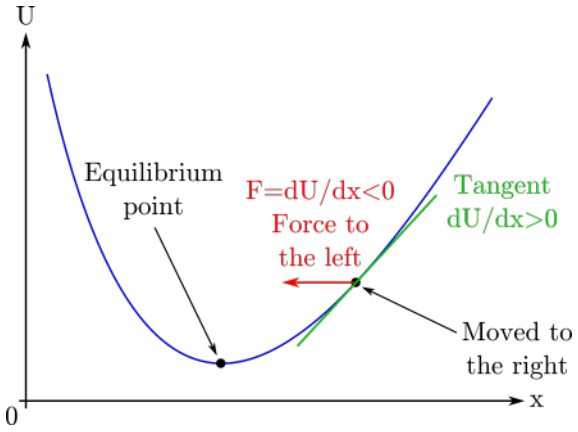
Problem 17: 3D charged ring: relationship between force and energy.

1. Derive the potential energy obtained in problem 16 with respect to z. Compare with the force obtained in problem 14. What is the relationship between the two?
2. Use the curve shown in problem 16 to discuss the evolution of the speed of the charge as it goes through the ring, first in terms of energy conservation, then in terms of the force exerted on it, which you'll infer from the potential energy curve. Show that the two approaches tell the same story.
3. What changes if the moving charge has charge $-q$ instead of $+q$? Write the new potential energy U and sketch it as a function of z. What condition must be satisfied for the charge to go through? Assuming it does go through, describe what happens to its speed.

2.3.4 Potential energy and stability

An object is said to be in mechanical equilibrium if the net force on it is zero. An equilibrium state is said to be stable if moving the object away from the equilibrium state results in a force that drives it back towards the equilibrium state. Conversely, an equilibrium state is said to be unstable if moving away from the equilibrium state results in a force pointing away from the equilibrium state.

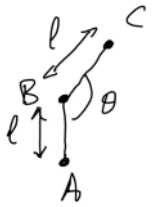
If the only forces are electric forces, the force is minus the derivative of the energy (more specifically, the $x/y/z$ component of the force is minus the derivative of the potential energy with respect to $x/y/z$). Thus, the force being zero means the potential energy has a horizontal tangent, i.e., it is extremal (either maximal or minimal). The figure below illustrates the case of a minimum as a function of the x coordinate. At the equilibrium point dU/dx, thus $F_x = -dU/dx = 0$. From there, increasing x moves the system to a point where $dU/dx > 0$, thus $F_x = -dU/dx < 0$. Since the force has a negative x component, it pushes the object towards lower values of x, i.e., back towards the equilibrium point. A similar reasoning can be used to show that moving to the left of the minimum results in a force towards the right, i.e., towards the equilibrium point. Thus a minimum of the potential energy is a stable equilibrium.



Near a maximum of the potential energy, the situation is reversed: moving to the right results in a force towards the right, and moving to the left results in a force towards the left. In both cases, the force pushes the system further away from its equilibrium point, meaning that the equilibrium is unstable.

Problem 18: A very short charged polymer.

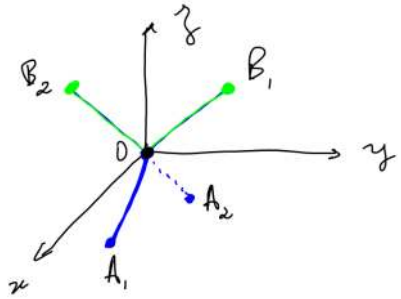
Three identical charges q located at points A, B, C are attached in a chain by two identical bonds of fixed length ℓ. The bond angle θ is free to change.



1. Compute the electric potential energy U of the system as a function of q, ℓ, θ.
2. Sketch $U(\theta)$. Discuss the stability of every equilibrium point.
3. If the system starts with $\theta = \pi/2$, which way does the electric force make it rotate? Discuss the torque acting on the hinge at B as a function of θ.
4. Summarize in plain English the implication of the nodes being charged, i.e., what does it change that $q \neq 0$ rather than $q = 0$?

Problem 19: Tetravalent molecule.

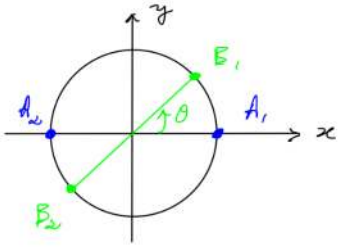
Molecules made of a central atom connected to four other usually take a tetrahedral shape like the one shown in the figure below:



$$A_1 = \begin{bmatrix} 1 \\ 0 \\ -1/\sqrt{2} \end{bmatrix}, A_2 = \begin{bmatrix} -1 \\ 0 \\ -1/\sqrt{2} \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ 1 \\ 1/\sqrt{2} \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ -1 \\ 1/\sqrt{2} \end{bmatrix}.$$

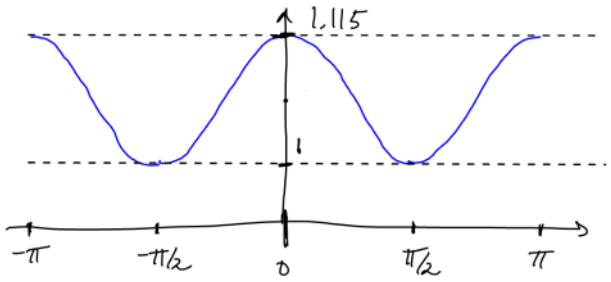
The goal of this problem is two-fold. First, show that the tetrahedral configuration is indeed the most stable one. Second, analyze what it takes to swap B_1 and B_2. Assuming the four peripheral atoms are distinct, the latter corresponds to converting one enantiomer into the other.

We make two major simplifications. First, we assume that each of the four peripheral atoms carries the same charge q and that the energy of the system is the electric potential energy of those four charges[1]. Second, we assume that the only way the molecule can deform is by rotating B_1B_2 around the z axis.



The sketch on the right shows molecule from above. θ is the angle between A_2A_1 and B_1B_2. Initially $\theta = \pi/2$. Swapping B_1 and B_2 means making $\theta = -\pi/2$.

1. Compute the coordinates of B_1 and B_2 during the transition as a function of θ.
2. Explain why the potential energy of the pairs A_1A_2 and B_1B_2 has no impact on the rotation dynamics.
3. Show that $A_1B_1 = A_2B_2 = \sqrt{4 - 2\cos\theta}$ and $A_1B_2 = A_2B_1 = \sqrt{4 + 2\cos\theta}$. Compute the potential energy.
4. The graph of the function $\dfrac{1}{\sqrt{4 - 2\cos\theta}} + \dfrac{1}{\sqrt{4 + 2\cos\theta}}$ looks like this:



What are the equilibrium values of θ? Discuss their stability. Describe the corresponding molecule shapes.

5. What is the minimum amount of kinetic energy needed to go from one stable state to the other. This is known as the height of the potential barrier.
6. In computing the potential energy, we did not discuss the contribution of the bond forces, i.e., the forces that ensure a constant bond length ($OA_1 = OA_2 = OB_1 = OB_2 =$ constant). What is their direction? Why is it ok to not include them when writing the conservation of energy?

[1]The real interactions are more complicated than that, with quantum mechanics playing a big role, but more detailed approaches still involve computing the energy as a function of the shape and looking for its extrema.

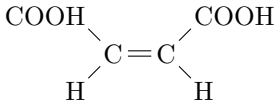
2.3.5 Potential energy and thermal agitation

At the molecular scale, thermal agitation plays a major role. Molecules are constantly colliding with each other, gaining kinetic energy, loosing it, being kicked into configurations they wouldn't have naturally explored, etc. To make matters worse, it's impossible to predict which collisions a specific molecule will undergo or when. In a word, it's chaos.

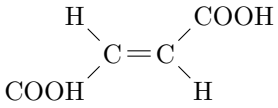
On the other hand, any system that's large enough to be observable has so many molecules that the law of large numbers apply. Statistical physics tells us that even though the behavior of a single molecule is as good as random, the probability of observing a specific configuration is proportional to $\exp\left(-\frac{U}{k_B T}\right)$ where U is the potential energy of that configuration, $k_B = 1.28 \times 10^{-23}\,\mathrm{m^2\,kg\,s^{-2}\,K^{-1}}$ is the Boltzmann constant, and T is the temperature in Kelvin. The law of large numbers then tells us with very good accuracy that if a molecule has, say, a 30% chance of being in a certain configuration, then 30% of the trillions upon trillions of copies of that molecule in the system will be in that configuration. For example, if a molecule has two configurations A and B with potential energies U_A and U_B respectively, the ratio of the number of molecules in configuration A to the number of molecules in configuration B will be $\exp(-U_A/(k_B T))/\exp(-U_B/(k_B T)) = \exp(-(U_A - U_B)/(k_B T))$.

Problem 20: Cis/trans isomerism.

We want to compute the energy difference between two following two isomers due to electric interactions:

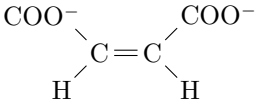


cis-butenedioic acid
(aka maleic acid)

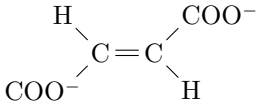


trans-butenedioic acid
(aka fumaric acid)

In water at pH> 7 both are overwhelmingly found in their dibasic form:



cis-butenedioate



trans-butenedioate

For the purpose of this problem, we simplify this to:

1. Compute the electric potential energy of each isomer.
2. Compute the ratio of the concentration of the cis isomer to the trans isomer. According to your result, which isomer is more abundant? Why?

Note: By definition, the ratio of the concentrations is the equilibrium constant of the reaction that transforms one isomer into the other. The interaction energy of the two charges does play a role, but there are a number of other factors we didn't discuss. For example, the bond lengths and angles are probably a little different in the two configurations. The solvent also plays a role. We'll say a few things about that later in this chapter. Then there's the contribution of entropy, which is a whole separate discussion.

2.3.6 Electric force and circular motion

Let's start with a refresher on curved motion. The acceleration is the rate of change of the velocity. In other words, assuming dt is very small, between times t and $t+dt$ the velocity vector changes by $d\vec{v} = \vec{a}\,dt$. The change $\vec{dv}$ can be split into two parts: $\vec{dv}_{||}$, which is parallel to $\vec{v}$, and $\vec{dv}_{\perp}$, which is perpendicular to $\vec{v}$. $\vec{dv}_{||}$ changes the magnitude of $\vec{v}$ but not its direction. Conversely, $\vec{dv}_{\perp}$ changes the direction of $\vec{v}$ but not its magnitude.

Looking at the change of direction, from the triangle formed by $\vec{v}$ and $\vec{dv}_{\perp}$ we see that the angle $d\theta$ by which $\vec{v}$ turns between t and $t+dt$ is $\tan(dv_{\perp}/v) \approx dv_{\perp}/v$ where $dv_{\perp} = ||\vec{dv}_{\perp}||$, $v = ||\vec{dv}||$, and we used $dv_{\perp}/v \ll 1$ to approximate the tangent. Thus the rate of turning of $\vec{v}$ (the angle turned per unit time) is $d\theta/dt = (dv_{\perp}/dt)/v = a_{\perp}/v$.

To relate this to the radius of curvature of the motion, we need to compute the rate of turning of the velocity on a circle with radius R. The distance traveled between t and $t+dt$ at speed v is $v\,dt$. The corresponding angle (measured from the center of the circle) is $d\theta = v\,dt/R$, and the corresponding rate of turning is $d\theta/dt = v/R$. Thus in a circular motion $d\theta/dt = v/R = a_{\perp}/v \implies a_{\perp} = v^2/R$.

If the shape of the trajectory is unknown, but the current speed and force are known, the formula can be used to predict the radius of curvature, thus the shape of the trajectory. This situation arises, for example, when analyzing the trajectory of a charged particle in a magnetic field in a mass spectrometer.

If on the other hand we know that the motion is circular with radius R, we can use the formula to know what the normal acceleration, thus the normal force, must be to maintain that circular motion. This arises for example when looking at the way an electron orbits around a nucleus.

Problem 21: Ionization energy.

A classical model of the hydrogen atom has the electron follow a circular orbit with radius R at constant speed v around the static proton. The electron has mass m and charge $-e$ where $e > 0$. The proton has charge e. A simple way to introduce quantum mechanics into this is to add the following constraint: the angular momentum of the electron around the proton must be a multiple of the reduced Planck constant $\hbar$. Mathematically, $mvR = n\hbar$ where n is any integer between 1 and infinity and each value of n corresponds to different orbit.

1. Write the electron's acceleration and the electric force exerted by the proton on the electron as a function of v, R, k_e, and e in the polar basis, then use Newton's second law to obtain a relationship between v, R, k_e, e, and m.
2. What angle does the force make with the velocity? Explain why a different angle would make uniform circular motion impossible.
3. Use the relationship you obtained in question 1 and the quantification relationship $mvR = n\hbar$ to compute v and R as functions of k_e, e, m, $\hbar$, and n. Compute the kinetic energy K of the orbit, its electric potential energy U, and its total mechanical energy (kinetic+potential) as a function of those same variables.

   Each integer value of n from 1 to ∞ corresponds to one possible orbit. Describe in words how each of the quantities above (v, R, U, K, E) varies with n.
4. Show that the electric potential energy and the kinetic energy obey a simple relationship that doesn't involve any other variable.
5. Sketch the electric potential energy U as a function of the distance R between the electron and the proton.

6. Pick a proton-electron distance R on the sketch. Use your answer to question 4 to draw the horizontal line corresponding to the mechanical energy E of a circular orbit with radius R.
7. How much additional energy does this electron need to have any chance of escaping the proton entirely (i.e., move arbitrarily far from the proton)? This is the ionization energy of that electron.
8. Compute the ionization energy I_1 for the lowest available orbit ($n = 1$). Look up the values of e (charge of a proton), m (mass of an electron), $\hbar$ (reduced Planck constant), and the ionization energy of hydrogen. How far off is this model?
9. Consider now a single electron in a circular orbit around a nucleus containing n_p protons (rather than 1). What is the new I_1? How does it depend on n_p?
10. What if there are two electrons? What happens to the reasoning we used in question 1? Why is this not a readily solvable problem?

2.3.7 Electric potential

We defined the electric field $\vec{E}$ created by one or more charges as the force those charges would exert on an hypothetical additional charge Q, divided by Q. It can be computed at any point whether or not there is an actual charge there.

Similarly, the electric potential created by one of more charges is the electric energy of an hypothetical additional charge Q, divided by Q. Specifically, the electric energy of a charge Q located at point B due to its interaction with a set of charges q_1,q_2,...,q_N located at points A_1,A_2,...,A_N is

$$U_B = \sum_{i=1}^{N} \frac{k_e Q q_i}{A_i B} = Q V_B \text{ where } V_B = \sum_{i=1}^{N} \frac{k_e q_i}{A_i B}$$

V_B is the electric potential created at point B by the charges at A_1,A_2,...A_N. Just like the electric field, it can be computed at any point B whether or not there is an actual charge there.

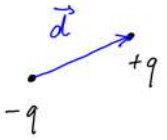
Since $V = U/Q$ and $\vec{E} = \vec{F}/Q$ where Q is the hypothetical charge, the relationship between $\vec{E}$ and V is the same as the relationship between $\vec{F}$ and U. In particular, the components of $\vec{E}$ can be obtained by deriving V with respect to the coordinates of B.

The benefit of thinking about the energy in terms of V will hopefully become clearer after we talk about dipoles (next section) and electric circuits (next chapter). In particular the electric voltage (one of two concepts at the core of electric circuits, the other being the electric current) is defined in terms of the electric potential.

2.4 Electric dipoles

2.4.1 Definition

A "physical dipole" consists of two opposite charges close to each other. $\vec{d}$ is called the separation vector. It goes from the $-q$ charge to the $+q$ charge. The distance between the two charges is $||\vec{d}||$. It is normally small, although the exact meaning of "small" is very context dependent. Unless specified otherwise q is positive. I will sometimes refer to it as the "dipole's charge". This is an abuse of terminology, but not a very ambiguous one in the sense that the actual net charge of a dipole is always $+q - q = 0$, and q is the next most logical thing to call the "dipole's charge".



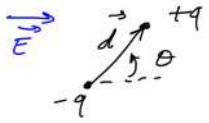
The "dipole moment" of a physical dipole is defined as $\vec{p} = q\vec{d}$. As it turns out, most of a dipole's properties only depend on q and $\vec{d}$ through $\vec{p}$. In other words, dipoles with the same $\vec{p}$ tend to behave the same even if their q and $\vec{d}$ are different. For example, a dipole with charge q and separation $\vec{d}$ and a dipole with charge $2q$ and separation $\vec{d}/2$ have the same dipole moment $\vec{p} = q\vec{d} = (2q)(\vec{d}/2)$, therefore their behavior in an electric field is mostly the same. *Note: Many chemistry texts use the opposite sign convention, i.e., $\vec{p}$ points from the + charge to the - charge. The convention I'm using here is the one used in physics texts as well as some physical chemistry texts.*

Real dipoles are a little more complicated. Most atoms and molecules have more than two charges. Even if there were only two charges, they would be spread around because of quantum mechanics. Fortunately, it is possible to compute the dipole moment $\vec{p}$ of any distribution of charge, no matter how complicated, and the behavior of the real dipole in an electric field is essentially the same as that of a physical dipole with the same dipole moment. For this reason, it is enough to understand the physics of physical dipoles, and they're the ones we'll focus on.

2.4.2 Force and torque on a dipole

Problem 22: Dipole in a uniform electric field.

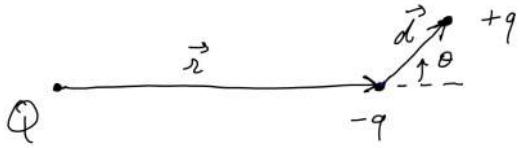
Consider a physical dipole with charges $+q$ and $-q$ and separation $\vec{d}$. q is positive (remember, unless specified otherwise q is always positive in dipole problems). We assume the charge are attached rigidly, i.e., the distance between them is constant. They are subject to the electric field $\vec{E}$ created by every charge in the universe not part of the dipole. We assume $\vec{E}$ is uniform in the vicinity of the dipole, i.e., it has the same value at every point. We dismiss the forces exerted by the two charges on each other as their effect is entirely canceled by whatever it is that keeps the distance between the two charges constant.



1. Write the electric force exerted on each of the dipole's charges by the charges creating $\vec{E}$ as a function of q and $\vec{E}$. Draw them on the sketch.
2. Compute the net force on the dipole. Why do the forces exerted by each dipole charge on the other not matter here?
3. Use the forces you drew in question 1 to predict the direction (clockwise or counterclockwise) of the torque resulting from the action of the electric field on the dipole.
4. Repeat the analysis of question 3 to discuss the direction of the torque as a function of the angle θ between the dipole and the electric field. What are the equilibrium positions (values of the angle for which the torque is zero)? Are they stable or unstable?
5. Summarize the torque result in one sentence (the torque always works to align the dipole moment with the same direction; which one?).

Problem 23: Dipole-charge interaction.

Instead of a uniform electric field, the dipole of problem 22 is now subject to the electric created by a third charge Q. Both q and Q are positive.



1. Sketch the forces exerted by Q on q and $-q$.
2. Identify the stable equilibrium orientation. What is the corresponding θ (Note: the dashed line used to define θ is the line joining Q and $-q$, which only happens to be horizontal in the sketch).
3. Assume the dipole has reached its stable equilibrium orientation. Sketch the force exerted by Q on each of the dipole's charges. Compare their magnitudes. What is the direction of the net force on the dipole?
4. Answer questions 2 and 3 when $q > 0$ and $Q < 0$.

2.4.3 Dipole energy

The behavior of an electric dipole in an electric field can also be understood in terms of the dipole's electric energy. It can be shown that this energy is $U = -\vec{p} \cdot \vec{E}$ where $\vec{p}$ is the dipole's moment and $\vec{E}$ is the electric field at the location of the dipole. The dipole is assumed to be small enough that the electric field at one end of it is not very different from the electric field at the other end.

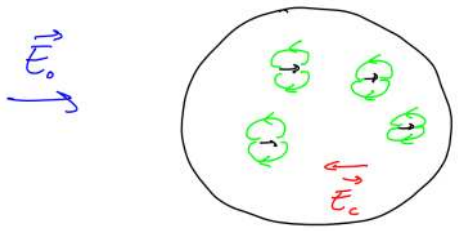
Problem 24: Dipole in a uniform field: energy approach.

1. The dipole consists of charges $+q$ and $-q$ separated by a distance d. The electric field has magnitude E. The angle between the dipole's separation vector and the field is θ. Write the dipole's energy as a function of those quantities.
2. Sketch the interaction energy as a function of θ. What dipole orientation has the lowest energy? What orientation has the highest energy?
3. Use the energy curve to discuss the direction of the torque as a function of θ.

2.5 Electric interactions in materials

2.5.1 Dielectrics

When a material that contains dipoles, like water, is subject to an electric field $\vec{E}_0$, the dipoles align with the field and create a counter-field $\vec{E}_c$. The actual field in the material is the sum of the original field and the counter-field. It can be shown that, at the end of the day, the effect of this is to divide the field by a dimensionless constant called the relative electric permittivity, noted ϵ_r, which depends on the material. In other words, the field that is actually felt in the material is $\vec{E}_0 + \vec{E}_c = \vec{E}_0/\epsilon_r$.



As it turns out, most materials exhibit this weakening of the electric field by a material-dependent factor ϵ_r. This is because even atoms and molecules that don't normally have dipole moment tend to acquire one when subjected to an electric field. The phenomenon is called *polarizability*, and the dipoles created this way are called *induced dipoles*. Unlike *permanent* dipoles, like that of the water molecule, induced dipoles disappear when the field $\vec{E}_0$ that created them is turned off.

ϵ_r is positive and larger than 1. In vacuum (no material) there's no weakening so $\epsilon_r = 1$. $0 < \epsilon_r < 1$ would correspond to a strengthening. $\epsilon_r < 0$ would mean the counter-field overpowers the original field such that their sum points opposite the original field. There are materials that behave like that, but they're not called dielectrics, and we won't discuss them here.

Values of ϵ_r for common materials are easily found online. Try typing "electric permittivity of water" into a search engine; there's a good chance the first link has the answer. Materials that have permanent dipoles tend to experience a much stronger weakening, i.e., they are a larger ϵ_r. A larger density of dipoles also increases ϵ_r. Water has both so its ϵ_r is quite large, around 80 at room temperature.

This effect is not limited to uniform electric fields. Inside a dielectric material, k_e is effectively replaced with k_e/ϵ_r. For example, the force exerted by a point charge at A on another at B and the interaction energy of that pair are given by

$$\vec{F}_{A\to B} = \frac{k_e q_A q_B}{\epsilon_r} \frac{\overrightarrow{AB}}{AB^3}, \quad U_{AB} = \frac{k_e q_A q_B}{\epsilon_r AB}$$

2.5.2 Ionic solutions

In addition to dipoles, water usually contains ions which are relatively free to move around. If a molecule has a positively charged region, it will attract the negative ions in the solution and repel the positive ones. This crowding of the charged regions with ions of the opposite sign also creates an even stronger type of counter field which makes the interaction energy of a pair of charges decay exponentially with the distance between them:

$$U_{AB} = \frac{k_e q_A q_B}{\epsilon_r AB} e^{-AB/\lambda}$$

where λ is the *screening length*, which depends on the concentration and charge of the ions present in the solution and the temperature. The electric interaction is then said to be *screened*.

We've discussed before the connection between the potential energy and the force. In the case of a potential with spherical symmetry (one that depends on the distance AB between the charges but not on the direction of $\overrightarrow{AB}$, which is the case here), the force is obtained by deriving the potential with respect to AB, putting a minus sign in front, and multiplying by the unit vector $\overrightarrow{AB}/AB$. The result is:

$$\vec{F}_{A\to B} = -\frac{d(U_{AB})}{d(AB)}\frac{\overrightarrow{AB}}{AB} = \frac{k_e q_A q_B}{\epsilon_r}\left(\frac{1}{AB^3} + \frac{1}{\lambda AB^2}\right) e^{-AB/\lambda}\ \overrightarrow{AB}$$

This correction is very significant. Exponentials decay slowly at first, then very quickly. When AB is larger than a few times λ, the potential and the force are essentially zero. Therefore the range of the interaction is effectively limited to a few λ's.

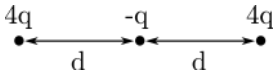
On top of that, λ can vary a lot depending on ionic concentrations. In pure water at $pH = 7$, $\lambda \sim 1\,\mu m$. That's the size of a smallish organelle. In a 0.1 M KCl solution, $\lambda \sim 3\,nm$. That's about 3000 times less, the size of a small biomolecule, say a phospholipid, or ATP. It makes a huge difference.

To take a slightly more concrete example, protein folding relies a lot on finely tuned electric interaction between charged residues. Changing ionic concentrations disrupts the balance of those interactions and can prevent the protein from folding correctly, i.e., denature it.

Problem 25: Screened force.

Redo problem 2 assuming a relative permittivity $\epsilon_r = 2$ and (1) no screening, (2) a screening length $\lambda = 5\,cm$, and (3) a screening length $\lambda = 0.5\,cm$. Compare the three results. Are they consistent with the general notion that the interaction decays quickly beyond the screening length? Why or why not?

Problem 26: Screened potential.



1. Compute the electric potential energy of this object in a dielectric without screening. Compare with the electric potential energy of the if it had no charges ($q = 0$). Which of the charged or uncharged form is more stable (has the lowest electric potential energy)?
2. Do the same with a screening length λ. What does that mean for the stability of that object as a function the concentration of ions in the solution hosting it?

Chapter 3: Electricity

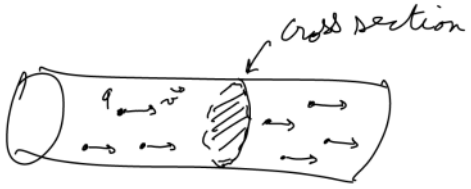
3.1 Electric current

3.1.1 Definition and microscopic origin

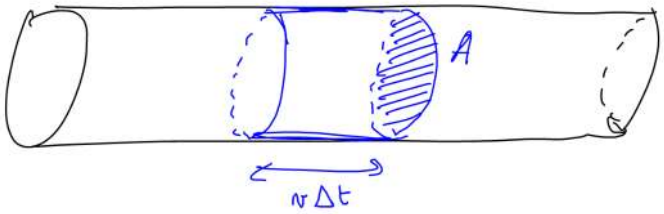
Electric current is the motion of charges through a material. It is the defining feature of electric circuits.

All material have charges: they're made of atoms, which are each made of a positively charged nucleus and some negatively charged electrons. Not all of those charges can move through the material though. In fact, in a lot of materials none of the charges are free to move across the material. They can vibrate a little bit around their original position, but that's it. Those types of materials are called insulators. Materials that do have freely moving charges are called conductors. Even then, only a small fraction of the charged particles that make up the material can move this way. In the context of electricity, though, the moving charges are the only ones that matter. They're the ones that carry the current. The fixed charges merely hold the material together and slow down the motion of the moving charges by acting as obstacles.

More precisely, the electric current is a number that quantifies the amount of charge traveling through a material. Consider the wire below. Only the moving charges are represented. Each individual moving particle has a the same charge q and travels at the same speed v along the wire. The *cross section* is an imaginary surface running across the material perpendicular to the motion of the particles which divides the material into an *upstream* side and a *downstream* side.



In this particular sketch upstream is left and downstream is right. Each side has a net charge, which is the sum of the charge of all the particles in it. Every time a moving charge crosses the divide, the net charge of the downstream side increases by q while the net charge of the upstream side decreases by q. By definition, the electric current is the rate at which the net charge of the upstream side is transferred to the downstream side. To compute it, let's think about the charges crossing between a time t and a later time $t+\Delta t$. During that time, every charge moves downstream by a distance $v\Delta t$. The charges that started more than $v\Delta t$ upstream of the divide will still be on the upstream side at $t+\Delta t$. The charges that started less than $v\Delta t$ upstream of the divide will cross. The charges that started downstream of the divide will move further downstream. Therefore, the number of charges that cross the divide between t and $t+\Delta t$ is exactly equal to the number of charges contained in a length $v\Delta t$ of conductor. In the sketch below, the charges that cross the red surface between t and $t+\Delta t$ are the ones that were initially in the blue region.



Let A be the area of the cross section. Let n be the density of freely moving charges in the material (number of charges per unit volume). The volume of the region whose charges will cross during Δt is $A(v\Delta t)$. The number of charges in that volume is $nAv\Delta t$. You can also think of nA as the number of moving charges per unit length of the wire. The net charge of all those particles is $qnAv\Delta t$. That's the net charge that has been transferred from the upstream side to the downstream side during Δt. Finally, the electric current, noted I, is the rate at which net charge is transferred, i.e., the net charge transferred divided by the duration Δt of the transfer:

$$I = qnAv$$

This formula contains a fair bit of information about what makes a good conductor. It needs to have a lot of free charges (large n), ideally each carrying a large charge (large q), it needs to be thick (large cross sectional area A), and it needs to give easy passage to the charges (large speed v).

The SI unit of electric current is the *Ampere*, symbol A. Since an electric current is a charge divided by a time, the SI unit of current is equal to the SI unit of charge divided by the SI unit of time: one Ampere equals one Coulomb per second ($1\,\mathrm{A} = 1\,\mathrm{C/s}$).

The current can be positive or negative. The sign depends on three things:

- The sign of the individual particles' charge. If q is negative, then each particle crossing the divide decreases the net charge of the downstream side while increasing that of the upstream side.
- The direction of the velocity. If $\vec{v}$ points upstream, then the number of particles crossing per unit time is still nAv, but each crossing removes a charge q from the downstream side and adds it to upstream side. In terms of net charge, this is the same as adding a charge $-q$ to the downstream side and removing it from the upstream side. Therefore, the current is $I = -qnAv$. In other words, the speed v is counted positively if the charges are moving downstream and negatively if they are moving upstream.
- The "current arrow". Since the definition works regardless of whether the charges are actually moving downstream (as opposed to moving upstream), it's entirely up to us which side we want to call upstream. If the net charge transferred from left to right during a time interval Δt is Q, then the net charge transferred from right to left during that same time is $-Q$. Therefore, changing the definition of upstream (swapping upstream and downstream) changes the sign of the current. This is consistent with the previous bullet point. For positive charges ($q > 0$), the current is positive if $\vec{v}$ points downstream and negative if $\vec{v}$ points upstream. Flipping the direction of $\vec{v}$ changes the sign of I. So does swapping upstream and downstream. Flipping $\vec{v}$ and swapping sides doesn't change I as the two sign changes cancel each other.

  The most important thing here is that it doesn't make sense to compute a current if you haven't first decided which side is going to be upstream/downstream. The convention when drawing electric circuits is to draw wires as lines and use an arrow head to indicate which side is downstream:

Problem 27: Sign of the current.

In each of the following cases, predict the sign of the current and explain your reasoning. In one of the cases the sign of the current cannot be predicted; explain why. Drawing a sketch will help.

1. Positive charges moving along the current arrow.

2. Positive charges moving against the current arrow.

3. Negative charges moving against the current arrow.

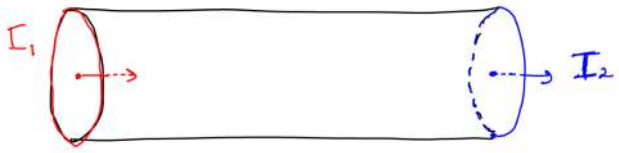
Same question for a material with both positive and negative charges:

4. Positive charges moving against the current arrow, negative charges moving along the current.

5. Positive and negative charges moving against the current arrow.

3.1.2 Charge conservation and current

Problem 28: Charge conservation and current through a wire.

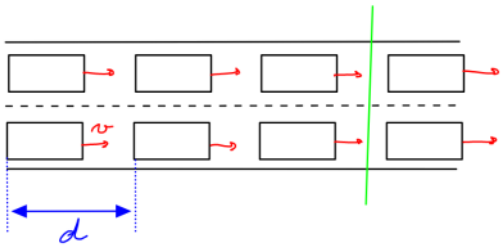


The piece of conductor above, hereafter "the system", has an initial charge Q_0 at $t = 0$. The electric current I_1 is the rate of net charge transfer through the red cross section in the direction of the red arrow. I_2 is the rate of net charge transfer through the blue cross section in the direction of the blue arrow. I_1 and I_2 are constant, i.e., they do not change through time. Charges cannot enter or leave through the curved sides (curved sides=everything except the red cross section and the blue cross section). Charge cannot appear or disappear spontaneously either. In other words, the only way the net charge inside the wire can change is because charge entered or exited through either the red or the blue cross section.

1. What is the total amount of charge $Q(\Delta t)$ in the system after a time Δt.
2. What happens to Q after a long time $\Delta t \to \infty$? What relationship must I_1 and I_2 satisfy for things to remain reasonable?

The same concepts apply to other conserved quantities, like the mass of a system or its number of object of a certain type. Here is an example with cars on a road.

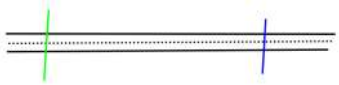
Problem 29: Conservation, current, and car traffic.



The sketch above shows a two-lane road. Every car has the same constant speed v and the same constant rear-to-rear distance d to the car in front of it. We want to compute the "car current" across the green line, i.e., the number of cars that cross the green line per unit time. Over short periods of time, the current jumps back and forth between 0 and 1 depending on whether there's a car crossing at that exact time, but that's not what we're interested in. What we want is the average car current over a longer period of time, say, how many cars on average pass through the line in a minute.

1. How many cars does a region of length ℓ contain on average? *Hint: How many cars in a stretch of road of length d? $2d$? How many stretches of length d in a stretch of length ℓ?*
2. How much distance does a car travel between $t = 0$ and $t = \Delta t$? Where does a car need to be at $t = 0$ in order to cross the green line at any time between $t = 0$ and $t = \Delta t$? What is the length of that region? What is the average number of cars in it?
3. How many cars cross the green line per unit time on average?

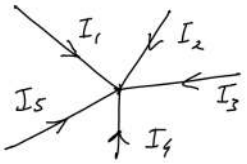
Consider now a steady traffic jam. By that I mean that although the car in it are changing, seen from far above the traffic pattern does not change through time. Specifically, we assume that v and d can take different values at different locations along the road, but their value at any given location remains constant through time. The total number of cars between the green line and the blue line also remains constant (it may go down by one when a car exits at the blue line then go back up shortly after as a new car enters through the green line, but we don't worry about that; think big picture). At the green line, the speed and distance are v_1 and d_1. At the blue line, they are v_2 and d_2.



4. Let N_0 be the total number of cars between the green line and the blue line. What is the new number after a time Δt?
5. What relationship must v_1, d_1, v_2, d_2 obey if the number of cars between the two lines is going to stay constant?
6. Imagine the green line is inside the traffic jam and the blue line is after the traffic jam. v_2 is larger than v_1. What can you say about d_1 and d_2? Is that the result you expected?

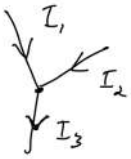
3.1.3 The junction rule

In practice the energetic cost of changing the net charge of any part of a circuit is almost always prohibitive. As a result the sum of the current into any subset of a circuit has to cancel at all times. For a straight wire, this means that the current is the same at both ends, as discussed in problem 28. The more general rule is that the sum of all the currents entering a region must vanish. It is called either the *junction rule* or *Kirchhoff's first law*. For example:



$$I_1+I_2+I_3+I_4+I_5=0$$

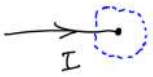
Currents point away from the region must be counted negatively (see current I_2 in problem 28). Alternatively you can write that the sum of the currents entering the region is equal to the sum of the currents exiting the region; it leads to the same equation. For example:



incoming − outgoing = 0 : $I_1+I_2-I_3=0$.

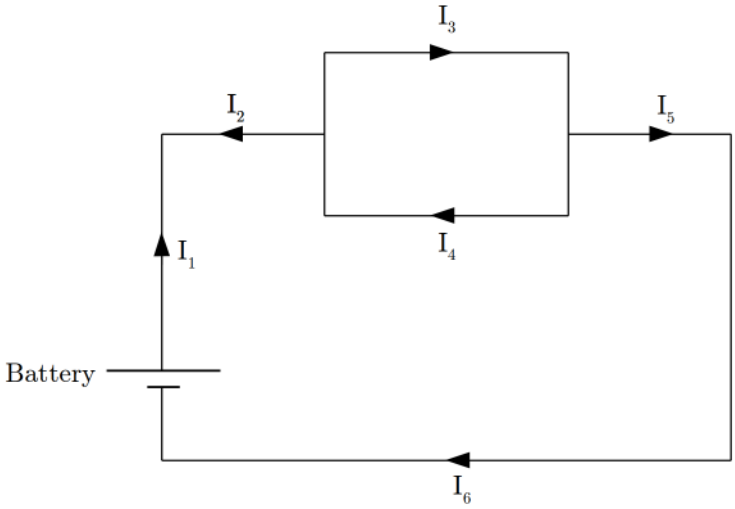
incoming = outgoing : $I_1+I_2=I_3$.

An important consequence of this rule is that current can only flow if the circuit forms a closed loop. Consider this open wire:



The blue region has an incoming current (I) but no outgoing current. The junction rule yields $I = 0$.

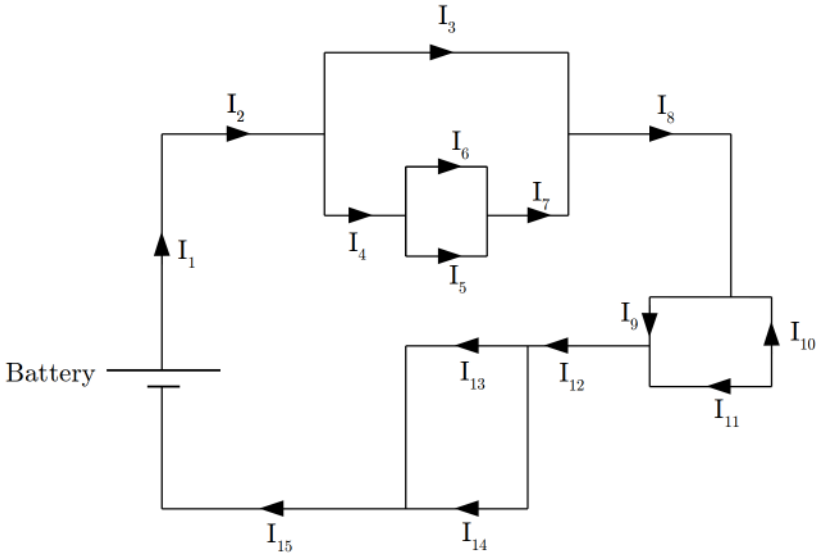
Problem 30: Junction rule.



1. Apply the junction rule to obtain as many independent relationships as possible between the currents.
2. Find the smallest possible set of currents that every other current can be computed from. How many are there? (There are many correct minimal sets of currents, but only one correct minimal number of currents.)
3. Redraw the circuit only keeping a minimal set of independent currents, then write the current in every other branch as a function of those currents.

Problem 31: Junction rule 2.

Answer the same questions as in problem 30 for the circuit below.

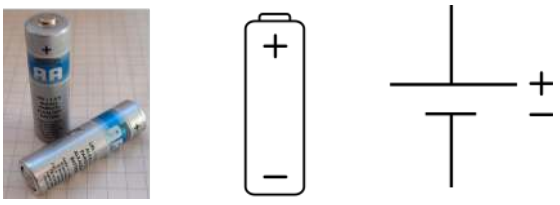


3.2 Voltage

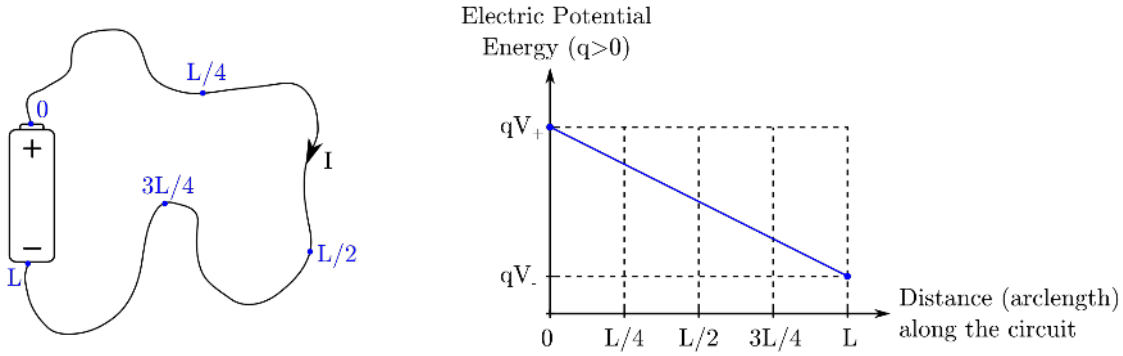
3.2.1 Electric potential and current

What pushes charges through a circuit is the electric force. What sources of electricity (a battery, a power outlet, etc) tend to impose, however, is not the force but the electric potential.

Say the electric potential is V_+ at the positive terminal of a battery and V_- at its negative terminal, with $V_+ > V_-$. By definition of the electric potential, the electric potential energy of a charge q is qV_+ at the positive terminal and qV_- at the negative terminal. Thus a positive charge ($q > 0$) has a higher energy at the positive terminal than at the negative one ($qV_+ > qV_-$). That in turn means there is an electric force pushing the charge in the overall direction of the negative terminal. Specifically, the general result that the force is the derivative of the potential energy is to be interpreted as follows: the electric force is always parallel to the circuit, and its value is the derivative of the potential energy with respect to the distance along the circuit.



The figure below illustrates the behavior of the potential energy and the force along a simple circuit made of a battery and a wire with length L:



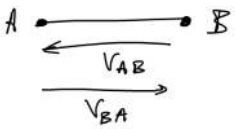
When the wire is homogeneous (same material properties and same cross sectional area throughout), the electric potential energy decreases at a constant rate, i.e., it is a line, and the magnitude of the electric force is $F = \frac{q}{L}(V_+ - V_-)$.

If the moving charges are negative, then $qV_+ < qV_-$ therefore the electric force pushes the charges from the negative terminal (the electric force always points in a direction in which potential energy decreases).

In terms of the current I, it doesn't matter whether it's positive charges moving from the negative terminal to the positive one or negative charges moving the other way, the sign of the current is the same. If the current arrow points from the positive terminal to the negative one, then $V_+ > V_- \implies I > 0$. If we had drawn the current arrow pointing from the negative terminal to the positive one (remember the direction of the current arrow is ours to choose), then the current would be negative. Either way, the current "flows" from the higher electric potential (V_+) to the lower one (V_-). The quotes around "flows" are there to remind you that the current is not a physical thing that actually flows, it's an abstract concept (net charge crossing a surface in a specified direction per unit time), that the actual charges may very well be moving the other way (if they're negative; see problem 27), and that the only way to fully grasp the meaning of the current flowing in a certain direction is to understand the reasoning that led us to this point.

3.2.2 Voltage

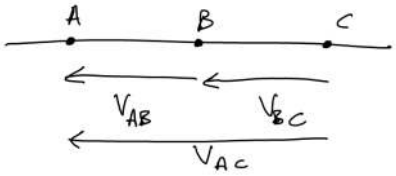
As the formula $F = \frac{q}{L}(V_+ - V_-)$ above suggests, the value of the electric potential at either terminal doesn't really matter, what matters is the difference of potential $(V_+ - V_-)$ between the two that sets the force F. If there's no difference of potential, there's no force. A voltage is simply that: a difference of potential between two points. More precisely, the voltage V_{AB} between any two points A and point B is defined as the difference between the electric potential V_A at point A and the electric potential V_B at point B: $V_{AB} = V_A - V_B$. When sketching a circuit, a voltage is represented by an arrow along the side of the circuit and pointing from B to A:



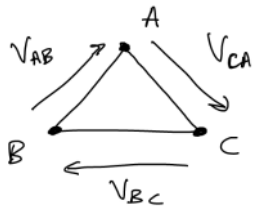
One has to be careful with notations here. A first source of confusion is that the same upper case "V" is used for the electric potential and the voltage. Whenever possible I write electric potentials with a single subscript representing the point where it's computed and voltages with two subscripts representating the two points between which it's computed, but that's not always possible. To make matters worse, electric potentials and voltages have the same SI unit, the Volt, whose symbol is also an upper case V. On the bright side, we mostly need to talk about electric potentials to understand what a voltage is and how to handle it. Once we start talking about real circuits we'll essentially stop talking about the electric potential and every "V" will be a voltage. A second source of confusion is that, when writing a voltage, the order of the points matters. Specifically, $V_{AB} = (V_A - V_B) = -(V_B - V_A) = -V_{BA}$. That means the direction of the arrow also matters. Just like current arrows, reversing a voltage arrow flips the sign of the voltage (see figure above: if the arrow goes from B to A it represents V_{AB}, if it points from A to B it represents $V_{BA} = -V_{AB}$).

3.2.3 Voltage additivity and the loop rule

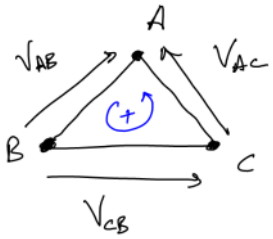
The loop rule is one of three key steps in analyzing any circuit (the junction rule is another). The point of this is not going to be obvious at first, but it will come together soon. Let's start with voltage additivity. A consequence of the definition of voltage as a difference of potential is that voltages add up a little bit like vectors: the same way you can write $\overrightarrow{AB} + \overrightarrow{BC} = \overrightarrow{AC}$, you can write $V_{AB} + V_{BC} = (V_A - V_B) + (V_B - V_C) = V_A - V_C = V_{AC}$.



If we apply this to a closed loop, the voltages add up to zero: $V_{AB} + V_{BC} + V_{CA} = (V_A - V_B) + (V_B - V_C) + (V_C - V_A) = 0$.



This is called either the *loop rule* or *Kirchhoff's second law*. If the voltage arrows don't all point in the same direction. Since reversing the arrow simply flips the sign of the voltage, the loop rule still works as long as we put a minus sign in front of any voltage pointing the "wrong" way. Which way you choose to be the "right" way to go around the loop is of no consequence, what matters is to pick a right way at the beginning (clockwise or counterclockwise; blue circular arrow with a "+" sign in the figure below) and be consistent in counting voltages positively if they point the right way and negatively if they point the wrong way.



Problem 32: Loop rule.

Write the loop rule for the loop above, but pick clockwise as the positive direction around the loop. How does the result differ from the result obtained above. Why does it not matter which direction gets picked as positive?

3.3 Ohm's law

Ohm's law is a relationship between the voltage between the two ends of a piece of conductor and the amount of current going through it. In other words, it tells us how many charges pass through the material as a function of how hard the power source is pushing them (or, if used the other way, how hard one should drive the charges to get them moving in a certain amount).

Let's go back to the relationship between the current and the speed of the moving charges: $I = qvnA$. The density of moving particles n and the individual charge those particles are properties of the material the charges are moving through (e.g. copper, salt water, etc). The cross sectional area A has to do with the geometry of the material (How big is the piece of copper? How wide is the cuvette holding the salt water?). The speed v deserves a little more attention.

Let's focus on a single charge q traveling at speed v. It is driven by the electric force $F = qV/L$ where V is the voltage between the two ends of the conductor (aka the voltage *across* the conductor) and L is the conductor's length (see the voltage section). If that was the only force, the charge would keep accelerating indefinitely: $m\dfrac{dv}{dt} = F \implies v \sim Ft$. The reason it does not is because there is some friction between the charge and the surrounding material (all the nonmoving nuclei and electrons in the material). The details of that friction are a bit complicated, but the friction force ends up being proportional to the speed, same as an object moving through a viscous fluid (e.g. a marble falling through oil). Let's call α the friction coefficient, which is a property of the material, then the equation of motion reads $m\dfrac{dv}{dt} = \dfrac{qV}{L} - \alpha v$. This is a very solvable differential equation for $v(t)$, but for now let's just say that whatever v is initially, it very quickly relaxes towards a constant value. Being constant implies $dv/dt = 0$, which in turn sets the value of the speed: $qV/L - \alpha v = 0 \implies v = \dfrac{qV}{\alpha L}$. Finally we plug this into $I = qnAv$ to get

$$V = \left(\frac{\alpha L}{q^2 n A}\right) I$$

To help make sense of this result, we introduce the conductor's resistivity $\rho = \dfrac{\alpha}{q^2 n}$ and the conductor's resistance $R = \dfrac{\rho L}{A}$ and rewrite the equation above as

$$V = R\,I = \frac{\rho L}{A}\,I$$

What this is saying is that:

- The more current we want to push through the conductor, the more voltage we need (the "stronger" the power source).
- The voltage required is actually proportional to the amount of current we want. The proportionality constant is called the resistance R. The smaller R, the easier it is to drive current through the conductor, the better it conducts electricity.
- The resistance is a property of the piece of conductor, i.e., different pieces of conductor have different values of R, but once we've chosen a specific piece of conductor R is constant.
- There are two factors contributing to a conductor's resistance: what it's made of (its material properties q, n, and α, summarized by the resistivity $\rho = \alpha/(q^2 n)$), and its shape (length L and cross sectional area A).
- A good conducting material needs lots of moving charges and few obstacles to their motion (small friction coefficient).
- Once a material has been chosen, the resistance can be decreased by increasing the cross sectional area. In other words, a thicker wire conducts electricity better.
- The longer the wire, the larger the resistance, the harder it is to conduct electricity through it.

Note that you can look up the resistivity of a material, say copper, online, but you can't look up the resistance of a conductor because there would be as many values as there are possible shapes. In other words, to predict the resistance of a specific piece of a known conducting material, one needs to look up the resistivity of that material, measure its shape (length and cross sectional area), then use $R = \rho L/A$. Another way this can be used is to determine the length and area you need to give a piece of a known conducting material to get the resistance you need for a given application. Of course you can also buy a piece of conductor with just about any desired resistance, called a resistor. In that case the calculation above is still happening, you've just outsourced it to the manufacturer of the resistor.

Ohm's law

The proportionality between the voltage across a conductor and the current going through it, $V = RI$, is known as *Ohm's law*.

SI units

Currents are measured in Amperes (A), which are the same as Coulomb per second ($1\,\mathrm{A} = 1\,\mathrm{C/s}$). Voltages are measured in Volts (V). A voltage has the same dimension as an electric potential, and an electric potential times a charge is an energy, therefore a Volt is the same as a Joule per Coulomb ($1\,\mathrm{V} = 1\,\mathrm{J/C}$). Resistances are measured in Ohms (Ω). By Ohm's law, $V = RI$, an Ohm is the same as a Volt per Ampere ($1\,\Omega = 1\,\mathrm{V/A}$).

Electric current is a fundamental quantity just like length, time, and mass. It follows that the Ampere is a fundamental SI unit just like the meter, the second, and the kilogram. In other words, 1A cannot be written in terms of m, s, and kg, however any unit you'll encounter in mechanics or electricity can be written as a combination of m, s, kg, and A.

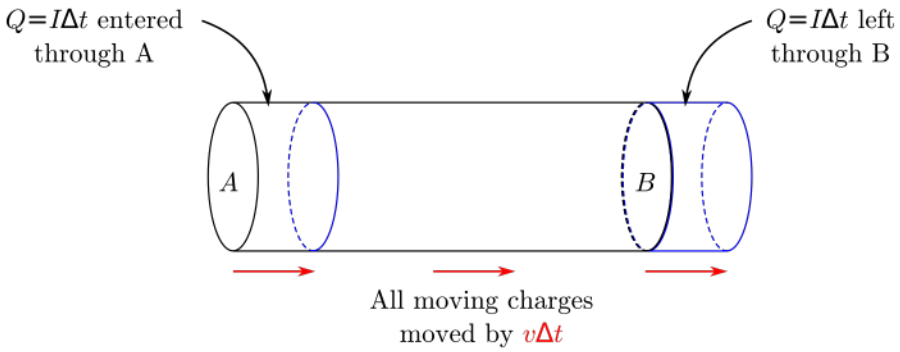
Problem 33: Resistance of a copper wire.

Look up the resistivity of copper online. Compute the resistance of a 1 m-long cylindrical copper wire with diameter 1 mm. Compute the current going through it when connected to a standard AA battery (1.5 V).

3.4 Electricity and energy

3.4.1 Energy received by a circuit

Let's think about a stretch of circuit extending from A to B and traversed by a current I. Let's call that stretch of circuit and whatever charges it contains at a given time the *system*. The sketch shows it as a cylindrical wire, but the argument holds for just about any circuit with one entrance and one exit, no matter what's in between (forks, resistors, capacitors, inductors, engines, radio antennae, etc).



To keep it simple, let's assume there's only one kind of moving charge, with individual charge q, moving left to right at speed v. As time passes, new charges enter the system at A while others leave the system at B. Each charge q that enters at A comes in with an electric energy qV_A. Each charges that exits at B leaves with an energy qV_B. Every time a new charge enters the system while another leaves it, the energy of the system changes by $qV_A - qV_B = qV_{AB}$ where $V_{AB} = V_A - V_B$ is the voltage across the stretch of circuit. Over a time Δt, every moving charge moves to the right by $v\,\Delta t$ the net charge added to the system at A is $Q = I\,\Delta t$, and the same net charge is removed at B (because of the junction rule, which imposes that the current at B is the same as the current at A). Therefore, the current I passing through the system during Δt has changed the system's energy by $Q\,V_{AB} = V_{AB}\,I\,\Delta t$. This is called the electric work received by the system during Δt, usually noted W (same letter as mechanical work, because it's closely related to the mechanical work of the electric force).

In summary, the energy W received from the electric force by the system (the stretch of circuit between A and B) during Δt is:

$$W = V_{AB} I \Delta t.$$

Like the work we defined in physics 1, W represents an energy transfer. The energy is coming from the power source, which supplies the voltage, thus the electric force. What becomes of that energy depends on the material the current is traversing, i.e., what exactly is between A and B. If it's vacuum, then the moving charges just accelerate, and the electric energy is converted into the kinetic energy of those charges. If AB is a conducting wire, the charges loose the energy almost instantly by bumping into the atoms of the material which start to rattle and vibrate. That's also kinetic energy, but of a disorganized kind, which appears to us as an increase of temperature of the conductor. We say that the energy has been dissipated, or turned into heat.

3.4.2 Power received by a circuit

Just like in mechanics, we define the power received as the energy received per unit time:

$$P = \frac{W}{\Delta t} = V_{AB} I.$$

3.4.3 Energy dissipation in conductors

In order to discuss temperature changes caused by electricity, we need to talk about the relationship between energy and temperature. When a object receives energy in the form of heat, its temperature increases. The amount of energy W the object needs to receive to increase its temperature from T to $T + \Delta T$ is proportional to the temperature increase ΔT, the mass m of the object, and a property of the material the object is made of called its *specific heat* c:

$$W = m\, c\, \Delta T$$

Problem 34: Resistive heating.

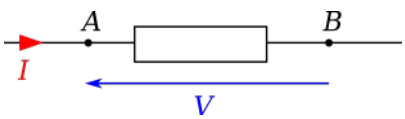
Don't try this at home. It has the potential to burn your fingers, or even start a fire.
The poles of a 1.5 V battery are connected with a copper wire of length 1 m and diameter 1 mm (same as in problem 34) for 1 minute. The wire is covered with a very good thermal insulator, which allows us to assume that all the energy dissipated by the electric current in the wire stays in the wire and goes toward heating it.

1. How much does the temperature of the wire increase during that 1 minute? The density of copper is 8.96 g/cm^3. The specific heat of copper is 0.385 J/g/°C (Joules per gram per celsius).
2. What happens if the battery remains connected forever and never runs out?
3. The battery advertises 1.5 V and 1350 mAh (mAh=milliAmpere hour). What is the latter in SI units? Based on its dimension, guess what does it represent? How long can this battery send the current we computed above through the wire? How much energy gets dissipated in total during that time?

3.4.4 Sign conventions

Energy and power

The work and power formulas above were derived assuming the voltage arrow is opposite the current and voltage arrow directions (the empty rectangle represents whatever the circuit contains between A and B: conductors, batteries, etc):



With those conventions, $W = VI\Delta t$ is the energy received by the circuit located between A and B during a time interval Δt, and $P = VI$ is the power received by that same circuit between A and B, i.e., the energy it receives per unit time.

Those formulas work regardless of the signs of V, I, P, and W. If P and W are positive, then whatever is between A and B is receiving energy from the charges that make up the electric current. This is what we expect for a conductor as the moving charges give their energy away to the material when they bump into its atoms. Conversely, if P and W are negative then the contents of the rectangle are giving energy to the moving charges. This is what we expect for a battery, which pushes the charges along the circuit.

Ohm's law

Ohm's law, $V = RI$, also assumes that the V is oriented opposite I. The resistance R is always positive. So is the resistivity ρ.

Problem 35: Sign conventions.

Let's delve into what happens to the energy, power, and Ohm's formulas when the current and voltage arrows are oriented differently. Assume the formulas hold for the sketch above.

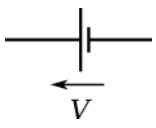
1. Let's define I' pointing left and V' pointing right. How do I' and V' relate to I and V? What is the correct form of the power and energy formulas if instead of writing them in terms of I and V we want to right then in terms of (1) I and V', (2) I' and V, (3) I' and V'. Sum up your results in terms of V being along I vs opposite I.
2. Assume $I > 0$, made up of positive charges moving to the right, and $V > 0$. Summarize, in words, why the circuit between A and B receives energy by discussing the sign of the change of energy of the moving charges (you may need to reread section 3.4.1).
3. Assume $I > 0$, made up of positive charges moving to the right, and that AB is a conductor. Explain, in words, why the fact that the conductor exerts a friction force directed opposite the motion of the charges implies that V must be positive.

3.5 Electric circuits

3.5.1 Drawing electric circuits

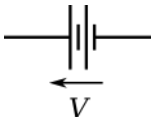
Here are the circuit components we need for now. We'll introduce more later.

- Voltage source:
  Example: a battery. Imposes a predetermined voltage V between its poles regardless of what the rest of the circuit does. Can deliver any amount of current required for the voltage to remain equal to V.

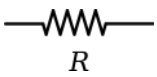


  The horizontal lines represent the wires that connect the voltage source to the rest of the circuit. The vertical lines represent the poles (also known as terminals) of the voltage source. The longer line (positive terminal) has a higher electric potential than the shorter line (negative terminal). Therefore, V is positive if the voltage arrow points toward the the longer line (as in the sketch) and negative if it points toward the shorter line. If the arrow is not drawn, it is implied that it points toward the positive terminal.

Sometimes the symbol is stacked like so:



- Resistor:
  Ohmic conductor of the kind discussed in section 3.3.



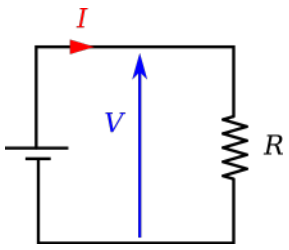
- Ideal wire:
  Idealized wire with no resistance at all ($R = 0$). Connects the other components of the circuit to each other. The voltage across an idealized wire is $V = RI = 0$, i.e., it is zero no matter how much current goes through it.

Comments:

- Real wires have a nonzero resistance. Often times that resistance is small enough to be neglected, and we model the real wire as an ideal wire. Sometimes, though, the wire's resistance does matter. In those cases, we model the wire as a resistor instead (still with ideal wires connecting it to the rest of the circuit).
- Since the voltage across any two points of an ideal wire is 0, the electric potential is the same all along the wire and the end points of voltage arrows can be moved along the ideal it without consequence.

3.5.2 Simple circuits

Problem 36: Simple circuit.

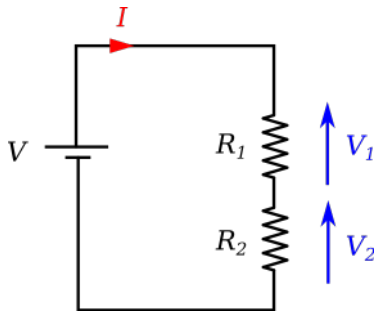


Note: When assembling a circuit like this, you typically have direct control V (by choosing the battery) and R (by choosing the resistor), but you have no direct control over I: it is determined by V and R. Accordingly, treat I as a unknown and V and R as parameters. In other word, write your answers in terms of V and R rather than I.

1. Based on the drawing of the voltage source, what is the sign of V? Which way does the electric field point in the resistor? Assuming the moving charges are negative, which way do they move in the resistor? Which way do they move around the circuit?

2. Write V in terms of R and I.
3. How much power does the source receive? How much power does the resistor receive? What is the sign of each? Summarize, in words, the overall flow of energy in this circuit.
4. What value of R maximizes the power dissipated in the resistor?

Problem 37: Two resistors.



Note: Again the parameters are the source's voltage and the resistances, which one typically has direct control over, whereas I, V_1, and V_2 are unknowns. Some of the questions ask for an answer in terms of an unknown, but they're only intermediate steps toward a later answer in terms of the parameters only.

1. Write V_1 and V_2 in terms of I, R_1, and R_2.
2. Write V in terms of V_1 and V_2.
3. Write I in terms of V, R_1, and R_2.
4. Write V_1 and V_2 in terms of V, R_1, and R_2.
5. Compute the power P_1 dissipated in the first resistor and the power P_2 dissipated in the second resistor in terms of V, R_1, and R_2.

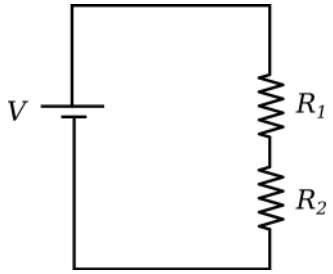
3.5.3 General method

Here are the typical steps one needs to go through to predict the behavior fo an electric circuit:

1. Name the current in every branch.
2. For every component in the circuit (source, resistor, capacitor, etc), write the relationship between the current through the component and the voltage across the component (also known as the component's *I-V characteristic*).
3. Apply the junction rule at every junction. Dismiss redundant junction equations (in simple circuits you can often dismiss the last equation).
4. Apply the loop rule to enough loops to include every component. Every loop should include at least one component that is not in any of the other loops, otherwise it will yield a redundant loop equation.
5. Solve the resulting system of equations.

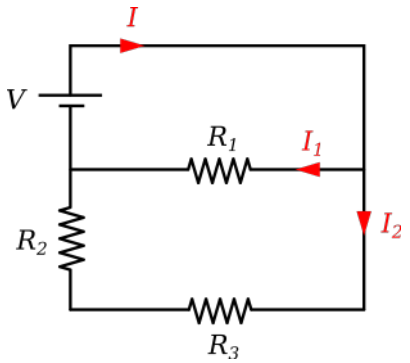
Example

Let's practice the rules above on the circuit from problem 37:



Problem 38: A two-loop circuit.

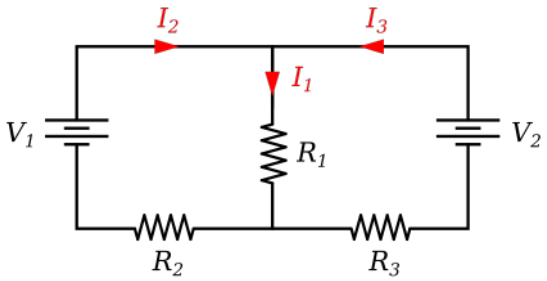
Compute the current in every branch of this circuit in terms of V, R_1, R_2, and R_3.



Problem 39

See problem 37 from `https://openstax.org/books/university-physics-volume-2/pages/10-problems`.
Notes:

- The components that look like two voltage source symbols on top of each other are just another way to represent voltage sources.
- The book's answer key assumes a specific current naming scheme that they don't share. Here it is:



- It's advantageous to plug in the numerical values before solving the linear system. The only parameter that can be useful to keep as a literal is V_2, because it allows you to get the currents in the second circuit without solving the linear system again.
- Sources deliver power. Resistors dissipate power. Therefore, the power dissipated by the circuit (question c) is the sum of the powers dissipated by all the resistors in the circuit, and the power delivered by the circuit (question d) is the sum of the powers delivered by all the sources in the circuit.

3.5.4 Equivalent resistances

Concept of equivalent component

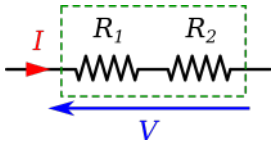
If you look at the general method for determining the voltages and currents in a circuit, it doesn't really matter what each component of the circuit is made of, or how it's built. The only property that matters is the component's I-V characteristic (the mathematical relationship between the current I going through the component and the voltage V across the component). For example, two resistors made from different materials but having the same resistance R have the same I-V characteristic: $V = R\,I$, therefore they are fully interchangeable: swapping one for the other in a circuit has absolutely no impact on what happens in the rest of the circuit, regardless of what the rest of the circuit is.

This concept of equivalent component extends to groups of components. If two arrangements of resistors have the same I-V characteristic, then they are said to be equivalent, even if their numbers of resistors,

resistance values, or geometries are different.

Problem 40: I-V characteristic of two resistors in series.

Two electronic components are said two be in series when they are on the same branch.

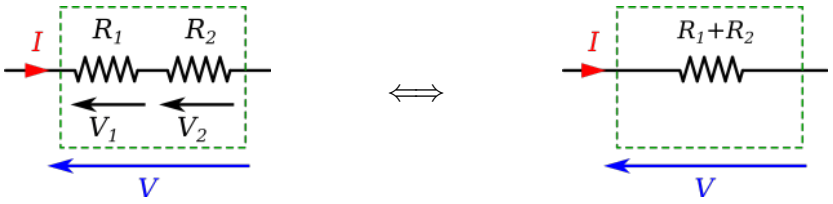


Imagine the two resistors in the sketch are encased in a box. That box has two wires sticking out, the one going into R_1 and the one coming out of R_2, which we can use to connect the box to a circuit. We can measure the current I going through the box, and the voltage V across the box. In other words, we can think of the box as a component in its own right.

What is that component's I-V characteristic? In other words, write V in terms of I, R_1, and R_2.

Equivalent resistance: Two resistors in series

The I-V characteristic we found in problem 40 is the same as that of a single resistor with resistance $R_1 + R_2$. Therefore, as far as the rest of the circuit is concerned, the two resistors R_1 and R_2 are indistinguishable from a single resistor $R_1 + R_2$. We say that $R_1 + R_2$ is the *equivalent resistance* of the two resistors in series.



The benefit of replacing R_1 and R_2 by $R_1 + R_2$ is that it simplifies the study of the circuit (the two resistors and whatever else they're connected to). As we'll see in a moment, there are more simplification rules which, when combined, allow us to solve problems much faster than we would have by applying the general rules.

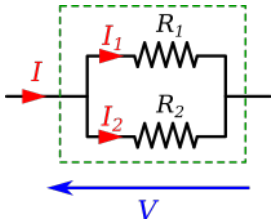
The one bit of information we loose by replacing R_1 and R_2 by $R_1 + R_2$ are the voltages V_1 and V_2, i.e., what happens inside the box. However, it's not as bad as it may sound. Imagine the two resistors are part of a circuit. We replaced them with a single resistor $R_1 + R_2$, then went on to predict V and I. Once we have V and I, we can use the usual rules to predict V_1 and V_2 through a sort of reverse problem **??**.

Problem 41

Compute I in the example from section 3.5.3 using an equivalent resistance.

Problem 42: I-V characteristic of two resistors in parallel.

Two electronic components are said two be in parallel when they are on separate branches that share the same end points.

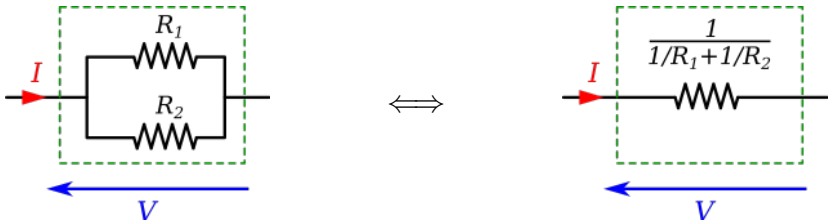


Write V in terms of I, R_1 and R_2.

Hint: Write the junction rule, then eliminate I_1 and I_2 using Ohm's law for R_1 and Ohm's law for R_2.

Equivalent resistance: Two resistors in parallel

As seen in problem 42, the I-V characteristic of two resistors in parallel is the same as that of a single resistor with resistance $\dfrac{1}{1/R_1 + 1/R_2} = \dfrac{R_1 R_2}{R_1 + R_2}$ (the two formulas are equivalent; which one is more convenient depends on the problem at hand).



Problem 43: Path of least resistance.

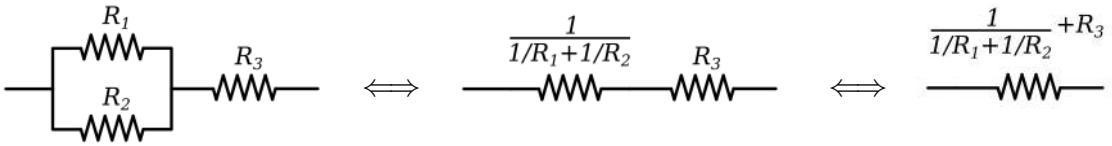
Show that when a current I has two resistive paths available to it (two resistors in parallel it can go through), the path with the lowest resistance has the largest current. The sketch from problem 42 may help.

Problem 44: Number of paths.

Show that adding more paths for the current (more resistors in parallel) always lowers the overall resistance.

Equivalent resistance: More than two resistors

The equivalency rules for resistors in series and in parallel are transitive, i.e., they can be applied multiple times. For example:



Problem 45

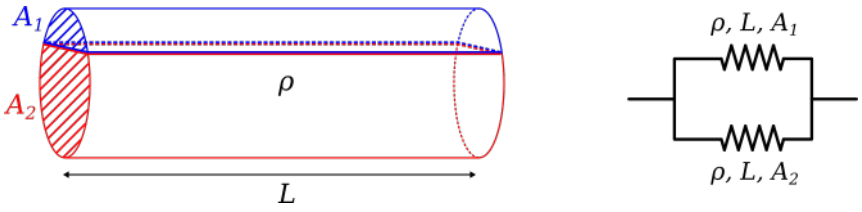
Compute I in problem 38 using a single resistor equivalent to all three resistors in the circuit.

Problem 46

See problem 69 from `https://openstax.org/books/university-physics-volume-2/pages/10-additional-problems`. Notes:

- The *e.m.f.* (electromotive force) of the voltage source is the same thing as its voltage. This problem denotes it ϵ rather than V.
- The *potential drop* across a resistor is the same thing as its voltage (measured opposite the current).
- The *power dissipated* by the resitor is the same thing as the power received by the resistor. The *power supplied* by a source is minus the power it receives.

Problem 47: Resistance: geometric formula vs parallel formula.



The sketch on the left shows a wire with resistivity ρ and length L, divided into an upper region (in blue) with cross sectional area A_1 and a lower region (in red) with cross sectional areas A_2. The charges that make up the current move parallel to the boundary between the two regions, therefore they do not cross from one region to the other. You may even say that the two regions behave like two distinct resistors. At the ends of the wire, where it connects to something else, those two resistors are connected to each other. In other words, they're in parallel (sketch on the right).

1. Use the formula for the resistance as a function of the resistivity and geometry to compute the resistance of the blue region, the resistance of the red region, and the resistance of the entire wire (blue region+red region).
2. Pretend you don't the resistance of the entire wire yet. Use the formula for the equivalent resistance of two resistors in parallel to compute the resistance of the entire wire from the resistances of the blue and red regions. Do you get the same result?

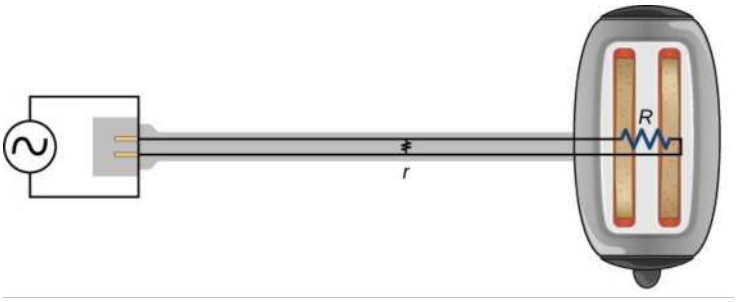
3.5.5 Application to household wiring and electrical safety

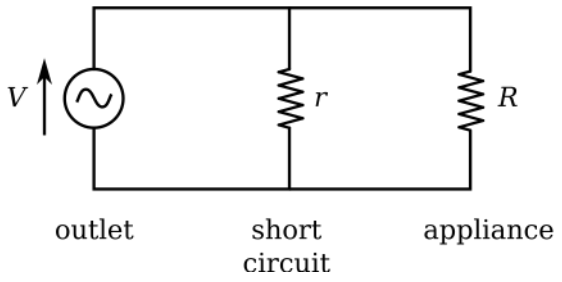
Note about AC current

We're about to discuss household circuits in which the source delivers an oscillating voltage $V(t)$ with amplitude $\approx$170 V and frequency 60 Hz (60 oscillations per second). The symbol for such a source is a circle with a tiny sinusoid inside. For the purpose of computing the power dissipated in resistor-based circuits, which is what this subsection is about, we can pretend that the source delivers a constant voltage $120V$ instead.

Problem 48: Voltage-controlled vs current-controlled dissipation. $P = \dfrac{V^2}{R}$ or $P = RI^2$?

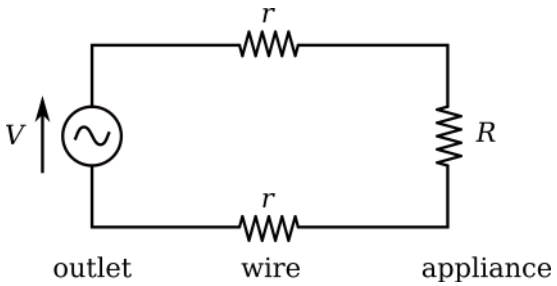
1. Power dissipated in a short-circuit.





   (a) Draw the circuit without short-circuit. Compute the current through the appliance, then the power delivered by the source and the power dissipated in the appliance.
   (b) In the circuit with the short-circuit, compute the current in each branch, then the power delivered by the source and the power dissipated in each resistor.
   (c) Summarize, in words, the impact of the short-circuit on the power delivered by the source and the power dissipated in the appliance.
   (d) Can we predict the power dissipated in the short-circuit if we only know the short-circuit's resistance r and the outlet's voltage V? What if we only know r and either I or I_R?

2. Power dissipated in a power cord.



This could be the same toaster, except there is no short-circuit here, and we no longer neglect the resistance of the wires in the power cord.

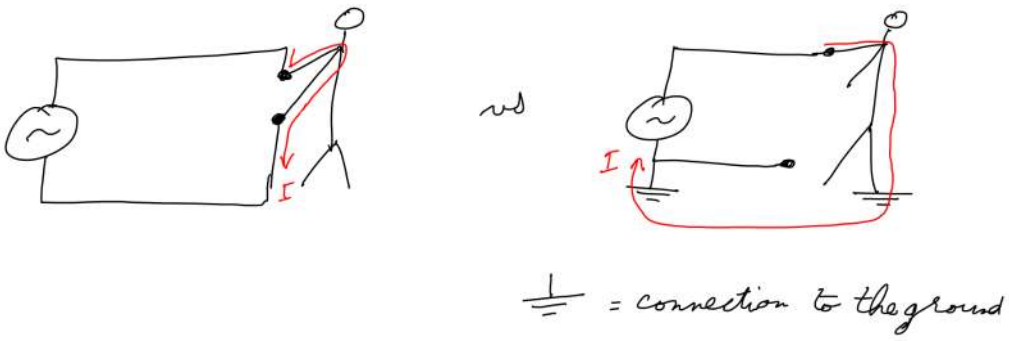
(a) Compute the current in the circuit. Use $r \ll R$ to simplify your answer. What do you notice?
(b) Compute the energy dissipated in the cord (the two r's).
(c) Can we predict the power dissipated in the cord if we only know the cord's resistance r and the outlet's voltage V? What if we only know r and the current I going through the appliance?
(d) If an appliance's cord is overheating, what specific property of the cord can we realistically change to address the problem?

Problem 49: Cord rating.

An appliance is plugged into a 120 V power outlet with a cord rated for 5 A. The appliance consumes 1000 W. Is the cord at risk of overheating?

The ground

The ground is a pretty good conductor, with a typical resistance of 1 Ω to 10 Ω. This is because the ground contains water and lots of ions, which gives it an ok to good conductivity (depending on the weather), and it has a very large cross sectional area (lots of different paths for the current to take). The ground is also part of most household electrical circuits. For reasons I won't delve into, it's beneficial for every outlet and every appliance to have one terminal connected to the ground. As a result, electric shocks often involve current passing through the ground:



The body

The human body also contains lots of water with lots of ions. The inside of the body has a low resistance, about 1 Ω. The main obstacle to the current's passage through the body ends up being the skin, whose resistance varies from about 5 Ω (wet) to about 100 Ω (dry).

Shock hazard

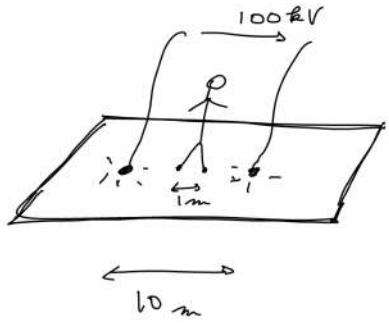
The most important quantity in assessing the risk posed by an electrical situation is the amount of current going through the body, followed by the current's frequency (every current has a frequency; for a constant current it's 0 Hz), its duration, and path through the body. Currents below 5 mA are normally harmless regardless of frequency and duration. Above that it depends on the situation.

When analyzing a potentially dangerous electrical situation, it's important to remember that the current depends on the resistance of everything the current has to go through. For example, the current going through the ground in the last sketch above may have to go through some wire ($< 1\,\Omega$), hand skin ($5\,\text{k}\Omega - 100\,\text{k}\Omega$), hand skin ($5\,\text{k}\Omega - 100\,\text{k}\Omega$), body interior ($\sim 1\,\Omega$), foot skin ($5\,\text{k}\Omega - 100\,\text{k}\Omega$), shoe sole (depends wildly on shoe material and dryness), concrete (depends wildly on the building), and finally the ground ($1\,\Omega - 10\,\Omega$). All those resistances are in series, therefore it's their sum that controls the current. With such a wide range

of orders of magnitude, in practice it's the parts of the circuit with the highest resistance that matter the most. In this example the footwear and the type of floor would likely be most important.

Problem 50: Downed power line.

The two ends of a broken high tension power line touch the ground 10 m apart. The voltage between the two broken ends is 100 kV. A person stands between the two broken ends with their feet 1 m apart. The resistance of the body (interior+skin+shoes) is 100 kΩ. To keep it simple we assume that the resistance between two points on the ground (e.g., a broken end and a foot) is equal to the distance between the two points times the ground's resistance per unit length $\lambda = 1\,\Omega/\mathrm{m}$.



1. Sketch the corresponding electric circuit.
2. Compute the current through the body. Is it above the safety threshold (5 mA)?

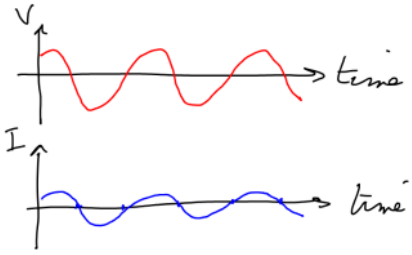
3.6 Time-dependent circuits

The voltage provided by standard power outlets is not constant. It oscillates sinusoidally. When the current and/or the voltage chage through time, computing quantities that add up through time, like the electrical work, requires an integral.

3.6.1 Ohm's law

$V = R\,I$ remains true when V and I vary through time.

Example:



V and I both vary, but $V/I = R$ is constant.

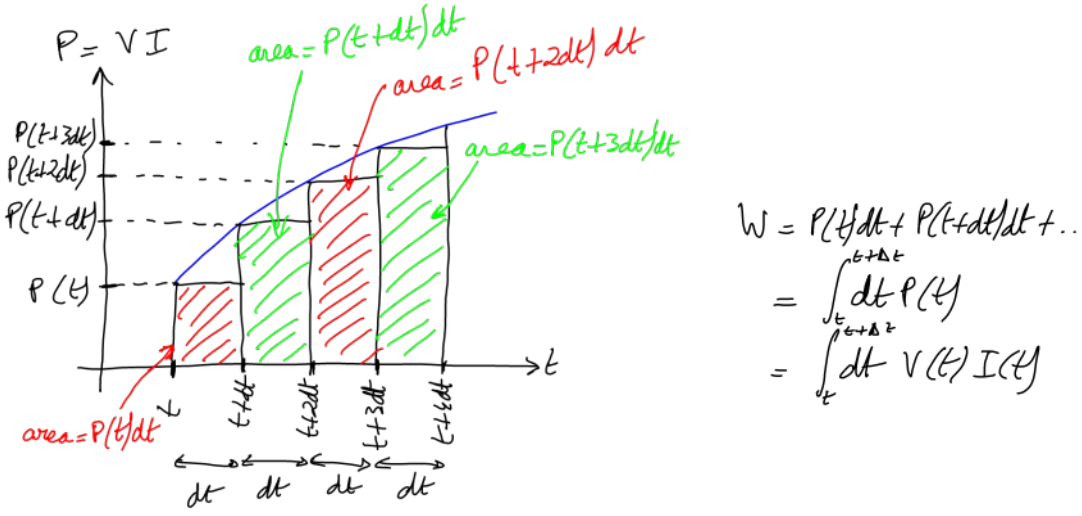
3.6.2 Work received

$W = VI\Delta t$ doesn't work well when V and I vary during the time interval Δt. Which of the many values V and I take during that time would you even plug into the formula?

On the other hand, when Δt is very small, then V and I don't have enough time to change much, so $W = VI\Delta t$ is approximately true. Since V and I don't change much during Δt, we can use their values at

any time during Δt; for example at the beginning of the interval. Going one step further, if Δt is *infinitely* small (in which case we often write dt instead of Δt), the formula is *infinitely* accurate, i.e., it's an exact formula rather than an approximate one.

If Δt is larger, the right way to compute the work received is to split Δt into infinitely many infinitely small time intervals dt. Each interval dt is short enough that we can compute the energy received between t and $t + dt$ as $V(t)I(t)dt$ (using the values of V and I at the beginning of the interval). To get the total energy received during *Deltat*, we add all of those individual contributions together. If you remember the Riemann sum definition of integrals, this is exactly what this is:



If you plot $P(t) = V(t)I(t)$ against time, $V(t)I(t)dt$ is the area of the rectangle with height $P(t)$ and width dt. If dt is small enough, it's the area under the curve between t and $t + dt$. Each rectangle's area represents the energy received during an interval dt. Splitting Δt into many such dt interval and adding all of their contributions amounts to computing the area under the curve, i.e., the integral of $P(t)$ with respect to time: $W = \int_t^{t+\Delta t} V(t')I(t')dt'$.

3.6.3 Power received

The power is still defined as $P = W/\Delta t$, however the way you compute it depends on Δt.

If Δt is infinitely small, then $W = VI\Delta t$, therefore $P = VI$. This is the *instantaneous power*. The formula is the same as before, but you have to keep in mind that P, V, I now depend on the time t, and $P(t) = V(t)I(t)$ only tells you about the amount of energy being received over an infinitely short period of time around t. Note that $P(t) = \lim_{\Delta t \to 0} \frac{W}{\Delta t}$ is also the derivative $\dfrac{dW}{dt}$ of W with respect to time. Conversely, P is the rate of change of W, i.e., the rate at which energy is received. This is not unlike the discussion we had in Physics 1 about the velocity being the rate of change of the position, then the acceleration being the rate of change of the velocity.

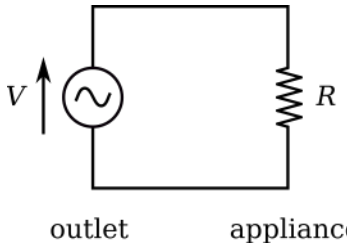
If Δt is not small, and W is measured between an initial time t_0 and a final time $t_0 + \Delta t$, then $P = W/\Delta t$ is the average power received during that time, i.e., the average rate at which energy was received during Δt:

$$P = \frac{W}{\Delta t} = \frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} V(t)I(t)dt.$$

This last form can also be interpreted as the average value of the instantaneous power $P(t) = V(t)I(t)$ between t_0 and $t_0 + \Delta t$. Overall the discussion of instantaneous vs average power is similar to the discussion of instantaneous vs average velocity or acceleration in Physics 1.

Problem 51: Power delivered by a wall outlet.

Standard wall outlets deliver a sinusoidal voltage $V(t) = V_0 \sin(2\pi f t)$ where $f = 60\,\text{Hz}$ is the frequency and $V_0 = 170\,\text{V}$ is the amplitude (in the US & Canada; different countries have different values). Say we plug an appliance with resistance R into such an outlet:



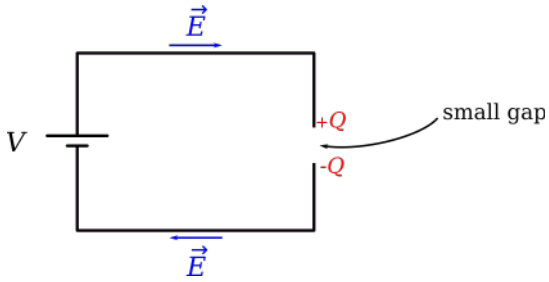
1. Sketch $V(t)$. What is its period T?
2. Compute the instantaneous power received by the resistor, then the energy received during a single period.
3. Imagine the oscillating source is replaced with a constant voltage source V_1. What does V_1 need to be to deliver the same energy over a time T as the the original source?
4. Why is the energy received over a single full period more significant than the energy received over a third, a quarter, etc of a period?

3.7 RC circuits

3.7.1 An exception to the junction rule

The basis for the junction rule is the fact that identical charges repel each other, therefore it takes a lot of effort to accumulate many of them in the same place. The more net charge there is in one place, the larger the repulsive force opposing the addition of new charges with the same sign. In practice this effect is so strong that in most cases only negligible amounts of charge can be accumulated, with the junction rule as consequence.

A notable exception is capacitors. On the most basic level, a capacitor is just a small gap in a circuit. The electric field $\vec{E}$ created by the battery pushes a small net positive charge Q to accumulate on the side of the gap connected to the positive terminal of the battery and the same amount of negative charge to accumulate on the side of the gap connected to the negative terminal of the battery.



As Q increases, this extra charge creates a repulsive force that makes it harder to bring more charges in. Eventually Q grows the the point of creating a counter-field that fully cancels the field $\vec{E}$ created by the battery, which causes the flow of charges to stop and Q to stabilize. As suggested in the first paragraph, this "saturation" effect is usually reached for such a low value of Q that we can ignore the phenomenon entirely. In some circumstances, however, Q can become significant. Three key factors can create such a situation:

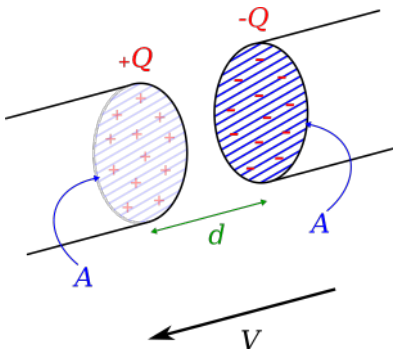
1. The size of the gap. What prevents bringing more positive charge to the positive side is the positive charge Q that's already there. However if the gap is very small, the repulsion caused by Q is partially offset by the attraction created by the charge $-Q$ on the other side of the gap. A similar reasoning applies to bringing more negative charge to the negative side. In both cases, the result is that a small gap helps build up more charge.
2. The more surface area there is for charges to accumulate, the further they can spread out, the weaker the repulsion. Thus a large surface area at the gap also helps build up a larger Q.

3. The larger the voltage, the stronger the electric field created by the battery, the stronger the force bringing charges to the edge of the gap in the first place.

To be more precise, if d is the size of the gap, A is the surface area at the edge of the gap, and V is the voltage measured across the gap, then Q is inversely proportional to d, proportional to A, and proportional to V:

$$Q \propto \frac{V\,A}{d}$$

("$\propto$" means "proportional to")



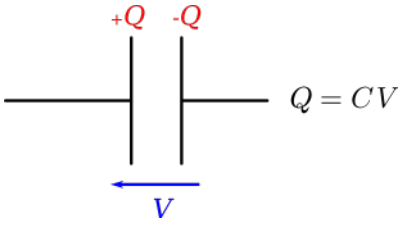
3.7.2 Capacitance

Although understanding the way the geometry of the gap controls Q is important to design capacitors, when it comes to studying circuits with capacitors in them the only thing that really matters is that Q is proportional to V. The corresponding coefficient of proportionality is called the *capacitance*, noted C:

$$Q = CV$$

The capacitance C is always a *positive* number. Its SI unit is the Farad, symbol F. If a capacitor's capacitance is $C = 1\,\mathrm{F}$, then applying a $1\,\mathrm{V}$ voltage to it will result in a charge $Q = +1\,\mathrm{C}$ on one side of the gap and $-1\,\mathrm{C}$ on the other. Most capacitors have a capacitance much smaller than $1\,\mathrm{F}$ though, so microFarads ($1\,\mu\mathrm{F} = 1 \times 10^{-6}\,\mathrm{F}$), nanoFarads ($1\,\mathrm{nF} = 1 \times 10^{-9}\,\mathrm{F}$), and even picoFarads ($1\,\mathrm{pF} = 1 \times 10^{-12}\,\mathrm{F}$) are much more common than full Farads.

As with Ohm's law and the power and energy formulas, the sign convention is important. $Q = CV$ assumes the voltage arrow points from $-Q$ to $+Q$:



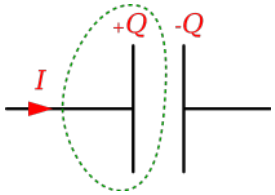
If V points the other way (toward $-Q$), the correct formula is $Q = -CV$.

3.7.3 I-V characteristic

As discussed in section 3.5.3, predicting currents and voltages in a circuit requires to know the I-V characteristic of every component. What we have so far for capacitors is a relationship between Q and V. Fortunately, the conservation of charge argument we used to obtain the junction rule provides us with a simple relationship between Q and I.

Charge conservation

Let's first review the junction rule argument. If I is constant, then the change of Q over a time Δt is $\Delta Q = I\Delta t$. This cannot go on forever, otherwise Q would become infinite. Therefore, I must be zero.



What's different now is that we no longer assume I is constant. In particular, I can be nonzero without building up an infinite amount of charge as long as it's only nonzero for a limited amount of time. I being allowed to change also means that we can no longer use $\Delta Q = I\Delta t$ as is. Instead we have to split Δt into a very large number of very small intervals dt, compute $dQ = I\Delta dt$ for each dt interval, then add all the dQ's together to get ΔQ. In the limit of an infinite amount of infinitely short dt intervals the formula becomes exact and the sum becomes an integral:

$$Q(t+\Delta t) - Q(t) = \int_t^{t+\Delta t} I(t')dt'$$

Deriving with respect to Δt on both sides yields

$$\frac{dQ}{dt} = I,$$

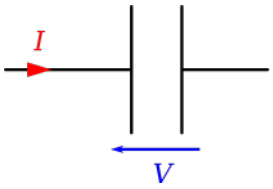
which is often more convenient to use. Another way to obtain this equation is to go back to $dQ = Idt$, which is valid when dt is very small even if Q varies, and divide both sides by dt to get $\frac{dQ}{dt} = I$. At this stage $\frac{dQ}{dt}$ is a division, but if you reinterpret it as a derivative you get the correct relationship between Q and I. It's not a very rigorous approach but it can help remember and/or make sense of the formula.

Sign convention

Again, the sign convention is important here. On the other hand there is no new "rule" to learn; the same logic we used to discuss charge conservation (or car conservation, or blood cell conservation) and the junction rule still applies. Namely, I is the only current entering the green region with charge Q, therefore the rate of change of Q is I: $dQ/dt = I$. If I pointed left, then it would exit the region with charge Q, therefore the rate of change of Q would be $-I$, which we would write $dQ/dt = -I$.

I-V characteristic of a capacitor

Combining $Q = CV$ and $I = dQ/dt$, we finally get the I-V characteristic of a capacitor:



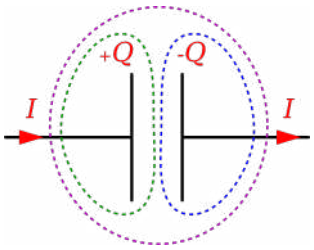
$$I = C\frac{dV}{dt}$$

where C is the capacitor's capacitance, I is the current going through the capacitor (it doesn't really go through, rather it stops on one side while an identical current leaves the other side, but from the point of view of the rest of the circuit it makes no difference), and V is the voltage across the capacitor with the voltage arrow pointing opposite the current arrow. If the voltage arrow points in the same direction as the current arrow then the I-V characteristic is $I = -C\frac{dV}{dt}$ instead.

Junction rule

Although neither side of the gap is bound by the junction rule (the sum of entering currents doesn't have to be 0 at all times), the capacitor as a whole still obeys the junction rule. Every time the charge changes on one side, the opposite change happens on the other side so that the charge of each sides remains minus the charge of the other side at all times. This implies that the same current must exist on both sides, which is what the usual junction rule predicts.

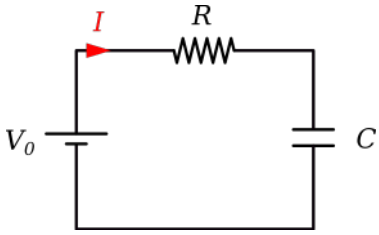
Here is another way to think about the same argument: since the sides of the gap have charges Q and $-Q$, the net charge of the capacitor is 0, which is constant. For the charge in the region of the circuit that contains the entire capacitor to remain constant, the sum of the currents entering the region must equal the sum of the current exiting it, i.e., the capacitor must obey the junction rule.



The significance of this is that we can apply the junction rule to circuits containing capacitors the same way we did in circuits without capacitors as long as we don't overthink what's happening inside the capacitor.

Problem 52: Charging a capacitor – Part 1.

Write the (differential) equation obeyed by the capacitor's charge Q. The equation should not contain V or I, only Q, C, R, and V_0.
Hint: Follow the general method from section 3.5.3, then use $I = dQ/dt$ to eliminate I in favor of Q.



3.7.4 Differential equations

The equation we obtained in problem 52 contains both the function $Q(t)$ and its derivative dQ/dt. This is known has a differential equation. More specifically, it is linear first-order differential equation with constant coefficients, meaning it has the form

$$\frac{dQ}{dt} = -\frac{1}{\tau}\left[Q(t) - Q_\infty\right]$$

where $Q(t)$ is the function we're trying to solve for, t is the variable this function depends on, and τ and Q_∞ are constants. As we will soon see, Q_∞ represents the value Q tends toward if we wait long enough while τ relates to the time it takes to reach (or at least get close to) the value Q_∞.

This type of equation shows up in a lot of different places in physics and mathematical modeling. We've actually already encountered it when writing Newton's second law for a falling object with viscous drag, except the function to solve for was the velocity $v(t)$.

Problem 53: Charging a capacitor – Part 2.

Write the equation we obtained in problem 52 in the form above, i.e., write τ and Q_∞ in terms of R, C, and V_0.

Problem 54: Fall with drag.

Write Newton's second law for an object of mass m and velocity $\vec{v}$ under the effect of its own weight $m\vec{g}$ where $\vec{g}$ is the acceleration of gravity and a viscous drag force $\vec{F} = -\alpha\vec{v}$ where α is the drag coefficient (a positive constant).
Assume the motion is one-dimensional along the z axis and $\vec{g} = -g\vec{z}$. Write the differential equation obeyed by the z component of the velocity, v_z. Put it in the generic form $\dfrac{dv_z}{dt} = -\dfrac{1}{\tau}[v_z(t) - v_\infty]$, i.e., write τ and v_∞ in terms of m, g, and α.

Qualitative analysis

A lot of information can be extracted from the differential equation without actually solving it. dQ/dt is the slope (or rate of change) of $Q(t)$. By analyzing its sign, specifically, the way its sign depends on the current value of Q, we can understand under what circumstances $Q(t)$ increases or decreases.

Looking at the general form $\dfrac{dQ}{dt} = -\dfrac{1}{\tau}[Q(t) - Q_\infty]$, we can see that:

- When $Q(t) = Q_\infty$, then $dQ/dt = 0$, therefore Q stays constant. In other words, once $Q(t)$ has reached the value Q_∞, it stops changing and stays at Q_∞ for ever. We'll call that the *equilibrium value*.
- If $Q(t) < Q_\infty$, then $dQ/dt > 0$, therefore Q increases, which gets it closer to Q_∞.
- If $Q(t) > Q_\infty$, then $dQ/dt < 0$, therefore Q decreases, which also gets it closer to Q_∞.

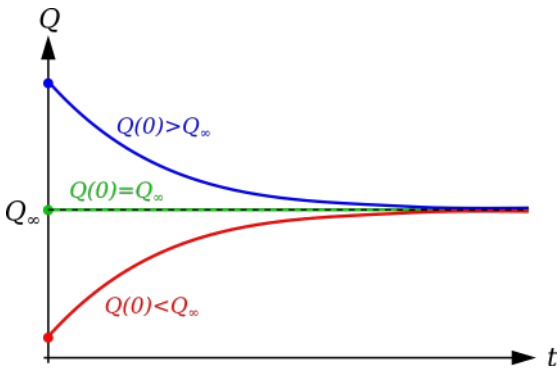
In summary, $Q(t)$ always changes to be closer to Q_∞, unless it's already at Q_∞ in which case it stops changing. If we wait long enough ($t \to \infty$), Q eventually reaches Q_∞ and stays there. For this reason Q_∞ we call Q_∞ the *equilibrium value*, or sometimes the *steady-state value*, of Q. We also say that Q relaxes toward Q_∞.

Another way to phrase this is to say that the distance between Q and its equilibrium value Q_∞ always decreases. Additionally, the rate at which this distance decreases is proportional to the distance itself: $\dfrac{d}{dt}(Q - Q_\infty) = -\dfrac{1}{\tau}(Q - Q_\infty)$. Therefore, the speed at which Q gets closer to Q_∞ slows down with time as Q gets closer to Q_∞.

Finally, the larger τ, the smaller dQ/dt, the slower the relaxation occurs. This is why τ, which has dimensions of a time, is called the *relaxation time* or the *time constant*.

Even though we haven't solved the differential equation yet, we now have enough information to sketch the solution, or rather some solutions, since the exact solution depends on the initial value of V (the initial condition):

- If $Q(0) > Q_\infty$ (Q starts above Q_∞; blue curve), then Q decreases toward Q_∞, first quickly, then slowly, until it reaches it.
- If $Q(0) < Q_\infty$ (Q starts below Q_∞; red curve), then Q increases toward Q_∞, first quickly, then slowly, until it reaches it.
- If $Q(0) = Q_\infty$ (Q starts at Q_∞; green curve), then Q remains constant and equal to Q_∞.



Full solution

The general solution of $\dfrac{dQ}{dt} = -\dfrac{1}{\tau}[Q(t) - Q_\infty]$ is:

$$Q(t) = Q_\infty + [Q(0) - Q_\infty]\, e^{-t/\tau}$$

where $Q(0)$ is the value of Q at $t = 0$ (in other words, the initial condition).

Here are some basic properties of this formula which I think (1) help makes sense of it, (2) as a result, help remember it, and (3) show that it's consistent with the insight gained from the qualitative analysis:

- By definition, $Q(0)$ is the value of Q at $t = 0$. Plugging $t = 0$ into the formula we get $e^{-t/\tau} = e^0 = 1$, therefore $Q(t) = Q_\infty + [Q(0) - Q_\infty] = Q(0)$. This is not a prediction, rather a self-consistency check. I also find that understand how this plays out helps remember the structure of the solution formula.
- In the $t \to \infty$ limit (long times), $e^{-t/\tau} = e^{-\infty} = 0$, therefore $Q(0) = Q_\infty$. As expected from the qualitative analysis, $Q(t)$ relaxes toward Q_∞. It never actually reaches it, but it gets infinitely close as time goes to infinity.
- We can check that the solution does obey the differential equation by computing the derivative of $Q(t)$:

$$\frac{dQ}{dt} = 0 + [Q(0) - Q_\infty] \times \left(-\frac{1}{\tau}e^{-t/\tau}\right) = -\frac{1}{\tau}\left([Q(0) - Q_\infty]e^{-t/\tau}\right) = -\frac{1}{\tau}\left[Q(t) - Q_\infty\right]$$

- The relaxation time (or time constant) τ controls how quickly $e^{-t/\tau}$ goes from 1 to 0, which determines how quickly Q goes from $Q(0)$ (when $e^{-t/\tau} = 1$) to Q_∞ (when $e^{-t/\tau} = 0$). To be more specific, let's think about the relative distance to the equilibrium value, i.e., how far $Q(t)$ is from the equilibrium value Q_∞ divided by how far it was initially. Let's call that α:

$$\alpha(t) \equiv \frac{Q(t) - Q_\infty}{Q(0) - Q_\infty} = \frac{[Q(0) - Q_\infty]e^{-t/\tau}}{Q(0) - Q_\infty} = e^{-t/\tau}$$

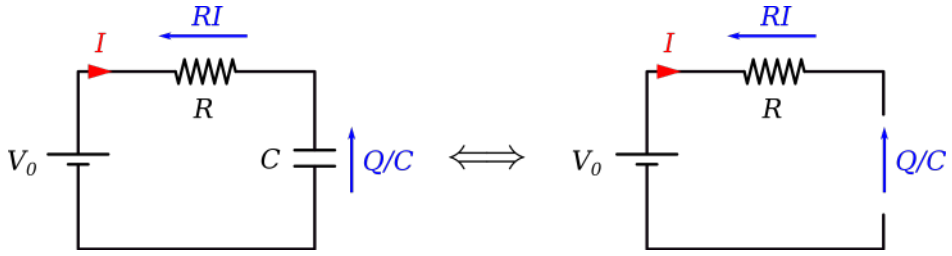
  Expectedly $\alpha(0) = 1$: when $t = 0$ the distance is the initial distance so the ratio is 1. Also expectedly $\alpha(+\infty) = 0$: if you wait forever ($t \to \infty$), Q gets infinitely close to Q_∞. Additionally, the meaning of τ becomes clearer: $\alpha(\tau) = e^{-1} \approx \frac{1}{2.718} \approx 0.37$, $\alpha(2\tau) = e^{-2} \approx 0.14$, etc. More generally, for any time t, $\frac{Q(t+\tau) - Q_\infty}{Q(t) - Q_\infty} = e^{-1}$, i.e., every time we wait for one extra relaxation time τ, the distance between Q and its equilibrium value Q_∞ gets divided by $e \approx 2.718$.

Problem 55: Charging a capacitor – Part 3.

1. The capacitor from problems 52 and 53 is initially uncharged ($Q(0) = 0$). Solve the differential equation to get $Q(t)$ in terms of R, C, V_0, and t. Sketch the graph of $Q(t)$.
2. How long does it take Q to reach 99% of its equilibrium value?

3.7.5 Final state analysis

A lot can be understood about an RC circuit without ever solving, or even writing, a differential equation. When the capacitor's charge has stabilized (Q no longer depends on time), the current through it is $I = \frac{dQ}{dt} = 0$, same as a regular gap in the circuit (one that does not act as a capacitor). Therefore, we can predict the eventual currents and voltages in the circuit (the ones attained once Q no longer depends on time) by studying instead a circuit in which the the capacitor has been removed (replaced with a gap):



Problem 56: Charging a capacitor – Final state analysis.

Use final state analysis to predict the final charge (the one we called Q_∞ in the differential equation).

3.7.6 Capacitors and energy

In section 3.6 we saw that the electrical energy received by a circuit component over a period of time, say from a time t_1 to a later time t_2, is $W = \int_{t_1}^{t_2} V(t)I(t)$ where I is the current going through the component and V is the voltage across the component, with the voltage arrow pointing opposite the current arrow. This applies to capacitors as well.

To gain insight into the way capacitors receive energy, it is useful to rewrite V and I in terms of Q:

$$W = \int_{t_1}^{t_2} V(t)I(t) = \int_{t_1}^{t_2} \left(\frac{Q(t)}{C}\right)\left(\frac{dQ}{dt}\right) = \frac{1}{2C}\int_{t_1}^{t_2} 2\frac{dQ}{dt}\ Q(t)$$

Per the chain rule, $2(dQ/dt)Q(t)$ is the derivative of $Q(t)^2$ with respect to time (you may be more familiar with this in the form $[u^2]' = 2u'u$ or something like that), therefore

$$W = \frac{1}{2C}\Big[Q(t)^2\Big]_{t_1}^{t_2} = \frac{Q(t_2)^2}{2C} - \frac{Q(t_1)^2}{2C}$$

What's remarkable about this result is that the amount of energy received in changing the capacitor's charge from $Q(t_1)$ to $Q(t_2)$ doesn't depend on the details of how the charge was changed (how much current, for how long, etc). It only depends on the initial state of the capacitor (its initial charge $Q(t_1)$) and its final state (its final charge $Q(t_2)$).

Imagine a capacitor has an initial charge Q_1. To raise its charge to Q_2 with $Q_2 > Q_1$, you need to provide an energy $(Q_2^2 - Q_1^2)/(2C)$, which will come from the source of electricity in the circuit. Now imagine the capacitor sends this extra charge back into the circuit. The energy it receives in going back from Q_2 to Q_1 is $(Q_1^2 - Q_1 2^2)/(2C)$, which is exactly minus the energy it received to go from Q_1 to Q_2. In other words, in going from the capacitor releases the entirety of the energy it had received. That's very different from resistors. The energy received by a resistor turns into heat, i.e., thermal agitation of the atoms in the resistor and the surroundings. It cannot be retrieved (not easily anyway, and definitely not entirely). In contrast, capacitors hold on to the electrical energy they receive when they receive charge and release it later when release that charge.

This is very reminiscent of potential energy. It is in fact the same thing, in a different context. The defining feature of a conservative force is that the work received from it only depends on the initial state (initial position) and the final state (final position). A corollary of it is that returning to the initial state releases the exact same amount of energy that was received on the way to the final state. This is the basis for saying that the energy was in some way "stored" in the object, ready to be returned. This stored energy is what we call potential energy.

In summary, a capacitor with capacitance C and charge Q "contains" a potential energy $U = Q^2/(2C)$. The electrical energy received by the capacitor over a period of time during which its charge changes from Q_1 to Q_2 is

$$W = \int_{t_1}^{t_2} V(t)I(t)dt = U_2 - U_1 \quad \text{where} \quad U = \frac{Q^2}{2C}$$

Sometimes it is more convenient to write U in terms of V rather than Q. This is easily done by substituting $Q = CV$ in $U = \dfrac{Q^2}{2C}$, which yields $U = \dfrac{CV^2}{2}$.

Problem 57: Charging a capacitor – Energetic aspect.

An initially uncharged ($Q(0) = 0$) capacitor with capacitance C is charge by a voltage source V_0 through a resistor with resistance R (see, e.g., figure in section 3.7.5).

1. How much energy does the capacitor store once fully charged? Write you answer in terms of the parameters of the problem (C, V_0, R).

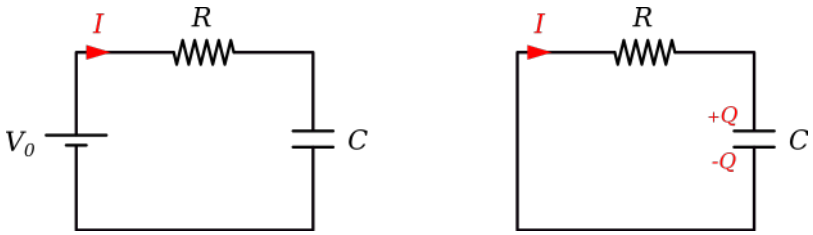
2. How long does it take for the capacitor to accumulate half of the energy it will have once fully charged?

3. Over the entire course of charging the capacitor, from $t = 0$ when $Q = 0$ to $t = \infty$ when $Q = CV_0$, how much energy gets delivered by the voltage source? How much get dissipated (lost) in the resistor?

4. Based on (1) the last question and (2) the time it takes to reach 99% charge (computed in *Charging a capacitor – Part 3*), what value of R seems optimal? Can you think of a potential downside to this choice?

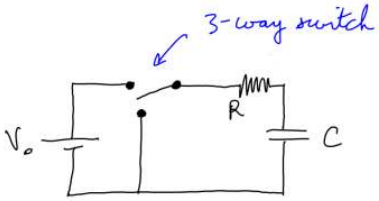
Problem 58: Discharging a capacitor.

The sketch on the left shows the circuit from the "Charging a capacitor" problems. Once the capacitor is fully charged ($Q = CV_0$), we swap the voltage source for a wire and start the clock. The resulting circuit is shown on the right. We count time from the swap, i.e., the swap happens at $t = 0$. The capacitor's charge Q is the same right after the swap as it was right before the swap.

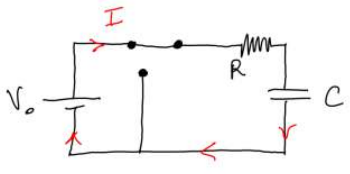


1. Use final state analysis to predict the final charge $\lim_{t\to\infty} Q(t)$.
2. Use the loop rule to obtain the differential equation obeyed by $Q(t)$.
3. Solve it. Write $Q(t)$ as a function of V_0, C, R, and t. Sketch $Q(t)$. Compare $Q(\infty)$ with your final state analysis.
4. Compute the current $I(t)$. What is the initial current? Where does it appear in the graph of $Q(t)$? Thinking about the capacitor as a voltage source of sorts, explain why it makes sense that the current would have the sign it has.
5. How long would it take to fully discharge the capacitor if the current remained equal to its initial value the whole time? In reality it takes longer, but this provides a reasonable order of magnitude.
6. Compute the energy released by the capacitor over the entire discharge, first by integrating VI, then using the formula for the energy stored in the capacitor. They should be the same.

Note: In practice one would use a switch to "swap" the voltage source with a wire without actually moving the voltage source or the wire:

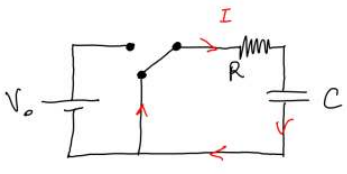


The 3-way switch can switch between "up" ( ) and "down" ( ).



"up"

The capacitor charges through R.



"down"

The capacitor discharges into R.

Chapter 4: Magnetism

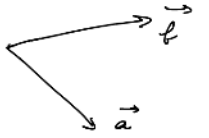
Cross Product

The cross product is a type of multiplication where you multiply two vectors to obtain a third vector. It is related to the dot product, but it is its own thing. The most obvious difference is that the dot product of two vectors is a scalar whereas the cross product of two vectors is a vector.

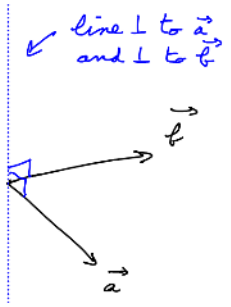
The cross product of two vectors $\vec{a}$ and $\vec{b}$ is noted $\vec{a} \times \vec{b}$.

- First go reread section 1.5 of "math.pdf" about the dot product. We'll need both.

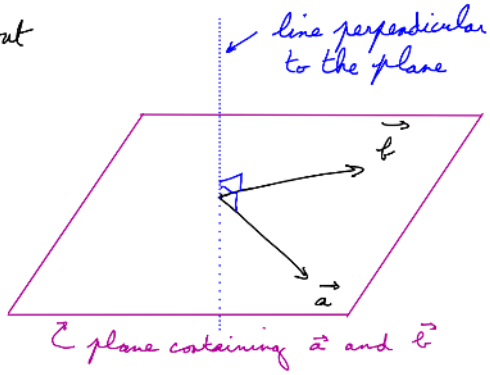
- Geometrical definition:
  Here is how to construct the cross product $\vec{a} \times \vec{b}$ of any two vectors $\vec{a}$ and $\vec{b}$:



① $\vec{a} \times \vec{b}$ is perpendicular to both $\vec{a}$ and $\vec{b}$.
There's only one direction that is perpendicular to both. $\vec{a} \times \vec{b}$ is along that line.

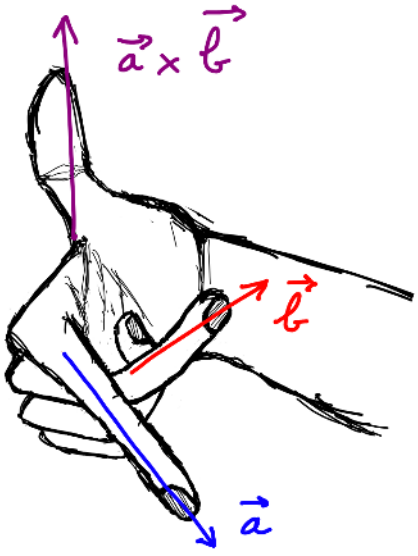


It may help to first think about the plane containing $\vec{a}$ and $\vec{b}$. The (only) direction that is perpendicular to both $\vec{a}$ and $\vec{b}$ is perpendicular to that plane.



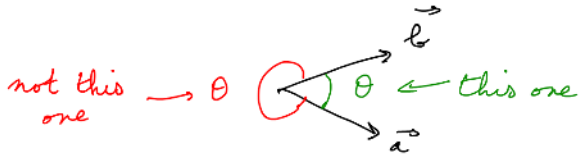
② At this point there are still two options for the direction as $\vec{a} \times \vec{b}$ could be pointing either side along the blue line. That's where the right hand rule comes in.

1. Point your right index finger along $\vec{a}$.
2. Point your right middle finger along $\vec{b}$. Depending on $\vec{a}$ and $\vec{b}$ you may have to contort your hand a bit.
3. Stick your right thumb out. That's the direction of $\vec{a} \times \vec{b}$.



Note: if you use your left hand, you will get the wrong direction. Same thing if you swap index and middle finger.

③ The magnitude of $\vec{a}\times\vec{b}$ is $\|\vec{a}\times\vec{b}\| = \|\vec{a}\|\cdot\|\vec{b}\|\cdot\sin\theta$ where θ is the angle between $\vec{a}$ and $\vec{b}$ (the smaller one).

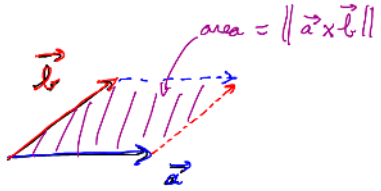


Special case: If $\vec{a}$ and $\vec{b}$ are parallel then θ is either 0 (if they point in the same direction) or π (if they point in opposite directions), therefore $\sin\theta = 0 \Rightarrow \|\vec{a}\times\vec{b}\| = 0$ $\Rightarrow \vec{a}\times\vec{b} = \vec{0}$. In other words the cross product of two parallel vectors is zero.

In that case you can skip the steps above. In fact you have to because when $\vec{a}\parallel\vec{b}$ there's an infinity of directions that $\perp$ to both $\vec{a}$ and $\vec{b}$ (any direction in the plane $\perp\vec{a}$ which is also $\perp\vec{b}$), so step ① doesn't work.

- Comments:

* $\|\vec{a}\times\vec{b}\| = \|\vec{a}\|\cdot\|\vec{b}\|\cdot\sin\theta$ is the area of the parallelogram formed by $\vec{a}$ and $\vec{b}$.



If $\vec{a}\parallel\vec{b}$ then the parallelogram collapses and its area is 0, consistent with $\vec{a}\times\vec{b} = \vec{0}$.

* Magnitudes of the cross product vs dot product:
  $\vec{a}\cdot\vec{b} = \|\vec{a}\|\cdot\|\vec{b}\|\cdot\cos\theta$ is maximal (equal to the product of the two vectors' magnitudes $\|\vec{a}\|\cdot\|\vec{b}\|$) when $\vec{a}$ and $\vec{b}$ are aligned, zero when they are perpendicular, and minimal (equal to $-\|\vec{a}\|\|\vec{b}\|$) when they are anti-aligned.
  $\|\vec{a}\times\vec{b}\| = \|\vec{a}\|\cdot\|\vec{b}\|\cdot\sin\theta$ is maximal (equal to $\|\vec{a}\|\cdot\|\vec{b}\|$) when $\vec{a}$ and $\vec{b}$ are perpendicular and zero when they are parallel (whether aligned or antialigned).
* The order of $\vec{a}$ and $\vec{b}$ matters: $\vec{b}\times\vec{a} = -\vec{a}\times\vec{b}$.
  (check it with the right-hand rule).

- Component definition: (not essential)
  We're not going to use this definition in this chapter, but it would definitely come up if we had a couple more weeks to spend on magnetism, and many of you have probably been exposed to it already.

$$\vec{a}\times\vec{b} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \times \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} a_y b_z - a_z b_y \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{pmatrix}$$

  It doesn't look like it, but this definition is fully equivalent to the geometrical one.

Magnetic interaction

- **Electric force vs magnetic force:**

A charge q_1 with position vector $\vec{r}_1$ and a charge q_2 with position vector $\vec{r}_2$ exert electric forces on each other. The force exerted by q_1 on q_2 is:

$$\vec{F}_{12}^{\,e} = k_e q_2 q_1 \frac{\vec{r}_{12}}{r_{12}^3} \qquad \text{where } \vec{r}_{12} = \vec{r}_2 - \vec{r}_1 \text{ goes from } q_1 \text{ to } q_2.$$

It is often useful to view this as a two-step process where

① q_1 creates an electric field $\vec{E}_1(\vec{r}_2) = k_e q_1 \frac{\vec{r}_{12}}{r_{12}^3}$ at $\vec{r}_2$.

② q_2 feels a force $\vec{F}_{12}^{\,e} = q_2 \vec{E}_1(\vec{r}_2)$.

If the charges are moving with velocities $\vec{v}_1$ and $\vec{v}_2$ respectively, there's a second force exerted by q_1 on q_2 called the magnetic force:

$$\vec{F}_{12}^{\,m} = k_m q_2 \vec{v}_2 \times \left(q_1 \vec{v}_1 \times \frac{\vec{r}_{12}}{r_{12}^3} \right) \qquad \text{where } k_m \text{ is a universal constant.}$$

Again it's often useful to view this as a two step process:

① q_1 creates a magnetic field $\vec{B}_1(\vec{r}_2) = k_m q_1 \vec{v}_1 \times \frac{\vec{r}_{12}}{r_{12}^3}$ at $\vec{r}_2$.

② q_2 feels a force $\vec{F}_{12}^{\,m} = q_2 \vec{v}_2 \times \vec{B}_1(\vec{r}_2)$.

Note: Many texts write $\frac{1}{4\pi\epsilon_0}$ instead of k_e and/or $\frac{\mu_0}{4\pi}$ instead of k_m. It doesn't matter for us, they're just fundamental constants with some value.

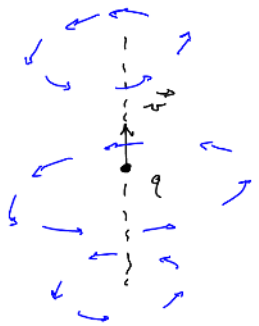
The magnetic force and the magnetic field obey the superposition principle just like their electric counterparts.

The sum of the electric and the magnetic force felt by a charge q with position $\vec{r}$ and velocity $\vec{v}$ is $\vec{F} = q\vec{E}(\vec{r}) + q\vec{v} \times \vec{B}(\vec{r})$ where $\vec{E}(\vec{r})$ and $\vec{B}(\vec{r})$ are the electric and the magnetic field created at $\vec{r}$ by all the other charges in the system. It is know as the electromagnetic force or the Lorentz force.

- Problem 1:

Show that the magnetic field of a moving charge swirls around the axis defined by $\vec{r}$ and $\vec{v}$.

In other words, show that the magnetic field lines are circles centered on and perpendicular to the axis.

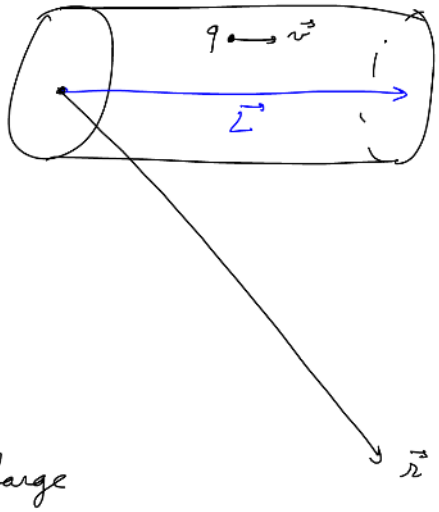


Hint: Construct the magnetic field at a few points around q using $\vec{B} = k_m q \vec{v}_1 \times \frac{\vec{r}_{12}}{r_{12}^3}$ where $\vec{r}_1$ = position of the charge and $\vec{r}_2$ = point where we compute $\vec{B}$. Start in the plane perpendicular to $\vec{v}_1$.

- Magnetic field created by a current:

Electric current consists of moving charged particles, therefore it creates a magnetic field.

Consider a small stretch of wire with cross section area A and end-to-end vector $\vec{L}$. It contains a density n of charges q moving at $\vec{v}$.



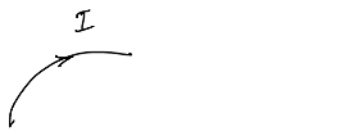
We want to compute $\vec{B}$ at position $\vec{r}$ measured from the basis of the wire.

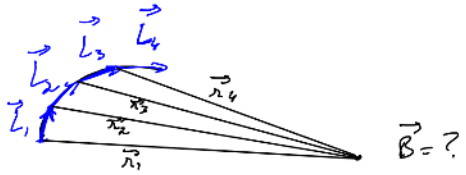
The total moving charge in the wire is $Q = qnAL$. Assuming the wire is small compared to r, the vector from any moving charge in the wire to the point where we want to compute the field is $\simeq \vec{r}$. Therefore the magnetic field of those charges is $\vec{B} = k_m Q \vec{v} \times \frac{\vec{r}}{r^3}$.

Since $Q\vec{v} = qnAL\vec{v} = (qnAv)\vec{L} \equiv I\vec{L}$, we can write $\boxed{\vec{B} = k_m I \vec{L} \times \frac{\vec{r}}{r^3}}$.

If the wire is too long then the vector going from an individual charge to the point where we want to compute the field varies along the wire so we can't just use $\vec{r}$ for everyone.

The solution is to divide the wire into a large number of very small bits, each with a different $\vec{r}$, compute the field created by each, then add them all up.





- **Problem 2:** Magnetic field of a current loop.

The drawing show magnetic field lines for a square current loop with current I. The field is roughly symmetric around the vertical axis through the center of the loop.

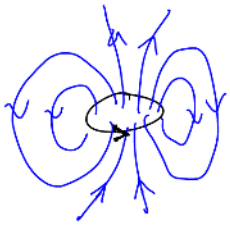
Discuss the direction of the field at each red dot using graphical vector addition.

Explain how they can be understood by adding up the field created by each of the four segments of the loop.

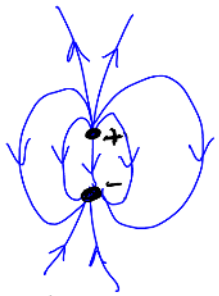
Hint: Like the electric field, the magnetic field decreases when the distance to the source increases. If two bits of wire create opposing fields, the closest one wins. We used similar reasoning to understand electric dipoles.

- Magnetic moment:

A small current loop is known as a magnetic dipole. The magnetic field far from the loop has the exact same form as the electric field far from an electric dipole.



magnetic



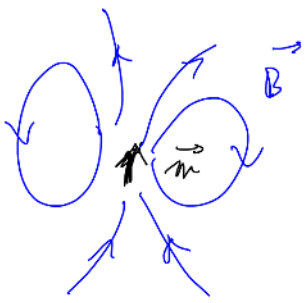
electric

Inside the dipoles however the fields point in opposite directions.

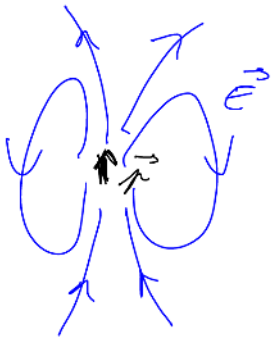
Earlier in the class discuss electric dipoles and the electric dipole moment $\vec{p} \equiv q\,\vec{d}$.

For a current loop with surface area A and current I. the magnetic dipole moment is defined as $\vec{m} = I A \hat{n}$ where $\hat{n}$ is a unit vector perpendicular to the loop whose direction is based on the direction of the current arrow using the "second right hand rule" (called like that in the textbook; not sure how widespread the name is): wrap the four fingers of your right hand around the loop pointing in the direction of the current arrow, then your thumb indicates which side $\hat{n}$ is on.

$\vec{m} = I A \hat{n}$

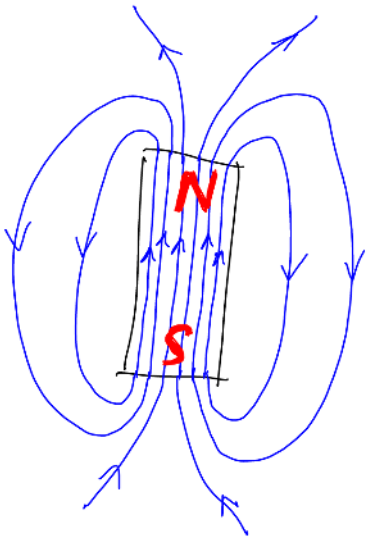


$\vec{p} = q\,\vec{d}$



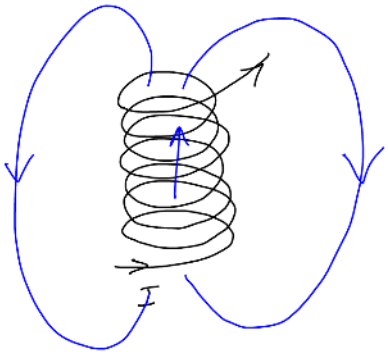
- Magnets:

  A magnet is essentially a big magnetic dipole whose dipole moment point from its south pole to its north pole.



- Solenoids:

  A solenoid is a stack of current loops made, e.g., by wrapping a wire around a cylinder. It creates a field very similar to a magnet. The field inside is fairly uniform. In fact this is a very common way to create a controlled uniform magnetic field (see MRI).



- **Force on a current carrying wire:**

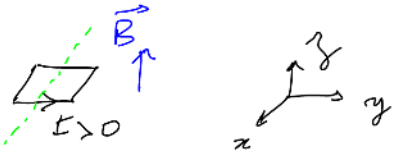
Earlier we showed that the magnetic field created by a small current carrying wire follows from the magnetic field created by a single moving charge by replacing $q\vec{v}$ with $I\vec{L}$:

$$\vec{B} = k_m\, q\vec{v} \times \frac{\hat{r}}{r^3} \longrightarrow \vec{B} = k_m\, I\vec{L} \times \frac{\hat{r}}{r^3}$$

Similarly the magnetic force on a current-carrying wire subject to a magnetic field follows from the single particle case:

$$\vec{F} = q\vec{v} \times \vec{B} \longrightarrow \vec{F} = I\vec{L} \times \vec{B}$$

- **Problem 3:** Magnetic torque on a current loop.



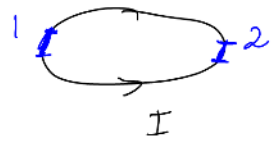
The square loop is in the xy plane. $\vec{B}$ is along z.

① Compute the magnetic force on each segment of the square loop. Discuss the net force and the net torque on it.

② The loop is rotated by an angle θ around the green dashed line. How does the torque depend on θ? Show that the torque always pushes the loop towards $\theta = 0$.

- Problem 4:

  What magnetic force do opposite ends of a current loop exert on each other? (qualitatively)



Sketch the forces between 1 and 2.

- Magnetic potential energy:

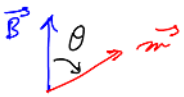
  An electric dipole with dipole moment $\vec{p}$ in an electric field $\vec{E}(\vec{r})$ has a potential energy $U = -\vec{p} \cdot \vec{E}$.

  Similarly, a magnetic dipole with dipole moment $\vec{m}$ in a magnetic field $\vec{B}$ has a potential energy $U = -\vec{m} \cdot \vec{B}$.

  As a result, all the phenomenology of dipole-field and dipole-dipole interactions we discussed in the context of electric interactions applies here as well:

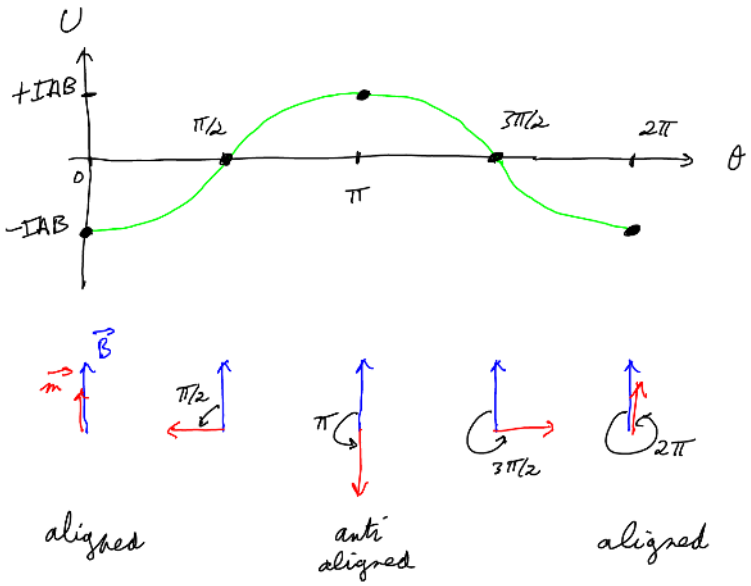
  * Dipoles align with field.
  * Once aligned, dipoles are attracted to stronger fields.
  * ...

In particular this applies to the situation of problem 3. Here $\vec{B}$ is fixed. As we discussed in the electric case, to get the torque we analyze U as a function of the orientation of the loop.



$$U = -\vec{m}\cdot\vec{B} = -IAB\cos\theta$$

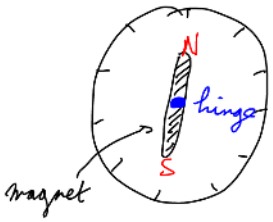
In this case what matters is the angle θ between $\vec{m}$ and $\vec{B}$.



The torque on the loop always pushes towards lower energy. Between $\theta=0$ and $\theta=\pi$, $U(\theta)$ increases, therefore decreasing θ decreases the energy, therefore the torque acts to decrease θ. Between $\theta=\pi$ and $\theta=2\pi$, $U(\theta)$ decreases, therefore increasing θ decreases U, therefore the torque acts to increase θ.

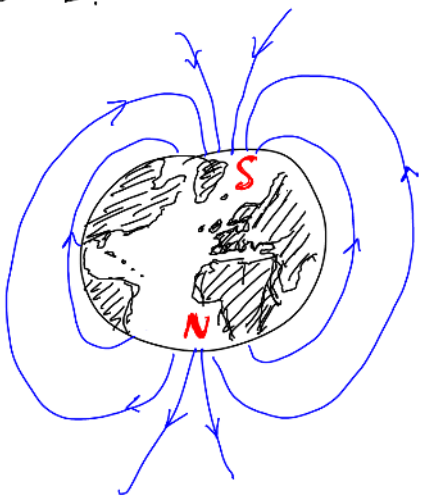
Note that $\theta = 0$ and $\theta = 2\pi$ are the same configuration. It's the aligned configuration, and it has the lowest energy. The torque always acts to rotate the loop towards it on whichever side is closest.

An application of this is compasses. They consist of a small magnet, i.e., a magnetic dipole, that is free to rotate. It aligns with the magnetic field $\vec{B}$ so as to minimize $U = -\vec{m} \cdot \vec{B}$, thus offering a way to visualize the direction of $\vec{B}$.



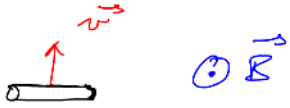
N and S refer to the north an south pole of the magnet. The dipole moment points from S to N so that's also the direction of the surrounding field $\vec{B}$.

- Earth magnetic field:
  The earth is also essentially a magnetic dipole. The geographical north pole is the south pole of the corresponding magnet.



- Problem 5: Interaction between 2 magnets.
  Explain as much as possible about magnet-magnet interactions by modelling each magnet as a magnetic dipole.

- Problem 6: Motional emf.
  ① A straight piece of conductor moves at velocity $\vec{v}$ in a uniform magnetic field $\vec{B}$. Sketch the force on the positive charges in the conductor and the one on the negative charges.



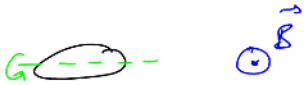
  The conductor is neutral (same number of + and - charges). What is the net force on the conductor? The free charge on the conductor are negative. How do they move within the conductor? How is charge distributed in the conductor once things have stabilized? Can there be a steady current in the conductor?

(2) Instead of a straigh wire it's now a circular wire moving at $\vec{v}$ in $\vec{B}$.

Sketch the magnetic force on the negative free charges at a few locations along the loop.

Can the magnetic force give rise to a current around the loop ?.

(3) the loop of question (2) is now rotating around the green axis.



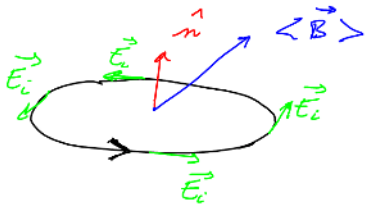
How does the magnetic force scale with the angular speed ω? Can it lead to a current around the loop ?

- Magnetic induction:

There is another, more indirect way, for a magnetic field to give rise to a force on a charged particle. Whenever the magnetic field changes through time, it gives rise to an electric field. The phenomenon is called magnetic induction. The electric field created thus is said to be induced by the changing magnetic field. If there are charges present as well, the full electric field is the sum of the electric field created by the charges, which we studied earlier in the class, and the electric field induced by the changing magnetic field.

The general expression of the induced electric field as a function of the changing magnetic field is a bit complicated so we'll focus on the case of a wire loop.

Let's consider a loop for which we've defined a positive direction (black arrow).



That in turn defines the direction of the normal unit vector $\hat{n}$ by the second right hand rule.

Let $\langle \vec{B} \rangle$ be the average magnetic field inside the loop.

When $\vec{B}$ changes through time, it induce an electric field $\vec{E_i}$ that swirls around the loop.

Let's focus on the tangential component of $\vec{E_i}$, i.e., the part of it that is along the loop. Let's call it E_i and count it positively if it's along the positive direction defined by the black current arrow and negatively if it points in the opposite direction.

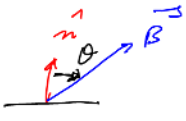
The law of induction tells us that the average value of E_i around the loop is proportional to minus the rate of change of $\langle \vec{B} \rangle . \hat{n}$:

$$\langle E_i \rangle \propto -\frac{d}{dt} \left(\langle \vec{B} \rangle . \hat{n} \right)$$

- Example:

Let's go back to question ③ from problem 6, but from the point of view of the loop. Instead of a fixed $\vec{B}$ and a loop rotating at ω, we now consider a fixed loop but $\vec{B}$ rotates at ω in the xy plane.

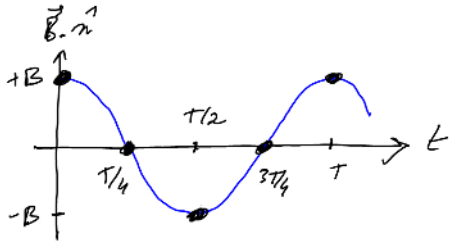
Looking along the y axis:



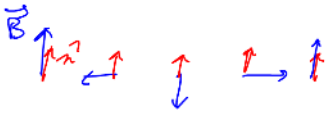
$\theta = \omega t$

We assume $\vec{B}$ is uniform so $\langle \vec{B} \rangle = \vec{B}$.

$\vec{B} \cdot \hat{n} = B \cos\theta = B \cos(\omega t)$.



$T = \frac{2\pi}{\omega}$ is the period of rotation



Between $t=0$ and $t=\frac{T}{2}$, $\vec{B}\cdot\hat{n}$ decreases therefore $-\frac{d}{dt}(\vec{B}\cdot\hat{n}) > 0$ therefore $E_i > 0$. This in turn drives a current $I > 0$.

Between $t=\frac{T}{2}$ and $t=T$, $\vec{B}\cdot\hat{n}$ increases therefore $-\frac{d}{dt}(\vec{B}\cdot\hat{n}) < 0$ therefore $E_i < 0$. This in turn drives a current $I < 0$.

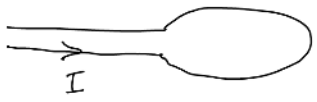
- Problem 7:

  Show that doubling the rotation speed in the example above doubles the induced electric field.

  Show that the magnetic force in problem 6 question 3 was also proportional to the rotation speed.

- Problem 8: Power generation.

  Each of the following set-ups can be used to generate AC power in the loop. In each case, discuss the direction of the induced electric field at various stages of the motion of the magnet(s).

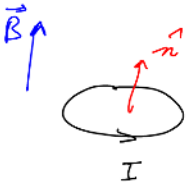
① 

The magnet rotates clockwise.

② 

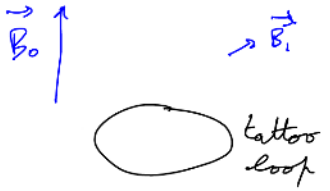
The magnet is translated up and down periodically.

- **Problem 9:** Lenz law.



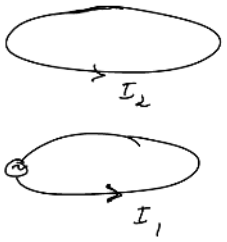
Consider a wire loop in a uniform field $\vec{B}$ parallel to $\hat{n}$. Show that if the magnitude of $\vec{B}$ change then it induces an electric field which drives a current which creates a magnetic field $\vec{B}'$ that opposes the changes (i.e., if $\frac{d\vec{B}}{dt}$ is along $\hat{n}$ then $\vec{B}'$ is along $-\hat{n}$).

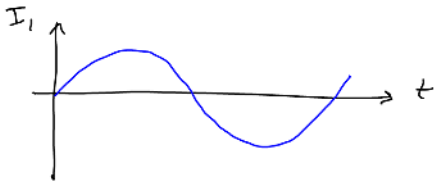
- **Problem 10:** MRI tattoo burn.



MRI involves a strong static magnetic field $\vec{B}_0$ and a much weaker but fast oscillating (tens of MHz) magnetic field $\vec{B}_1$. Both are uniform. The total field is $\vec{B}_0 + \vec{B}_1$. Imagine the patient has a tattoo made of somewhat conductive ink that forms a loop. Explain how that may lead to a burn. Is it caused by $\vec{B}_0$, $\vec{B}_1$, or both? How does it depend on the magnitude and frequency of the field?

- Problem 11: Mutual inductance.





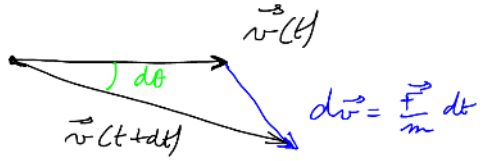
An AC source in loop 1 creates current I_1. Explain why this gives rise to a current I_2 in the second loop (made of wire).

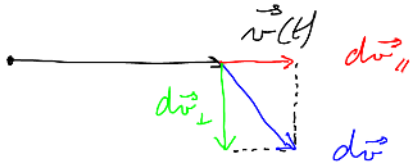
Sketch $I_2(t)$. First you'll need to sketch other quantities involved in predicting I_2.

- A useful result from kinematics:

This is largely a repeat of section 3.6 "Electric force and circular motion" from the "Electric Field" lecture notes.

Consider a particle with velocity $\vec{v}$.





Over an infinitesimal time interval dt, $\vec{v}$ changes by $d\vec{v} \equiv \vec{v}(t+dt) - \vec{v}(t) = \frac{d\vec{v}}{dt}dt = \frac{\vec{F}}{m}dt$ where $\frac{d\vec{v}}{dt}$ is the acceleration, $\vec{F}$ is the force on the particle, and m is its mass.

$d\vec{v}$ can always be decomposed into $d\vec{v} = d\vec{v}_{\parallel} + d\vec{v}_{\perp}$ where $d\vec{v}_{\parallel} \parallel \vec{v}$ and $d\vec{v}_{\perp} \perp \vec{v}$. As long as dt, thus $d\vec{v}$, is infinitesimally small, $d\vec{v}_{\parallel}$ only affects the speed $|\vec{v}|$, not the direction of motion $\hat{v}$. Conversely, $d\vec{v}_{\perp}$ only affects $\hat{v}$, not $|\vec{v}|$.

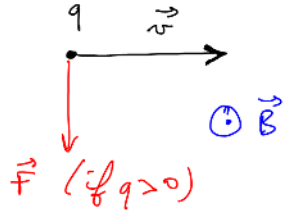
One can show that the angular speed due to $d\vec{v}_{\perp}$ is $\omega = \frac{1}{v}\frac{dv_{\perp}}{dt} = \frac{F_{\perp}}{mv}$ and the corresponding radius of curvature is $R = \frac{mv^2}{F_{\perp}}$.

You may be more familiar with this as an equality between the centripetal acceleration mv^2 and the radial force $F_{\perp}$ in circular orbits.

This is relevant to the magnetic force because $\vec{F} = q\vec{v} \times \vec{B}$ is always perpendicular to $\vec{v}$. If this is the only force then $\frac{d\vec{v}}{dt} = \frac{q}{m}\vec{v} \times \vec{B} \perp \vec{v}$ therefore the particle's speed remains constant and it turns towards the direction of $q\vec{v} \times \vec{B}$ with radius of curvature $R = \frac{mv^2}{|q|\,|\vec{v} \times \vec{B}|}$.

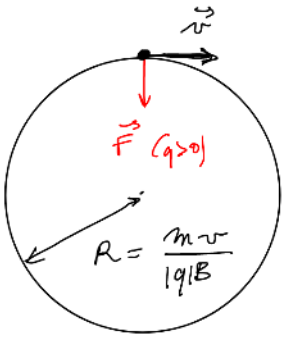
- Motion of a charged particle in a magnetic field:
What's important to realize is that $\vec{F} = q\vec{v} \times \vec{B}$ remains perpendicular to $\vec{v}$ even as $\vec{v}$ turns towards $\vec{F}$. There is no aligning with $\vec{F}$, only endless turning with radius of curvature $R = \frac{mv^2}{|q|\,|\vec{v} \times \vec{B}|}$.

The simplest case is when $\vec{v} \perp \vec{B}$.



Then $|\vec{v} \times \vec{B}| = |\vec{v}|\,|\vec{B}| \equiv vB$ and

$$\boxed{R = \frac{mv}{|q|B}}.$$



Note that the direction of $\vec{F}$, thus the turning direction, flips if $q < 0$.

If $\vec{B}$ is not uniform, or if it changes through time, the radius of curvature changes and the trajectory gets more complicated.

Note when a particle experiences a change magnetic field (whether because the field has change or because it's moved to a location where the field was different), its velocity experiences no abrupt change. It's the acceleration $\vec{a} = \frac{\vec{F}}{m} = \frac{q\vec{v}\times\vec{B}}{m}$ which changes in direct response to a change in $\vec{B}$.

If $\vec{B}$ is not perpendicular to $\vec{v}$ then we write $\vec{v} = \vec{v}_{\parallel} + \vec{v}_{\perp}$ where $\vec{v}_{\parallel} \parallel \vec{B}$ and $\vec{v}_{\perp} \perp \vec{B}$.

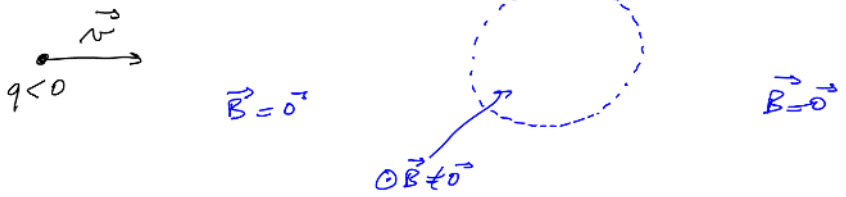
$\vec{v}\times\vec{B} = \underbrace{\vec{v}_{\parallel}\times\vec{B}}_{=\vec{0}} + \vec{v}_{\perp}\times\vec{B}$ so the turning essentially ignores $\vec{v}_{\parallel}$. What ends up happening is that we have the same circular trajectory as before in the plane of $\vec{v}_{\perp}$ and $\vec{F}$, i.e., the plane $\perp$ to $\vec{B}$, with curvature radius $R = \frac{mv_{\perp}}{|q|B}$, but in addition to that the particle moves steadily at $\vec{v}_{\parallel}$ along $\vec{B}$. The resulting trajectory is a spiral with its axis along $\vec{B}$. This is how the solar wind gets guided towards the poles (where $\vec{B}$ goes into/comes out of the ground) to give rise to northern lights.

(not essential)

- **Problem 12:**

In each case discuss the trajectory.

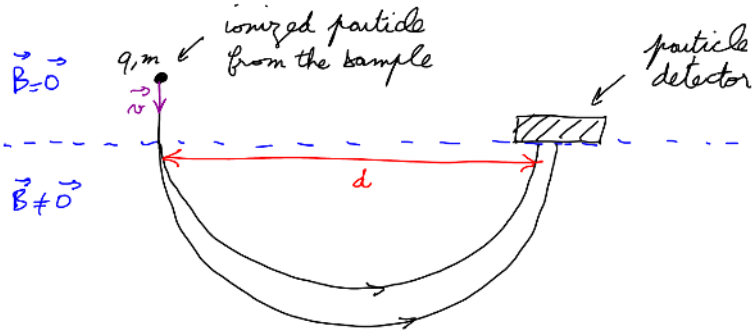
(1)



(2)

Draw three trajectories corresponding to three particles with the same mass but different charges: $q_1 < 0$, $q_2 = |q_1| > 0$, and $q_3 > q_2$.

- Mass spectrometer :



The idea behind the mass spectrometer is to vaporize the sample, ionize its atoms, accelerate them to a known velocity $\vec{v}$ (with an electric field), then have them travel through a region with a uniform magnetic field $\vec{B} \perp \vec{v}$. This makes each atom follow a circular path whose radius depends on the atom's properties. A detector then measures where each atom lands (the distance d) and infers what chemical element it is. By counting the number of atoms received at each landing point the device can also measure the relative abundance of each element in the sample.

- **Problem 13 : Mass spectrometer.**

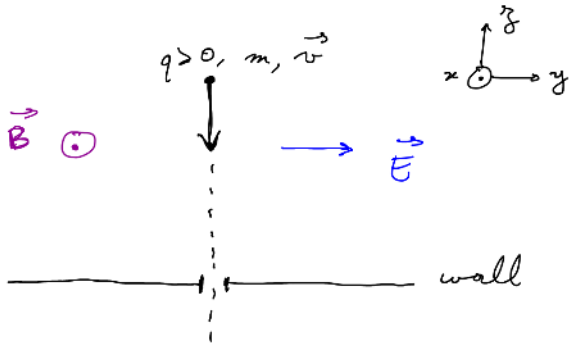
  (1) Assuming $q > 0$, what direction does $\vec{B}$ need to have to get the trajectories I drew?

  (2) Compute d as a function of m, q, v, B.

  (3) Assuming v and B are known but neither m or q is, what quantity does the device measure? Does that seem like enough information to uniquely identify which element it is?

- **Problem 14 : Speed selector.**

The mass spectrometer above relies on incoming particles having a known speed. Here is one way to screen charged particles based on their speed.



We assume $\vec{v}$ is always along $-\hat{z}$ but $\|\vec{v}\|$ is unknown (in other words $\vec{v} = -v\hat{z}$), $\vec{B} = B\hat{x}$, and $\vec{E} = E\hat{y}$. The only forces at play are the electric force due to $\vec{E}$ and the magnetic force due to $\vec{B}$.

To pass through the hole in the wall, particles must go straight.

(1) Write the equation of motion for the particle. What condition must $\vec{E}$ and $\vec{B}$ obey for the particle to go straight?

(2) How must one choose E and B in order to only let particles with a specific speed v through? Does the result depend on q? Why does (a'4) it?

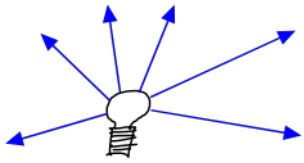
(3) Sketch the trajectory of a positively charged particle going too fast.

Chapter 5: Geometrical Optics

5.1 Basic concepts

5.1.1 Light rays

Geometrical optics is also known as ray optics. The basic idea is that light sources emit "light rays" that extend in straight lines away from the source.

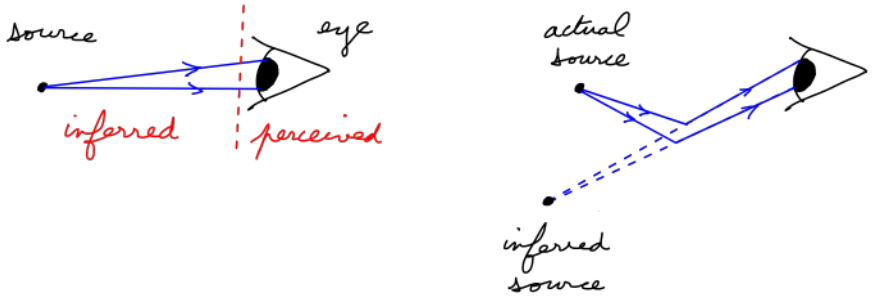


Another way to think about this is in terms of light particles called "photons". Souorces of light "throw out" photons which then proceed in straight lines by Newton's first law. The light rays are the trajectories of the photons.

On some occasions, lights rays do change direction. When they encounter a reflective surface, they bounce off of it. When they enter a new medium, they make a sharp turn. Geometrical optics is the study of those changes of direction, they consequences, and their applications. That includes devices made of lenses and/or mirrors like reflective paint, prescription glasses, the eye, cameras, telescopes, and more.

5.1.2 Perceived location of an object

The eye cannot tell where the source is, only which direction the ray(s) came from. If there are multiple rays, and their directions are consistent with them all coming from the same location, the brain infers that's where the source is. If the rays are deflected between the source and the eye, the brain infers the wrong location.



5.2 Reflection

5.2.1 Motivation

Here are a few systems the laws of reflection will allow us to understand better.

Retroreflectors

Those are devices that reflect light straight back at the source no matter where the source is. The simpler kind we'll study is used in bike reflectors, raised pavement markers, in radar reflectors, and to measure the earth-moon distance. The same principle is also is used in street signs, reflective clothing, although those also involve some refraction, which we'll study later. It's also involved in the eyeshine phenomenon as observed when taking a picture of a cat or dog with flash.

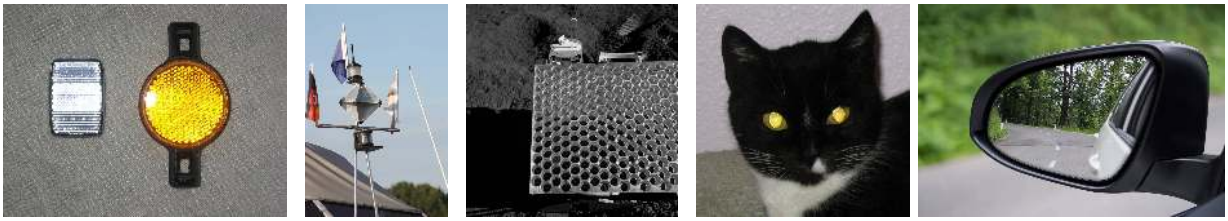


Figure 5.1: Left to right: bike reflector, radar reflector, moon reflector used to measure earth-moon distance, eyeshine phenomenon, field of view of a car mirror.

Field of view of a mirror

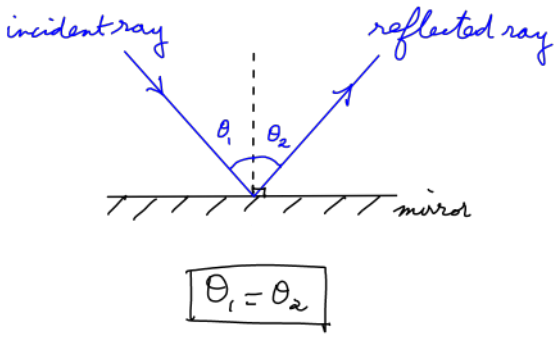
How can we predict what is or isn't visible in a mirror? How can we design/place/orient of mirror so that something is visible in it?

Towards refraction, lenses, and the eye

Reflection is the simplest type of change of direction a light ray can experience. It's the easiest way to introduce the central concepts of geometrical optics, which we'll then use to understand refraction, lenses, and the eye.

5.2.2 Law of reflection

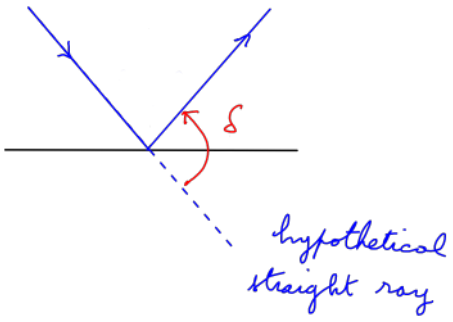
The angle of incidence (θ_1) is equal to the angle of reflection θ_2. Both are measured between the ray (incident ray for θ_1, reflected ray for θ_2) and the line perpendicular to the surface of the mirror going through the point where the ray strikes the mirror. The latter is known as *the normal*.



Note that the law of reflection is reversible, i.e., if you turn the reflected ray around (flip the arrow), it will retrace its step, hit the mirror at the same point of incidence, then continue along the incident ray above.

5.2.3 Deviation angle

The deviation angle is the angle by which the ray has turned. I usually call it δ.



5.2.4 Problems

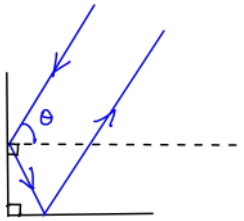
Problem 59: Deviation in a reflection.

A ray hits a mirror with angle of incidence θ. Compute the deviation angle δ.

Problem 60: Corner reflector.

A corner reflector is made of two mirrors attached together at a right angle. The blue ray bounces off of the vertical mirror first, then the horizontal one. The angle of incidence for the first reflection is θ.

1. Compute the deviation angle δ_1 for the first reflection.
2. Compute the angle of incidence for the second reflection.
3. Compute the deviation angle δ_2 for the second reflection.
4. Compute the total deviation angle δ, i.e., the total change of direction after both reflections. Summarize the result in plain English. What's remarkable about it?



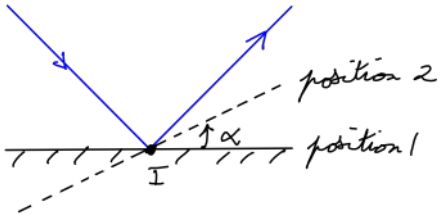
Problem 61: Imperfect corner reflector.

The two mirrors now make an angle α which is not quite $\pi/2$.

1. Compute the new total deviation angle δ.
2. If the right angle is off by 1°, how far is the total deviation from a perfect half turn?

Problem 62: Rotating mirror.

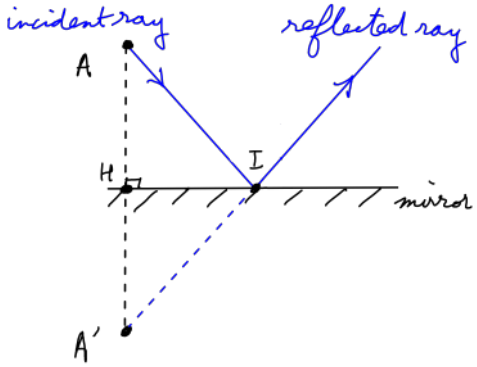
The mirror is initially in position 1. The blue ray bounces off of the mirror at point I. The mirror now rotates around I by an angle α, eventually reaching position 2. How much does the reflected ray rotate?



5.2.5 Law of reflection 2

Let A' be the mirror image of A with respect to the mirror. After bouncing off of the mirror, an incident ray coming from A looks like it's coming from A'.

This is fully equivalent to the equality of the angle of incidence and the angle of reflection. It's merely another way to express the same law. It allows us to construct reflected rays with a ruler. First, construct the mirror image A' using $\overline{AH} \perp \overline{HI}$ and $AH = HA'$, then draw $\overline{A'I}$ and extend it past I. It's convention to keep solid lines for actual rays and use dashed lines for $\overline{AA'}$ and $\overline{AI}$. More generally, use dashed lines for "construction lines", i.e., lines you need to draw to get to the result but do not correspond to any physical ray of light.

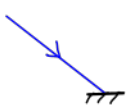


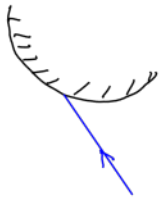
Important note regarding exams

Don't erase the dashed lines after constructing a ray. Use the appropriate perpendicular signs. On exams there are no points for lucky guesses, only for verifiably correct constructions. You will lose points every time I can't tell how you constructed your ray.

Problem 63: Reflected ray.

Construct the reflected ray of each incident ray below.

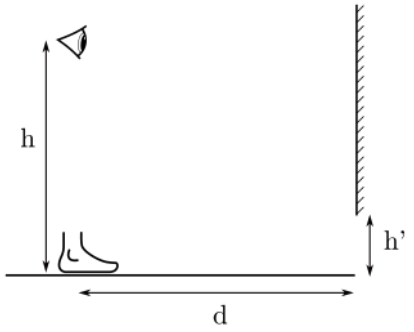




Problem 64: Seeing your feet in a mirror.

Use a graphical construction to show whether the eye can see the foot in the mirror. How far off the ground can the mirror be (largest possible h') for the eye to see the foot?

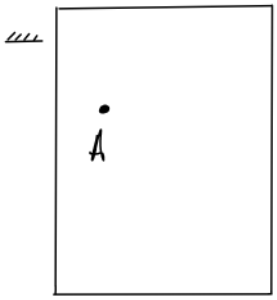
Info: The eye is right above the foot. The mirror is perpendicular to the ground. You can treat the eye and the foot as points.



Problem 65: Blind spot.

The rectangle is a car. A is the driver. Construct and shade the region visible in the side-view mirror.

Hint: A point is visible in the mirror if there is a ray from that point that bounces off the mirror and reaches the eye. Since light rays are reversible, this is equivalent to there being a ray from the eye that bounces off the mirror and reaches the point. In other words, a point is visible by the eye if and only if the light from a hypothetical light source located where the eye is would reach it.

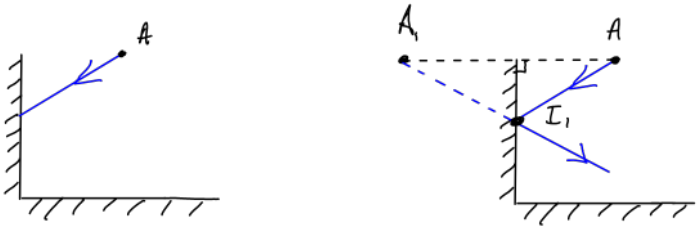


Variation: Replace the mirror with this one: 

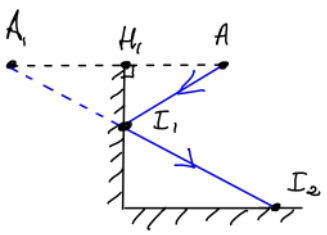
5.2.6 Multiple reflections

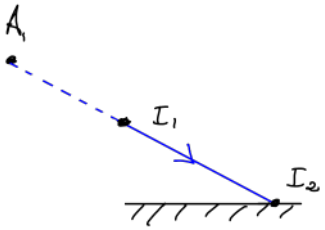
Let's apply the graphical approach to the corner reflector from earlier:

1. The ray first reflects off the vertical mirror at a point we'll call I_1. To construct the reflected ray, we first construct the mirror image A_1 or A with respect to the vertical mirror, then extend A_1I_1.

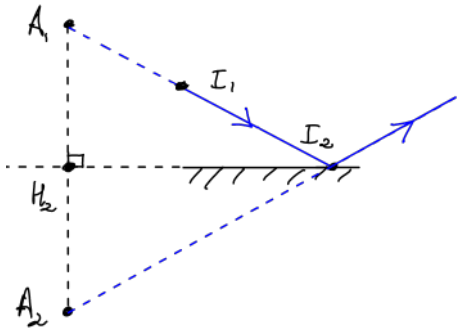


2. After reflecting off the vertical mirror, the ray hits the horizontal mirror at a point we'll call I_2. **As far as the horizontal mirror is concerned, the ray is coming from** A_1. The fact that the ray actually came from A along AI_1 is completely irrelevant to the construction of the ray's path after I_2. All that matters is the ray's direction at the time when it hits I_2, and that direction is along A_1I_2.



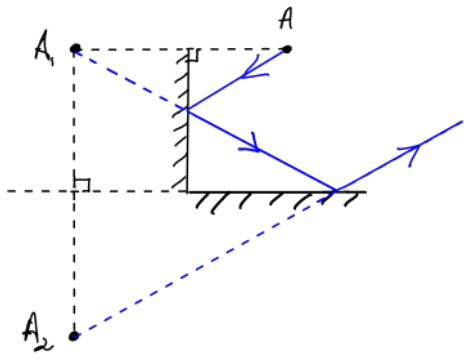


3. To construct the path of the ray after I_2, imagine A, I_1, and the vertical mirror are no longer there. Erase them in your mind. You're left with a ray coming from A_1 and hitting the horizontal mirror at I_2. Construct the mirror image A_2 of A_1 with respect to the horizontal mirror, then extend A_2I_2 past I_2. That's you're reflected ray.



Comments

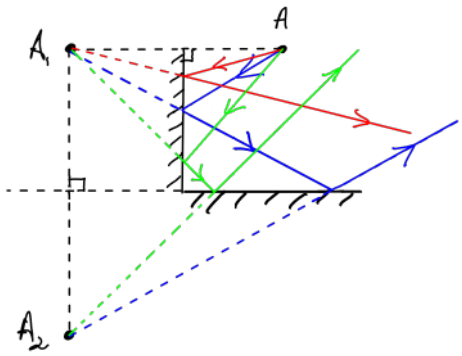
- At first it's a good idea to break the process down as above and draw multiple figures. The key is to identify which parts of the figure are relevant to each step of the reasoning and ignore the rest, at least while you perform that step. Once you're comfortable with the process you can do it all on the same figure. It will look like this:



  That being said, never feel bad about starting a new drawing if the current one gets overcrowded. This is especially true when there are multiple rays involved.
- Reminder: On exams you will lose points every time I can't tell how you constructed your ray. All your intermediate steps must be visible.
- In order to construct accurate rays you have to put some effort into every step. Make sure your right angles are truly right, your equal distances are truly equal, your lines go through the points they're supposed to go through, etc. Use an actual ruler. Inaccuracies tend to build up step after step.

5.2.7 Multiple rays from the same source

The beauty of the geometrical construction of section 5.2.6 is that the same A_1 and A_2 can be used to construct the path of any ray from A that hits the vertical mirror then the horizontal mirror. Every ray coming from A that hits the vertical mirror, no matter its exact orientation, comes out pointing away from A_1. Similarly, every ray coming (or looking like it coming) from A_1 that hits the horizontal mirror, no matter its exact orientation, comes out pointing away from A_2.



5.2.8 Objects and Images

We call *object* a point that has rays either coming out of it or looking like they're coming out of it. When rays coming from an object A go through an optical system (a set of mirrors, lenses, etc) and come out looking like they're all coming from the same point A', we say that A' is the *image* of A by the optical system.

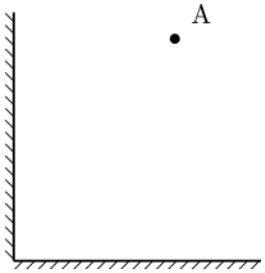
Objects and images are context-dependent. In the corner reflector, A_1 is the image of A by the vertical mirror, and A_2 is the image of A_1 by the horizontal mirror. When we construct A_1 from A, we think of A as the object and A_1 as the image, but when we construct A_2 from A_1 we think of A_1 as the object and A_2 as the image.

What about the image of A by the corner reflector as a whole. Well, it depends which mirrors they hit. A ray from A that hits the vertical mirror then the horizontal mirror will look like it's coming from A_2. However, a ray from A that hits the vertical mirror then leaves without hitting the horizontal mirror will look like it's coming from A_1. Thus, the corner reflector generates multiple images of A. Concretely, someone looking at the corner reflector may see multiple copies of whatever is at A.

Problem 66

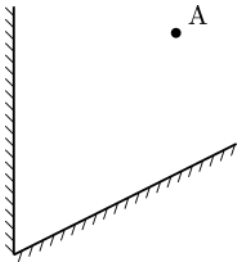
There are four distinct ways a ray from A can hit the mirror(s). Draw an example of each. Construct the path of each ray until it exits the corner reflector for good.

How many images of A can one see in this set of mirrors?



Problem 67

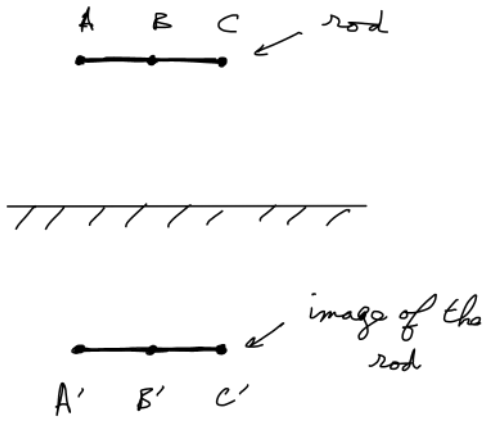
Same questions for the system below. What's different?



5.2.9 Extended objects

Not all objects are points, but an extended object can be treated a collection of points. Although we can't construct the image of every point (because there are infinitely many), often a few is enough to get a sense of what's going on.

In the example below, the image $A'B'C'$ of the line ABC is still a line. B' is still the middle of $A'C'$. The distance between A' and C' is the same as between A and C. From there it's not hard to convince yourself that the image of the rod has the exact same size and shape as the original rod.



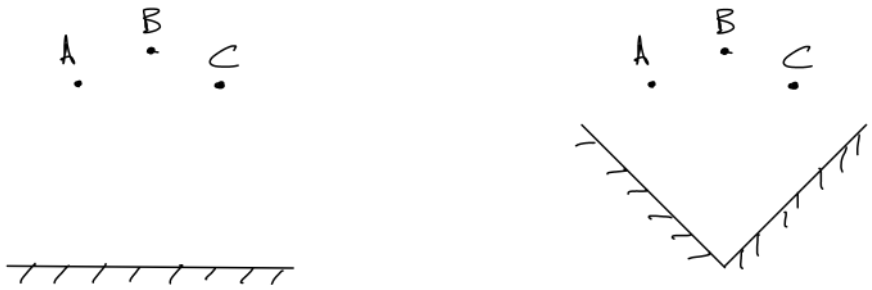
Note: This is only true for flat mirrors. Curved mirrors and lenses do magnify or shrink and sometimes deform, i.e., they produce images whose size and sometimes shape are different from the object.

Problem 68: Flipping vs non-flipping mirror.

Construct the images of A, B, C through the mirrors. Call them A', B', and C'. For the two-mirror set-up, assume the rays hit the left mirror then the right one (the mirrors are perpendicular, so you'd get the same result reflecting off the right mirror then the left one, but it's easier to help each other if we all do it the same way).

Does $A \to B \to C$ go around the triangle clockwise or counterclockwise? What about $A' \to B' \to C'$? Which mirror configuration flips the "clockwiseness" of the triangle?

Side note: You may have heard of chirality in Chemistry. You may have defined a chiral molecule as one whose mirror image cannot be overlaid onto the original molecule. The clockwiseness discussed above is a 2D version of chirality.

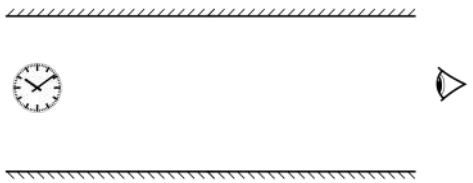


Problem 69: Reflected clock. [optional]

Construct the reflection of the clock in the bottom mirror. When looking at that reflection, do the hands of the clock look to be rotating clockwise or counterclockwise?

Same questions for the reflection of the clock in the bottom mirror then the top mirror.

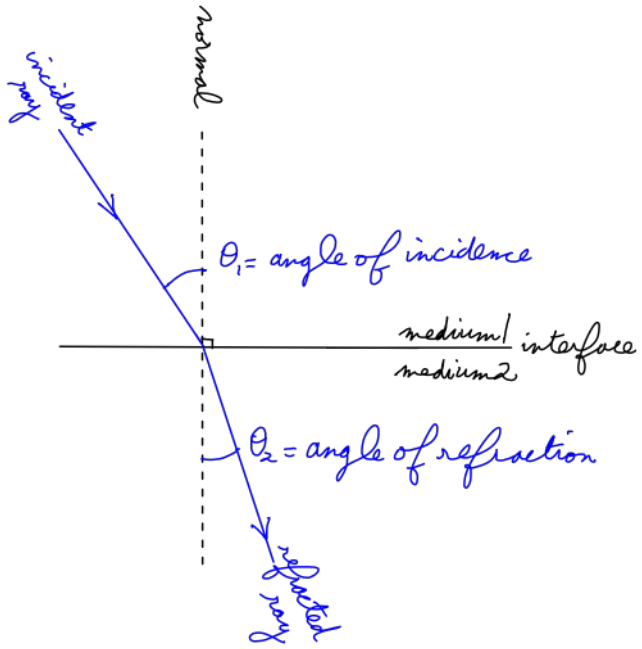
Hint: Replace the clock by two points corresponding to the base and tip of either hand. Construct their images for two positions of the hand about 15 minutes apart.



5.3 Refraction

5.3.1 Law of refraction

Refraction is the sharp change of direction a ray of light experiences as it crosses the interface between two transparent media. The direction of the refracted ray is controlled by three things: the direction of the incident ray, and the refractive index of the two media.



The refractive index n of a medium is equal to the speed of light in vacuum ($c = 3 \times 10^8\,\mathrm{m\,s^{-1}}$) divided by the speed of light v in the medium: $n = \dfrac{c}{v}$. From the theory of relativity we know nothing can go faster than c, therefore n is always ≥ 1. In vacuum $v = c$ therefore n is exactly 1. Gases tend to have $n \approx 1$ because their molecules are far from each other, so in a sense they're mostly made of vacuum. The table below shows its value in a few common materials.

Material	vacuum	air	water	window glass	diamond
Refractive index	1	1.0002	1.33	1.52	2.42

The law of refraction relates the direction of the ray before refraction (incident ray), its direction after refraction (refracted ray), and the refractive indices of the two media:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

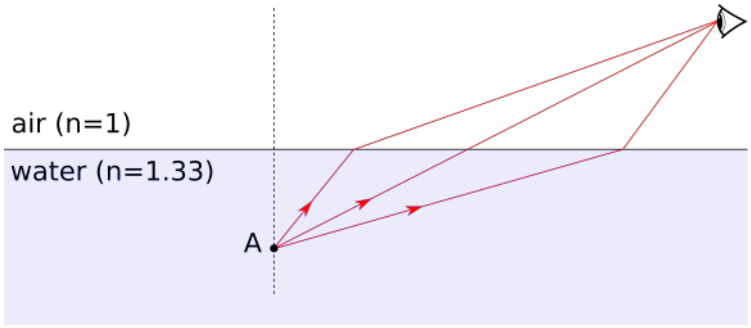
Problem 70: First consequences of the law of refraction.

1. According to the definition of the refractive index, what is its dimension (in the sense of dimensional analysis)? What is its SI unit?
2. What is θ_1 for an incident ray the normal to the interface? Once you have θ_1, use the law of refraction to compute θ_2. State the result in plain English.
3. Solve for θ_2 when the two media have the same refractive index ($n_1 = n_2$). State the result in plain English.
4. If $n_2 > n_1$, is the refracted ray closer or further away from the normal than the incident ray?
5. In the sketch above, which medium has the larger refractive index?

Problem 71: Looking into water.

1. Which of the three red rays is consistent with the law of refraction?w

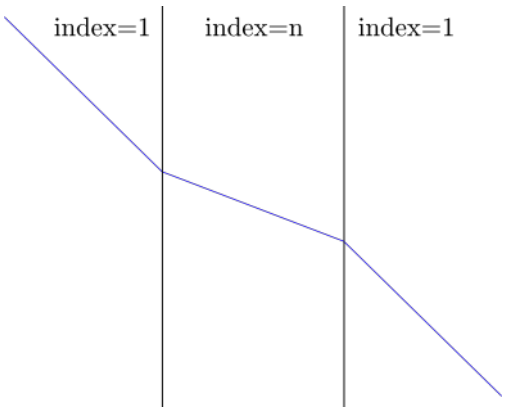
2. Let's assume A has an image A' located somewhere on the vertical dashed line. Is A' above or below A. Why?



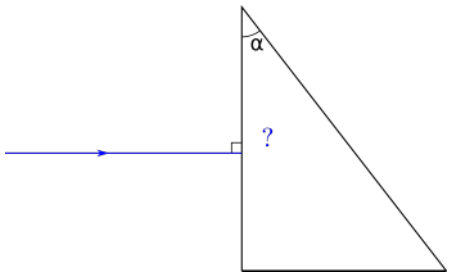
Problem 72: Slab.

A ray coming from air crosses a piece of material with refractive index n whose faces are parallel to each other.

1. Draw the angle of incidence θ_1 for the first refraction on the sketch. Express the other answers in terms of θ_1 and n.
2. Draw the angle of refraction θ_2 for the first refraction. Relate it to θ_1.
3. Draw the angle of incidence θ_3 for the second refraction. Relate it to θ_2, then to θ_1.
4. Draw the angle of refraction θ_4 for the second refraction. Relate it to θ_3, then to θ_1.
5. Show that the total deviation is zero, i.e., the ray exits the slab in the same direction it entered it.



Problem 73: Normal incidence on a prism.



The blue ray arrives perpendicularly on the left face. It is refracted once as it enters the prism, then again as it exits the prism. The prism is made of glass ($n = 1.5$).The surrounding medium is air ($n \approx 1$).

1. What is the angle of incidence θ_1 for the first refraction?
2. What is the angle of refraction θ_2 for the first refraction?
3. What is the angle of incidence θ_3 for the second refraction?
4. What is the angle of refraction θ_4 for the second refraction?

5.3.2 Dispersion

Problem 74: Dispersion by a prism.

In reality, the index of a material usually depends on the color (more specifically the wavelength) of the light traversing it. Imagine the ray of problem 73 is really made of three rays: one red, one green, and one blue. If they're right on top of each other, they will look like a single ray of white light. In the prism, though, each ray experiences a different refraction index: $n_{\text{blue}} = 1.5$, $n_{\text{green}} = 1.45$, and $n_{\text{red}} = 1.4$.

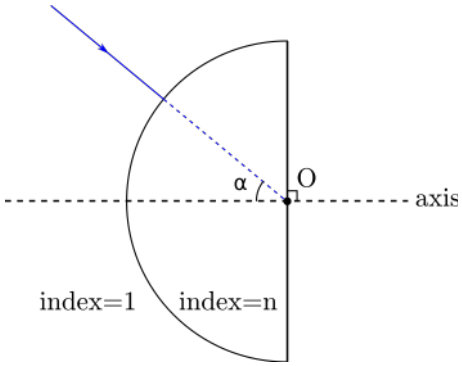
1. For each of the three rays, write the four angles (θ_1, θ_2, θ_3, θ_4) from the last problem in terms of α.
2. At which point do the rays diverge?
3. Which ray comes out above/in the middle/below. Explain your reasoning. Sketch the rays. The exact angles don't have to be correct on the sketch, but the rays should be on the correct side of the normals and in the correct color order.

5.3.3 Total reflection

Problem 75: Semi-circular prism.

The semi-circle is has a refractive index $n > 1$. It is surrounded by air (index 1). The ray points towards the center O of the circle.

1. Describe what happens to the ray as it enters then exits the prism.
2. Compute the angle between the exiting ray and the horizontal axis (thereafter the *exit angle*) as a function of α and n.



Problem 76: Total reflection.

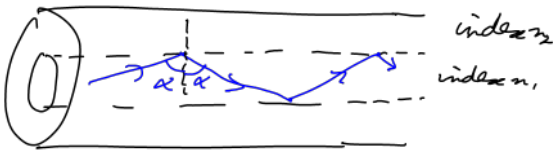
Assume $n = 1.5$ in the previous problem.

1. Compute the exit angle when $\alpha = 40°$.
2. Compute the exit angle when $\alpha = 45°$. What happens? What's causing it?
3. At what angle does it start to happen? This is called the critical angle.

Total reflection

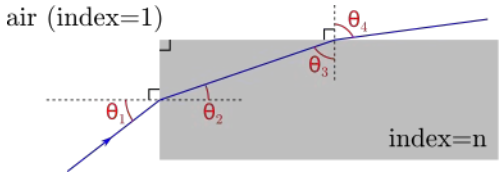
When there is no solution for the refracted angle in the law of refraction, there is no refracted ray. Instead the ray is reflected according to the law of reflection.

Application: Optical fiber



The central part of the fiber, called the core, has index n_1. The outer part, called the cladding, has index $n_2 < n_1$. If $\alpha > \arcsin(n_2/n_1)$, the ray experiences a total reflection every time it tries to exit the core. Once it's entered the fiber, it follows it until the other end.

Problem 77: Optical fiber.



If θ_1 is small enough, the ray experiences total reflection as it tries to exit the optical fiber and remains trapped thereafter. Conversely, if θ is large enough the ray does exit the optical fiber.

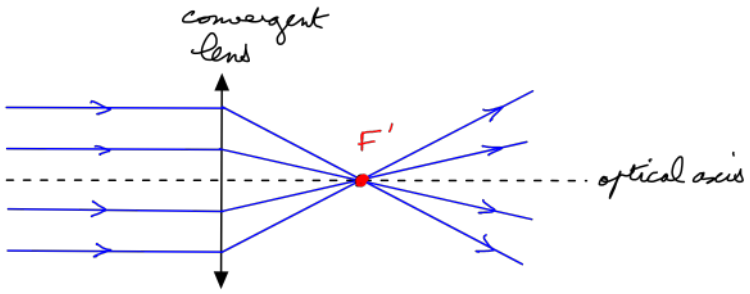
1. Write the law of refraction for the second refraction (as the ray exits the fiber). What is the value of θ_3 above which the ray experiences total reflection, thus failing to exit the fiber? The result should only depend on the index n of the fiber.
2. The ray remains trapped if θ_3 is larger than the value from the last question. What does that mean for θ_2, i.e., what inequality must θ_2 obey for the ray to be trapped in the fiber? The result should only depend on the index n of the fiber.
3. What does that mean for θ_1, i.e., what inequality must θ_1 obey for the ray to be trapped in the fiber? The result should only depend on the index n of the fiber.
4. Assuming rays are entering the fiber from every direction, and every direction is equally likely, what fraction of the rays entering the fiber remains trapped in it?

5.3.4 Towards lenses

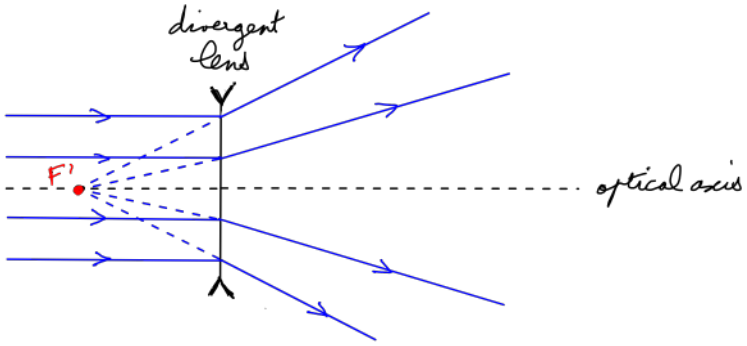
Definitions

The next few problems aim to give you a sense of how lenses work. For most intents and purposes you can think of a lens as a thin transparent disk whose faces are very slightly curved. The axis perpendicular to the surface of the lens and going through its center is called the *optical axis*. There's a special point on that axis called the image focal point, noted F'. Any ray that enters the lens parallel to its optical axis exits through F'. Depending on how this happens, the lens may be called convergent or divergent.

For a convergent lens, F' is located on the exit side and the rays literally go through it (they *converge* to F').

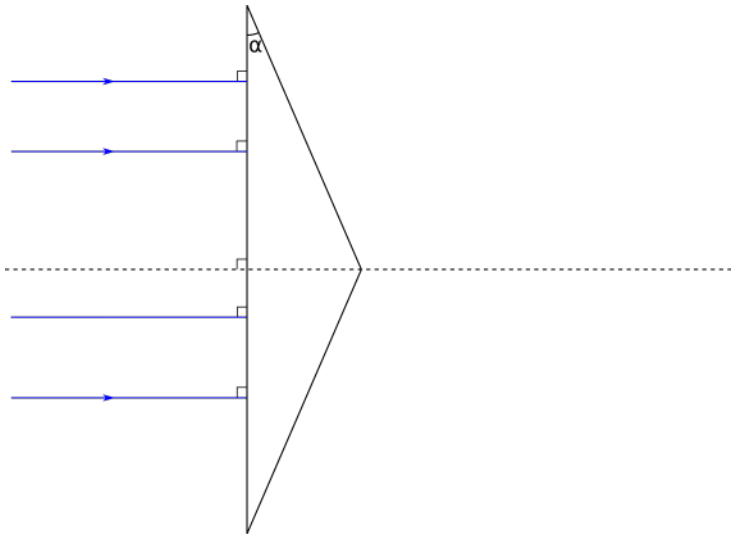


For a divergent lens, F' is located on the entry side. After being deflected the rays all look like they're coming from F', but they never actually go through F'.



Qualitative lens

Problem 78: Qualitative lens 1.



The rays are perpendicular to the entry face of the prism. The refractive index of the prism is larger than that of the surrounding medium.

1. Which way does the upper prism deflect the rays?
2. Without computing it, compare the deflection angles of the rays hitting the upper prism.
3. Answer questions 1 and 2 for the lower prism.
4. Sketch the path of the rays. What kind of lens is this device most similar to?
5. What's missing from this device to be a proper lens?

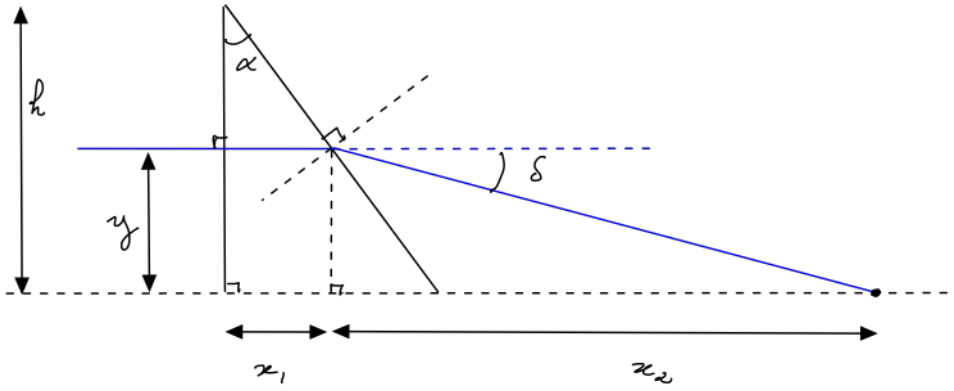
Paraxial approximation

When the angles involved in an optics problem are small, one can approximate trigonometric functions by their first order Taylor expansion around zero:

$$\theta \ll 1 \implies \sin\theta \approx \theta,\ \cos\theta \approx 1,\ \tan\theta \approx \theta$$

All the lens formulas we'll encounter are derived using this approximation. Among other things it simplifies the law of refraction to $n_1\theta_1 = n_2\theta_2$ (when θ_1 and θ_2 are small). For the rest of the section, unless specified otherwise, prisms are thin, rays make a small angle with the axis, and the paraxial approximation should be used with every small angle.

Problem 79: Paraxial prism.

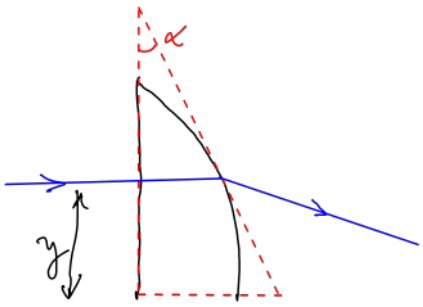


The refractive index of the prism is n. It is surrounded by air (index 1). α is very small. As a result x_1 is very small and we neglect it. Use the paraxial approximation whenever possible. Express everything in terms of α, h, y, and n.

1. Use the law of refraction to compute δ.
2. Compute x_2.

Curved lenses

Consider a lens with a flat, vertical entry face and a curved exit face. From the point of view of the blue ray, the lens might as well be the red prism. The angle α of the prism is the angle between the entry face and the tangent to the exit face at the location where the ray exits. Since the exit face is curved, α is not constant – it's a function of y. In other words, every ray "thinks" it's going through a prism, but rays with different values of y "see" different prisms (different values of α).

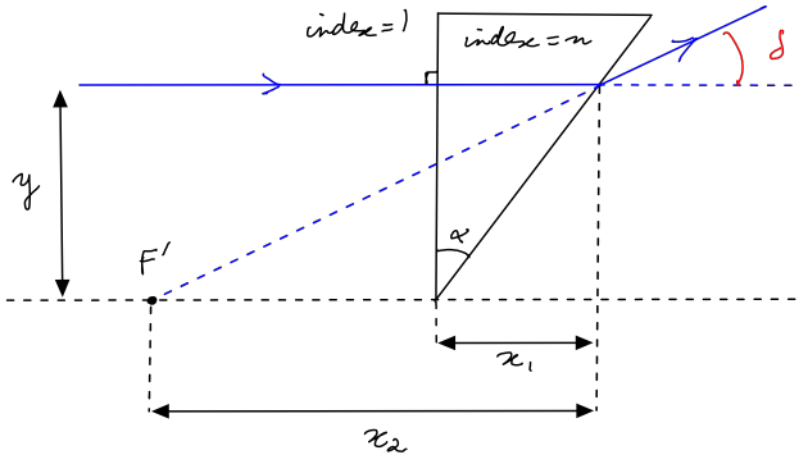


In problem 79, we computed how far a ray initially parallel to the axis hits the axis as a function of its initial distance from the axis y and the angle α of the prism. In order to build a convergent lens. We need every ray to cross the axis at the same point, regardless of its initial y. Therefore, we must find a function $\alpha(y)$ that makes $x_1 + x_2$ not depend on y. Better yet, since we work in the paraxial approximation, we can use $x_1 \ll x_2 \implies x_1 + x_2 \approx x_2$ and simply look for a function $\alpha(y)$ that makes x_2 not depend on y.

Problem 80: Plano-convex lens.

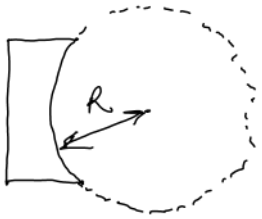
In problem 79, find the function $\alpha(y)$ that makes x_2 not depend on y.

Problem 81: Plano-concave lens.



Use the paraxial approximation whenever possible.

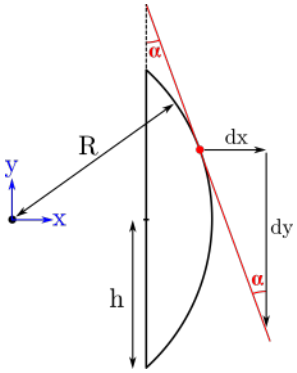
1. Compute δ as a function of n, α, y.
2. Compute x_2 as a function of n, α, y.
3. We now neglect x_1. In particular, we assume $x2 - x1 \approx x_2$. What condition does x_2 need to obey for every horizontal ray to emerge looking like it's coming from the same point F'?



4. When the exit face is a circle with curvature radius R, $\alpha = y/R$. Compute the corresponding x_2. Where is the corresponding F'?

Spherical lenses

The sketch shows a spherical cap. R is the radius of curvature of the spherical face. h is the radius of the flat face. As long as the cap is thin ($h \ll R$), it has just the right shape to be a lens, i.e., the angle α between the local tangent to the spherical face and the vertical axis is proportional to the height y.

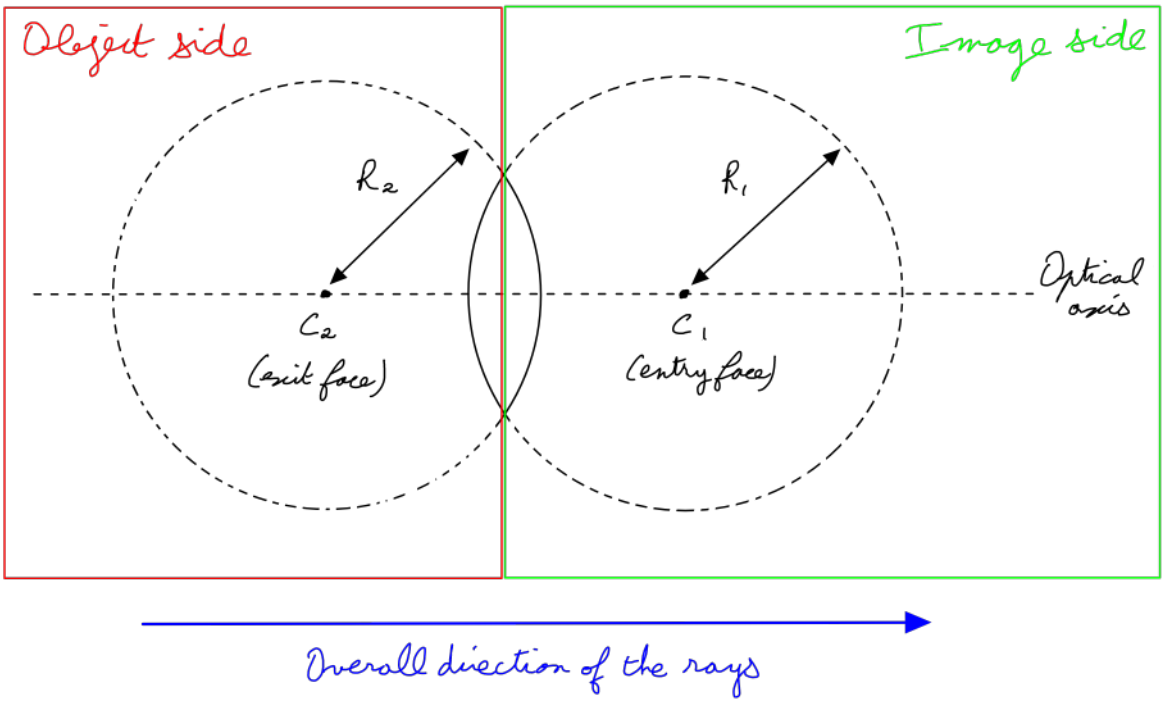


Here is a loose proof. α is related to dx and dy by $\tan\alpha = dx/dy$. The slope of the red line is also related to dx and dy: it is dy/dx. Another way to compute the slope of the red line is to write the function whose graph traces the spherical face and compute its derivative. The coordinates of a point on the spherical face obey $\sqrt{x^2+y^2} = R$ (their distance to the origin is R). Isolating y yields the corresponding function: $y(x) = \pm\sqrt{R^2 - x^2}$ (+ for the upper half, − for the lower half). Its derivative is $y'(x) = \mp x/\sqrt{R^2 - x^2} = \mp x/y$ (− for the upper half, + for the lower half). Thus $\tan\alpha = 1/y'(x) = y/x$. For a thin cap h is much smaller than R, x is almost equal to R, and α is very small. Thus we can simplify $\tan\alpha \approx \alpha$ and $y'(x) \approx y/R$ to get $\alpha = y/R$. As announced α is proportional to y, and the constant of proportionality is $1/R$.

5.4 Spherical lenses

5.4.1 Definitions

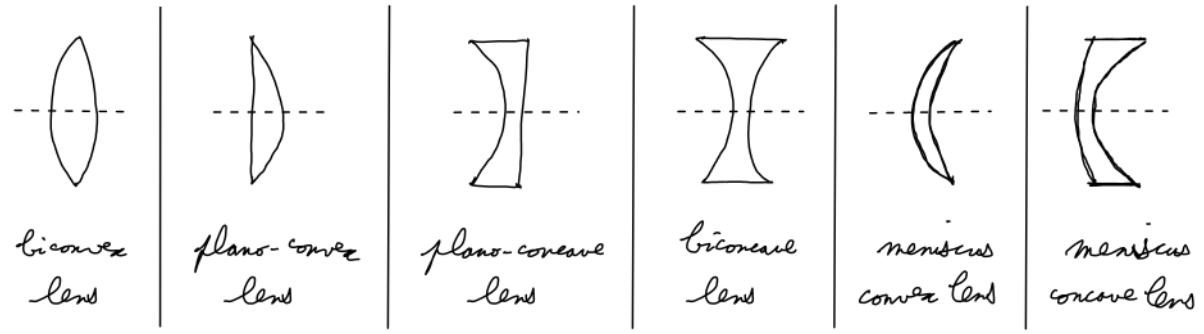
A spherical lens is a transparent object with two faces shaped like spherical caps. Each face has a center of curvature C and a radius R such that the distance between C and any point on that face is R. The entry face is the one the rays encounter first – the one through which they enter the lens. Its center and radius of curvature are called C_1 and R_1 respectively. The exit face is the one the rays encounter last – the one through which they exit the lens. Its center and radius of curvature are called C_2 and R_2 respectively. The side the rays come from, i.e., the side of the entry face, is called the *object side*. The side the rays exit on, i.e., the side of the exit face, is called the *image side*.



Whenever possible lenses are drawn vertically with the object side on the left, i.e., the overall direction of the rays is from left to right. If no rays are drawn, it is assumed they are coming from the left. In some rare instances, e.g., if rays are coming from both sides, one must remember that entry face/exit face/object side/image side are all defined relative to a ray. For example, one ray's object side can be another ray's image side, and one ray's R_1 can be another ray's R_2.

By convention R is positive if C is on the image side and negative if C is on the object side. In the sketch above $R_1 > 0$ and $R_2 < 0$. Here both faces protrude out. Their center sticks out more than their edges. Such faces are called convex. Conversely, a face that caves in, i.e., whose edges stick out more than its center, is called concave. The larger $|R|$ (absolute value of R), the less curved (the flatter) the face. In the limit $R \to \pm\infty$ the face is completely flat, or plane.

Lens shapes are named after the convexity/concavity of their faces:

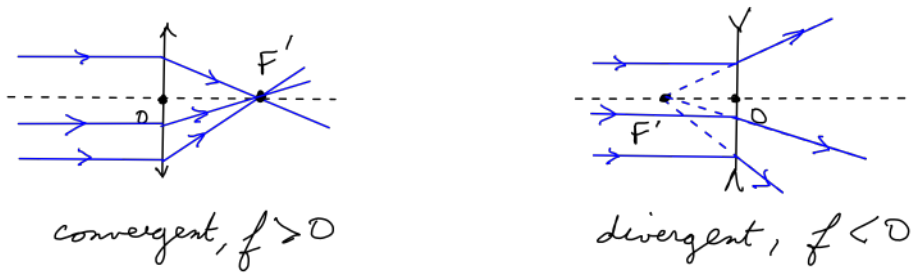


Problem 82: Curvature radius.

For each lens shape above, indicate which side C_1 and C_2 are, then determine the sign of R_1 and R_2.

Image focal point

Recall the image focal point F' we discussed in section 5.3.4. As long as the lens is reasonably thin and the rays are not too tilted on the optical axis, all spherical lenses have this ability to deflect every ray coming in parallel to the axis towards (or away from) F'.



The distance between the center O of the lens and the image focal point F' is called the focal length f. Like the radius of curvature of the faces, f is positive if F' is on the image side (convergent lens) and negative if F' is on the object side (divergent lens).

5.4.2 Lens maker's equation

The lens maker's equation relate the focal length f, i.e., the location of F', to the properties of the lens: its refractive index n and radius of curvature of its faces:

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

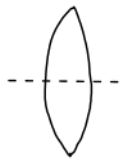
As we'll see, f is really all we need to know about a lens to predict the way it deflects any ray. Despite the variety of lens shapes and materials, all lenses work more or less the same. The only truly significant difference is between convergent ($f > 0$) and divergent ($f < 0$) lenses.

Example: Biconvex lens.

C_1 is on the image side therefore $R_1 > 0$. C_2 is on the object side therefore $R_2 < 0$. n is larger than 1 therefore $n - 1 > 0$.

$$\frac{1}{f} = \underset{>0}{(n-1)}\left(\underset{>0}{\frac{1}{R_1}} - \underset{<0}{\frac{1}{R_2}}\right) > 0 \implies f > 0$$

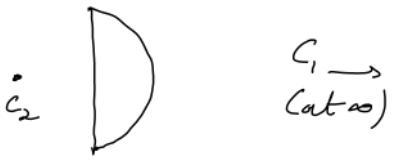
This proves that all biconvex lenses are convergent.



Example: Plano-convex lens.

The lens maker's equation works with flat faces too, you just have to set the corresponding radius, in this case R_1, to ∞.

$$\frac{1}{f} = \underset{>0}{(n-1)}\left(\underset{=0}{\frac{1}{\infty}} - \underset{<0}{\frac{1}{R_2}}\right) > 0 \implies f > 0$$



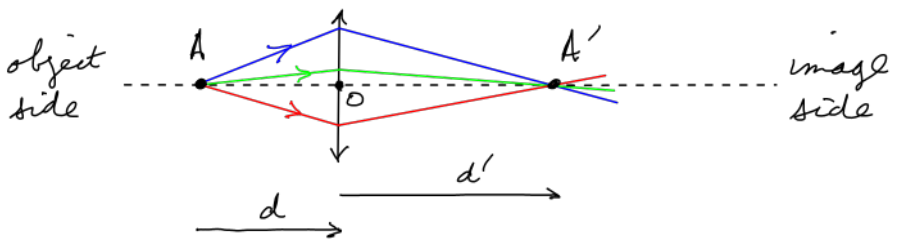
Note: C_1 could just as well be argued to be at infinity on the left, in which case $R_1 = -\infty$. Thankfully this has no consequence since R_1 only appears in the lens makers's equation as $\frac{1}{R_1}$ and $\frac{1}{\infty} = \frac{1}{-\infty} = 0$. More generally in geometrical optics the object-side and image-side "ends" of the optical axis are fully interchangeable in the sense that swapping them doesn't affect the result of any formula.

Problem 83

Determine the sign of f for every type of lens (biconvex, plano-convex, plano-concave, biconcave, meniscus convex, meniscus concave). For the last two you will need to compare the absolute values of R_1 and R_2.

5.4.3 Thin lens equation

It's not just rays parallel to the axis that lenses deflect towards (or away from) a common point. Any set of rays coming from the same point A gets deflected towards (or away from) the same point A'.



A is the object. $d = AO$ is the distance between the object A and the center O of the lens. By definition it is positive if A is on the object side and negative

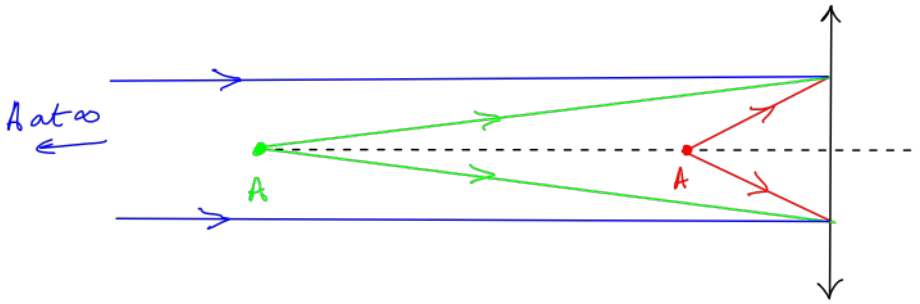
if A is on the image side. In the sketch above, d is positive if the arrow from A to O points to the right, negative if it points to the left.

A' is the image. $d' = OA'$ is the distance between the center O of the lens and the image A'. By definition it is positive if A' is on the image side and negative if A' is on the object side. In the sketch above, d' is positive if the arrow from O to A' points to the right, negative if it points to the left.

With those notations, the positions of A, A', and F' are related by the thin lens equation:

$$\frac{1}{d} + \frac{1}{d'} = \frac{1}{f}$$

Just like the lens maker's equation has no problems with the center of curvature C of a flat face being at infinity, the thin lens equation has no problems with either A or A' being at infinity. In fact we need $d = \infty$ to describe the set of rays parallel to the axis we used to define F'.



The further away A, the less divergent the rays from A are, the more parallel to each other they are. When A is infinitely far ($d = \infty$), the rays are actually parallel. Similarly, $d' = \infty$ corresponds to the rays being parallel after the lens.

In mathematics, parallel lines are usually said to not intersect at all. Here we're saying instead that they do intersect, but infinitely far. The two points of view are not mutually exclusive, they're just different ways to think about the same thing. Perhaps even more counter-intuitive, it doesn't matter whether they intersect infinitely far on the object side or infinitely far on the image side. Both describe the same situation: the rays being parallel.

Consistent with this idea and with the note at the end of the *Plano-convex* example above, the thin lens equation doesn't make any difference between a point being at infinity on the object side or at infinity on the image side. A, A', and F' only appear in the equation through the inverse of their distance to the lens ($1/d$, $1/d'$, and $1/f$ respectively), all of which have the exact same value (0) whether the distance is $+\infty$ or $-\infty$.

Example: Image focal point.

A set of incident rays parallel to the axis corresponds to an object A at infinity: $d = \infty$. Let's apply the thin lens equation:

$$\begin{cases} \dfrac{1}{d} + \dfrac{1}{d'} = \dfrac{1}{f} \\ \dfrac{1}{d} = \dfrac{1}{\infty} = 0 \end{cases} \implies \frac{1}{d'} = \frac{1}{f} \implies d' = f$$

We obtain $A' = F'$, i.e., F' is the image of an object located at infinity on the axis. In other words, any ray coming from a point at infinity on the optical axis

(i.e., coming parallel to the axis) emerges looking like it's coming from F', which is precisely how we first defined F' in section 5.3.4.

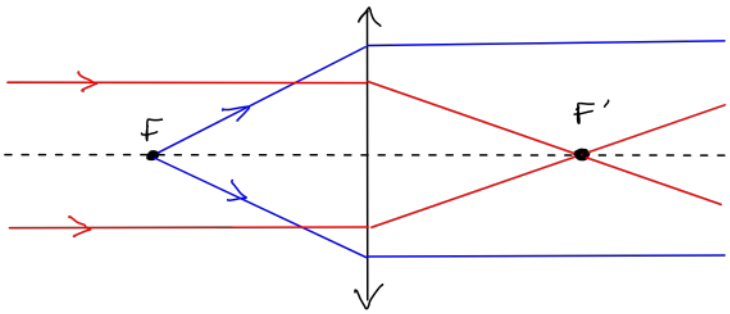
Example: Object focal point.

There is a second focal point called the *object focal point* F. Whereas F' is the answer to the question "Where do rays come from after going into the lens parallel to the axis?", F is the answer to the question "Where do rays need to come from to exit the lens parallel to the axis?". F' is the image of an object at infinity. F is the object whose image is at infinity. To find the location of F, where set $d' = \infty$ (image at infinity) in the thin lens equation and look for the position of the object:

$$\begin{cases} \dfrac{1}{d} + \dfrac{1}{d'} = \dfrac{1}{f} \\ \dfrac{1}{d'} = \dfrac{1}{\infty} = 0 \end{cases} \implies \frac{1}{d} = \frac{1}{f} \implies d = f$$

Since d is counted positively towards the object side whereas d' is counted positively towards the image side, $d = d'$ means that F is at the same distance as F' on the other side of the lens. In other words, F is the mirror image of F' with respect to the center of the lens.

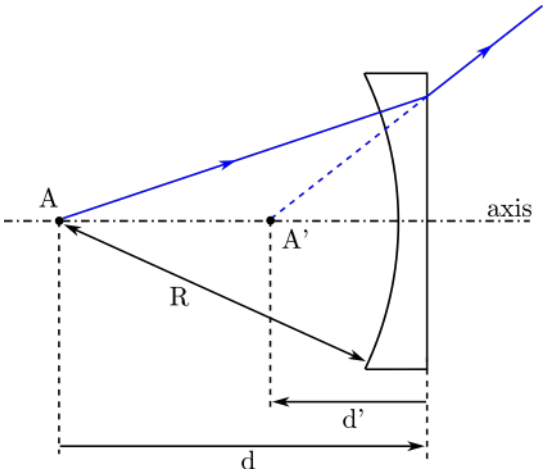
The meaning and locations of the two focal points are summarized below for a convergent lens:



For a divergent lens F and F' are swapped, i.e., F' is on the object side and F is on the image side.

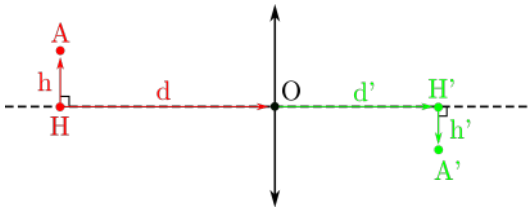
Problem 84: Another plano-concave lens.

An object A is located at the center of curvature of the plano-concave lens below. The goal of the problem is to determine the location of its image A'. The refractive index of the lens is n.



1. Use the lens maker's equation to compute the focal length f of this lens as a function of R and n. Is the result consistent with a divergent lens?
2. Use the thin lens equation to compute d'. Explain, in plain English, where the image is.
3. Imagine there's a small object at A. What will an observer looking at A through the lens see? Explain your reasoning.

5.4.4 Magnification



If the object A is not on the optical axis, we define $d = HO$ and $h = HA$ where H is the orthogonal projection of A onto the optical axis. Similarly we define $d' = OH'$ and $h' = H'A'$ where H' is the orthogonal projection of the image A' onto the axis. d and d' follow the same sign convention as before and obey the same equation as before the thin lens equation):

$$\frac{1}{d} + \frac{1}{d'} = \frac{1}{f}$$

h is positive if A is above the axis and negative if it is below. The same goes for h' and A'. In the sketch above $h > 0$ and $h' < 0$. They are related to d and d' by:

$$\frac{h'}{d'} = -\frac{h}{d}$$

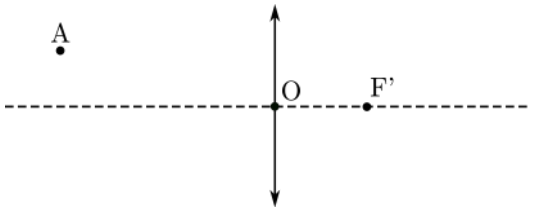
Together those two equations allow us to predict the exact location of A' (that is, d' and h') provided we know the location of A and the lens' focal length f.

Now imagine an object that extends in a line from H to A. The size of that object is h. The image of that object is the line $H'A'$, whose size is h'. The ratio of the two is called the magnification $m = \frac{h'}{h}$. If A and A' are on opposite sides of the axis then $m < 0$ and the image is said to be *inverted*. If $|m| < 1$, the

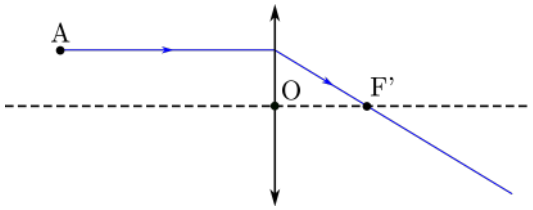
image is smaller than the object. If $|m| > 1$, it is larger. Designing a microscope is largely about creating a large $|m|$ (microscopes usually involve multiple lenses, but the idea is the same).

5.4.5 Graphical constructions

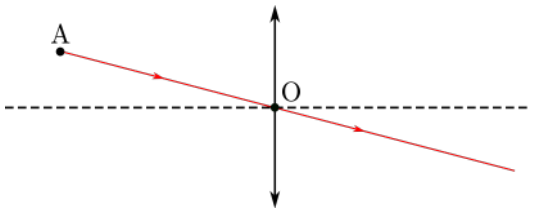
The sketch below shows an object A and a lens with center O and image focal point F':



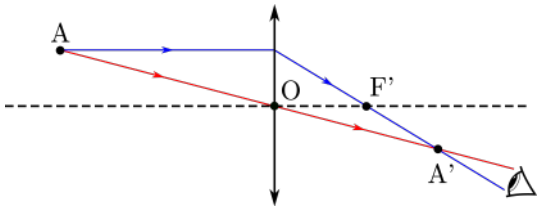
Consider the ray coming out of A parallel to the axis. Since it's parallel to the axis, it gets deflected to go through F':



Now consider a ray coming out of A and going through the center O of the lens. Near O both faces of the lens are perpendicular to the axis, therefore they are parallel to each other. In other words, near O the lens is like the slab of problem 72. A ray going through such a slab get shifted a bit but its direction after the slab is the same as before the slab. On top of that the shift is proportional to the thickness of the slab, so for a thin lens we can neglect it. Bottom line: rays going through the center O of the lens continue straight as if there was no lens at all:



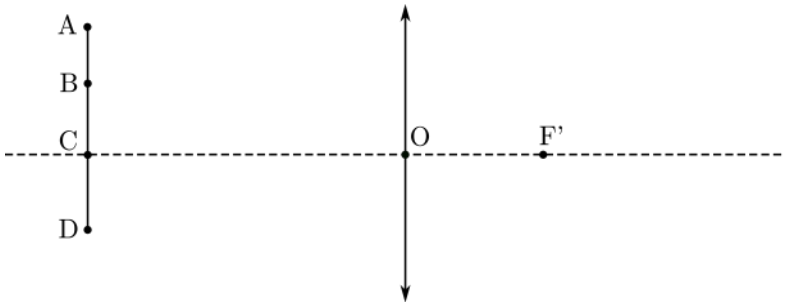
By definition of A', every ray coming from A must exit the lens looking like it's coming from (or going to) A'. In other words, A' is somewhere on the deflected blue ray, and also on the deflected red ray. In other words, A' is at the intersection of the two deflected rays:



You can check that an observer located after A' perceives two rays coming from A' and subsquently infers the object is located at A' as discussed in section 5.1.2 (perceived location of an object).

Problem 85: Image of an object perpendicular to the axis.

The rod AD is perpendicular to the axis. We want to show that its image is also perpendicular to the axis.



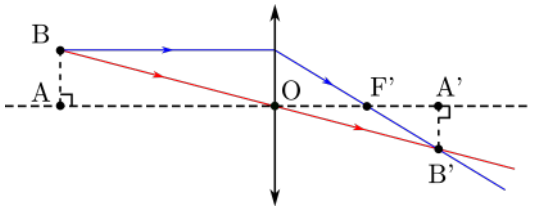
1. Use the graphical construction method to construct the images of A, B, and D. Call them A', B', C', respectively.

2. Try to use the same method to construct the image C. Explain why it doesn't work.

3. The segment AD is perpendicular to the axis. Is its image (the segment A'D') also perpendicular to the axis? Can you guess the location of the image of C?

Image of a point on the axis

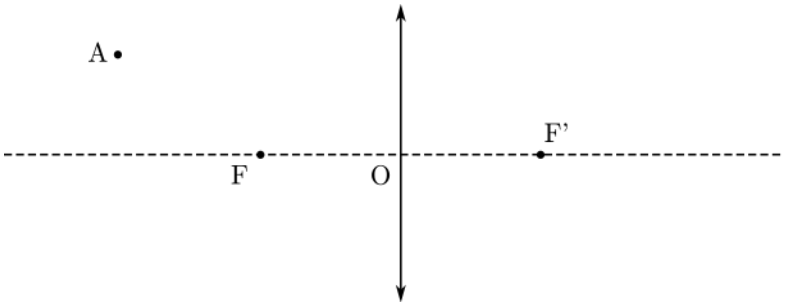
As illustrated in problem 85, the graphical construction method doesn't work when the object is a point on the axis. To get around this, we use the fact that the image of an object perpendicular to the axis is also perpendicular to the axis, which we already discussed in section 5.4.4. Imagine we want to construct the image of point A below. We would first construct a point B such that AB is perpendicular to the axis. Any point on the line perpendicular to the axis and going through A will do. Then, construct the image B' of B. Finally, construct the orthogonal projection of B' back onto the axis to get A'.



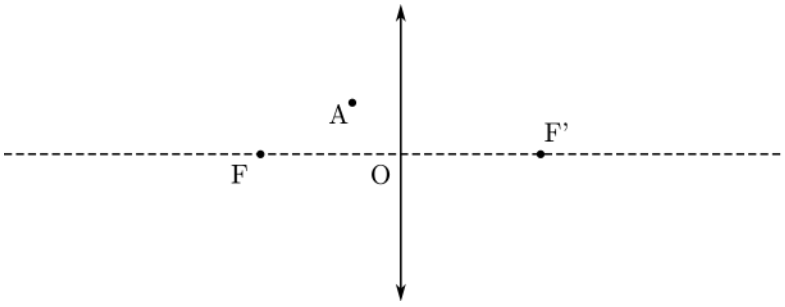
Problem 86: Graphical construction cases.

Graphically construct the image of point A in each of the following cases:

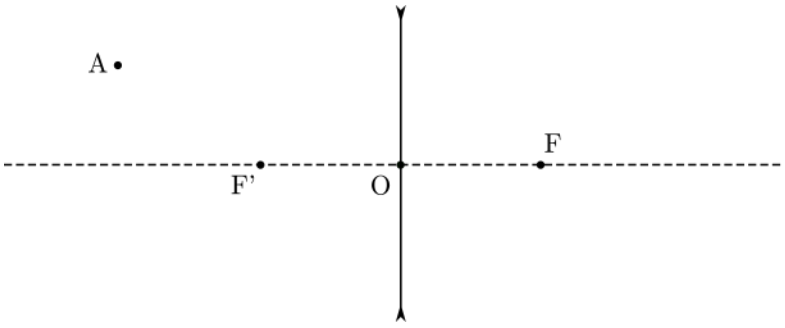
1. Convergent lens, A left of F.



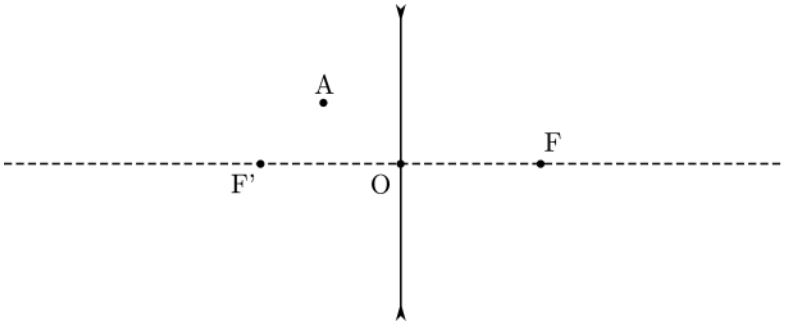
2. Convergent lens, A right of F.



3. Divergent lens, A left of F'.

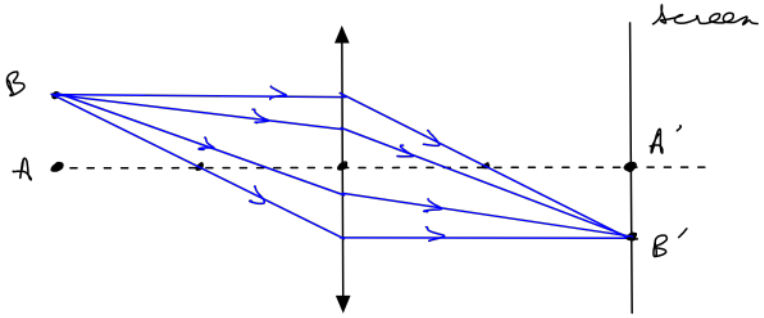


4. Divergent lens, A right of F'.



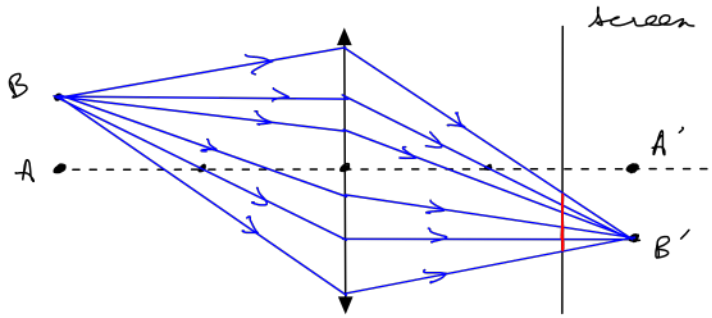
5.4.6 Making an image on a screen

Note: The principles discussed in this section are quite general. The lens could be any image-forming optical system, including the eye's lens. The screen could be the retina, a CCD, a piece of photographic paper, a wall, etc.



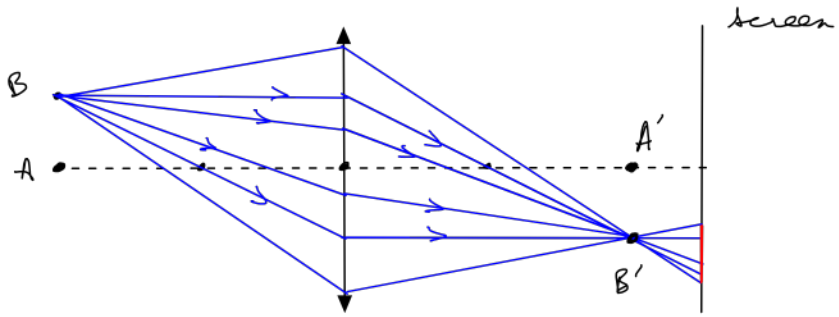
In the sketch above, every ray coming from B goes to B', and all the light at B' is coming from B (assuming there's no other source of light), therefore B' has the same color and relative brightness as B. Similarly, each point on the AB line has its own image on the screen with the same color and relative brightness as the original point. Every detail of the object, every nuance of shape and color, is reproduced in the image. That's what makes the object recognizable on the screen.

Contrast with what happens when the screen is a little too close:



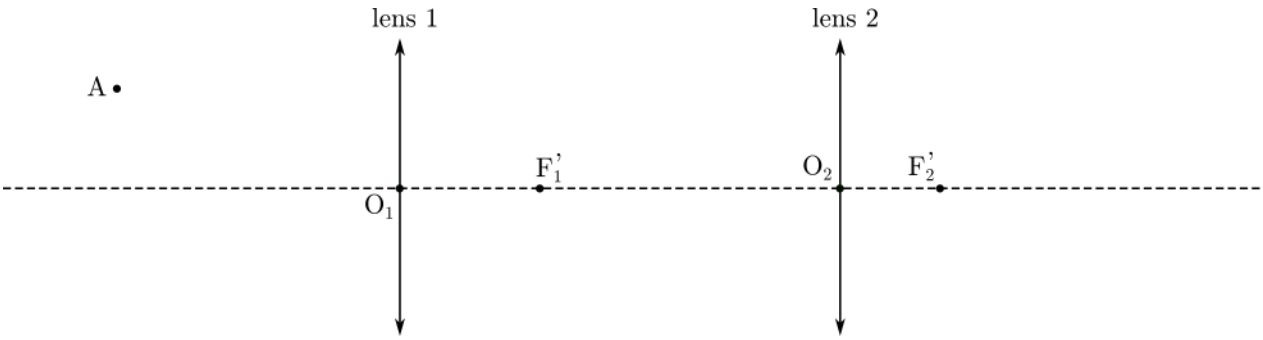
Now the light from B illuminates a spot (the red region) rather than a single point. The same goes for every other point of the object. Features of the image that are closer than the size of the spot are blended together. The image is blurry.

Something similar happens when the screen is too far:

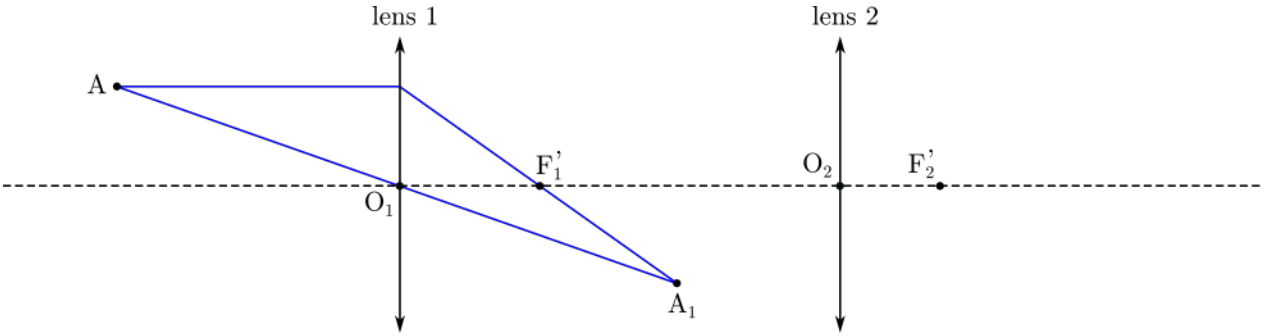


5.4.7 Multiple lenses

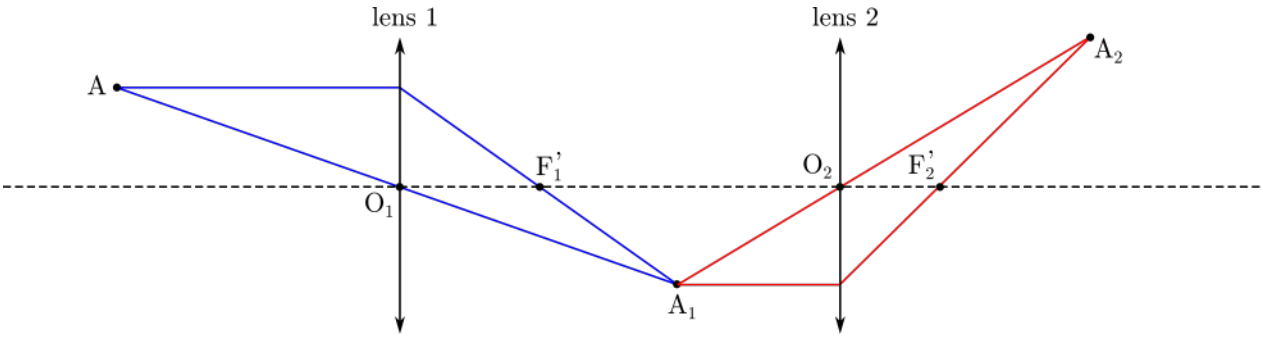
When there are multiple lenses, we apply the same strategy we used to deal with multiple mirrors. In the example below, the rays from object A go through lens 1, then lens 2.



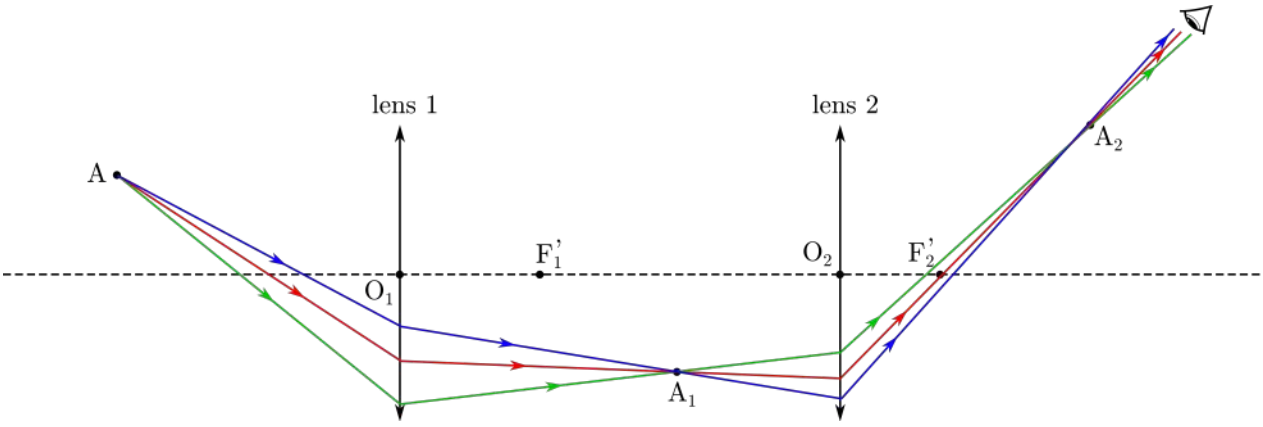
First, we construct the image of A by lens 1, which we call A_1:



Then, we construct the image of A_1 by lens 2, which we call A_2. The key point is that lens 2 has no way to know about A or lens 1. All it sees is rays coming from A_1. A_1, which was the image of A by lens 1, now plays the role of object for lens 2. Thus for this step we need to focus on A_1 and lens 2 and pretend (temporarily) that A and lens 1 don't exist.



Once we have A_1 and A_2, we can construct the path of any ray coming from A: after going through lens 1, it passes through A_1; after going through lens 2, it passes through A_2. At the end of the day, an observer located after lens 2 only sees rays coming from A_2; they don't see A or A_1, only A_2:

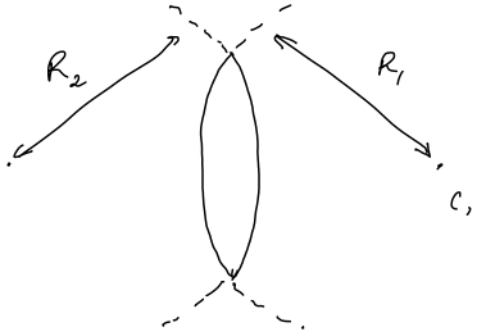


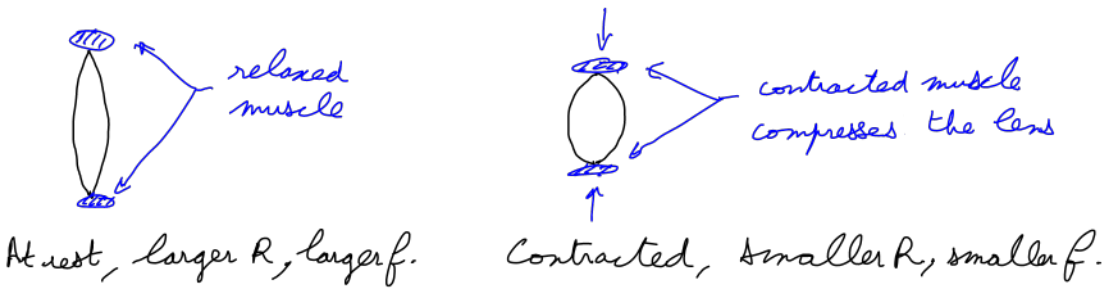
5.5 The eye

Roughly speaking the eye is a screen (the retina) and a lens like the ones in section 5.4.6. The special thing about it is that it can accommodate: the lens can adjust its focal length such that no matter how far the thing you're looking at is, its image always forms exactly on the screen.

5.5.1 Accommodation

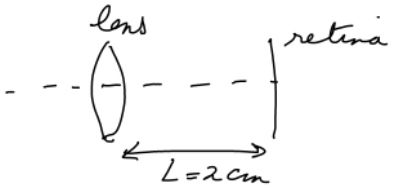
For the purpose of understanding the basics of accommodation, the eye's cornea and lens can be modeled as a single symmetric biconvex lens with curvature radius $R > 0$, i.e., $R_1 = R$ and $R_2 = -R$. It is encased in a ring-shaped muscle (the ciliary muscle). When the ciliary muscle contracts, it compresses the lens, thus decreasing R. This in turn decreases the focal length according to the lens maker's equation.





Problem 87: Normal accommodation.

We model the eye as a flat retina and a symmetric biconvex lens with refractive index $n = 1.4$ and curvature radius R located $L = 2cm$ before the retina.



1. Using the thin lens equation, compute the focal length f needed for the image of an object located at infinity to form on the retina.
2. Using the lens maker's equation, write the focal length of the lens as a function of R.
3. Compute the value R needed to achieve the focal length of question 1.
4. Answer questions 1 and 3 for an object located $50cm$ left of the lens.
5. When the ciliary muscle is fully contracted, R is $1.5cm$. This is the lowest R can go. What is the corresponding focal length? What distance does an object need to be to form an image on the retina in that case?
6. If the object is closer than that, say $20cm$, the best the lens can do is contract all the way to $R = 1.5cm$. Where is the image then? What's its position relative to the retina?

Note: The closest distance at which the eye can see clearly is called the *near point*. The further distance at which the eye can see clearly is called the *far point*. For someone with normal vision those are about $25cm$ and infinity, respectively.

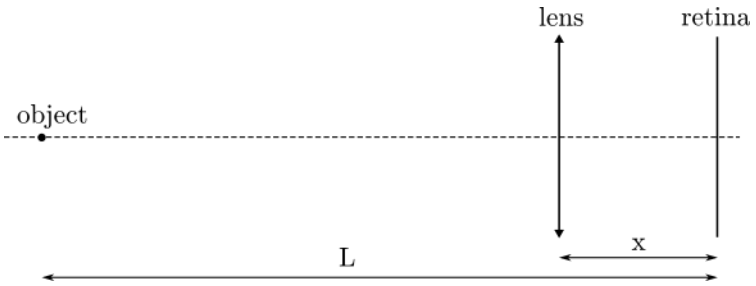
Problem 88: Defective accommodation.

We model the eye as in problem 87 except the retina is now $2.1cm$ behind the lens. The lens' curvature radius R is allowed to vary freely between 1.5 cm and 1.6 cm, but no further.

1. What focal length is needed for an object at infinity to have its image on the retina? Show that the corresponding curvature radius is not in the accessible range. What does that mean for the organism the eye belongs to?
2. When the ciliary muscle is fully relaxed, how far from the lens does an object need to be seen clearly (i.e., for its image to be on the retina)?
3. Same question when the ciliary muscle is fully contracted.
4. Write the distance d at which an object is seen clearly (image on the retina) as a function of the radius of curvature R of the lens. Show that $d(R)$ is a monotonically increasing function. Sketch it. Include the values computed in questions 2 and 3. In what range of distance d can this eye see clearly?

Problem 89: Fish accommodation.

Fish and amphibians use a different accommodation mechanism. Instead of adjusting the curvature of the lens, they adjust its distance to the retina. In the sketch below, a fish is looking at an object located at a distance L from its retina. The lens has a fixed, positive focal length f, but the fish can adjust the distance x.



1. What equation must x obey for the lens to form an image of the object on the retina? Rewrite it as a quadratic equation for x.

2. What condition must f obey for the quadratic equation to have real solutions? Can you make sense of why this condition is an upper bound (rather than a lower bound) on f?

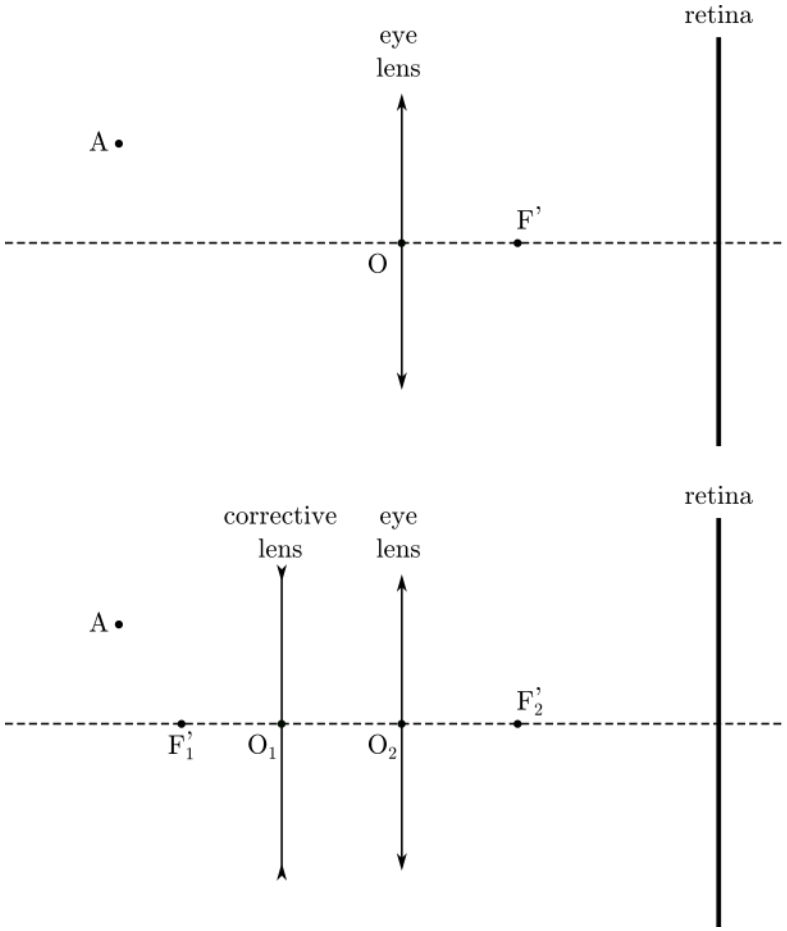
3. When there are real solutions, show that they both correspond to the lens being between the object and the retina, as it should be. Which of the two solutions makes the most sense in the context of an eye?

5.5.2 Corrective lenses

Some vision problems can be fixed by adding a lens in front of the eye's lens.

Problem 90: Corrective lens: Graphical approach.

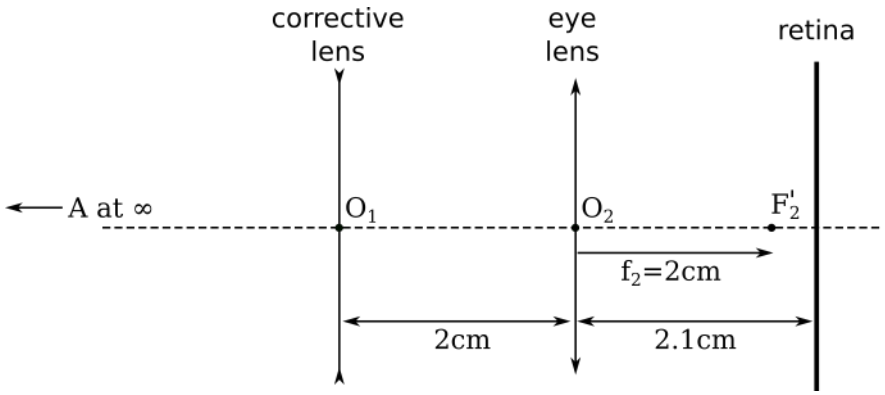
1. On the first sketch, construct the image of A by the lens. Which side of the retina is it on?
2. On the second sketch, we added a corrective lens before the eye. Is it convergent or divergent? Construct the image of A by the two lenses. Was the correction successful?
3. Can you make a qualitative case for this being the right type of lens (convergent vs divergent) when the uncorrected image A' of question 1 is on this side of the retina (before vs after the retina)?



Problem 91: Corrective lens: Analytical approach.

In a normal eye, the image of an object located at infinity ($d = \infty$) by the fully relaxed eye lens forms right on the retina ($d' =$[lens-retina distance]). In a defective eye, this is no longer the case. A good rule of thumb to determine the appropriate corrective lens is that it should correct this, i.e., the image of an object located at infinity (point A on the sketch) by the [corrective lens + fully relaxed eye lens] should form right on the retina. The purpose of this problem is to compute the corrective focal length required to achieve this.

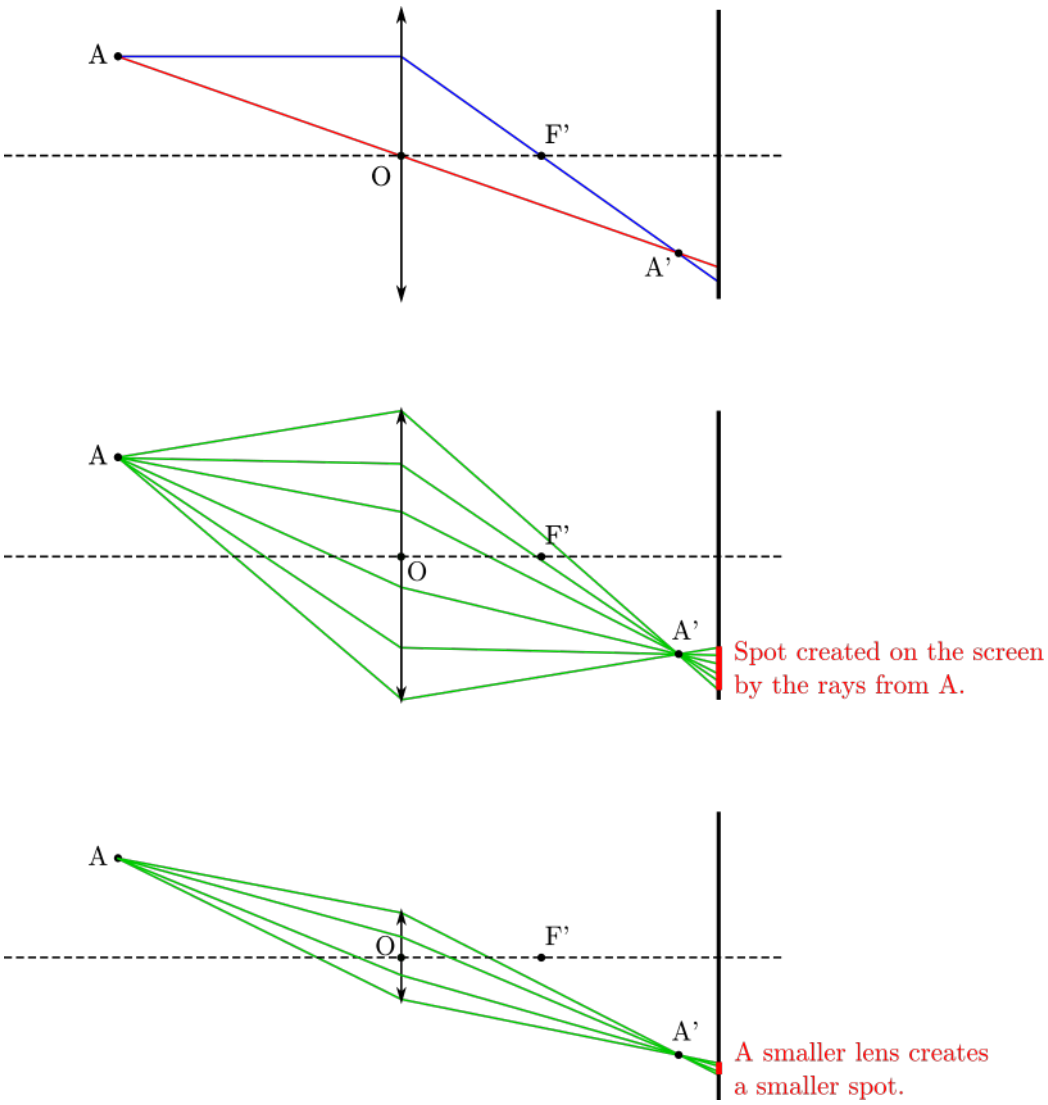
The focal length of the corrective lens is f_1. It is unknown for now; the purpose of the problem is to compute it. The focal length of the fully relaxed eye lens is $f_2 = 2\,\text{cm}$.



1. First we need to locate the image of A by the corrective lens. Let's call that image A_1. What is the (signed) distance d_1 from the object A to the corrective lens? Once you have d_1, use the thin lens equation to compute the distance d_1' from the corrective lens to A_1 as a function of f_1.
2. Next we need to locate the image of A_1 by the eye lens. Let's call that image A_2. Compute the (signed) distance d_2 from A_1 to the eye lens as a function of f_1. Keep in mind that d_1' is measured relative to O_1 whereas d_2 is measured relative to O_2. Once you have d_2, use the thin lens equation to compute the distance d_2' from the eye lens to A_2 as a function of f_1.
3. The ideal corrective lens is the one that makes A_2 be on the retina. What does that mean for d_2'? Compute the corrective focal length f_1 required to make it happen.

5.5.3 Aperture

Once we've constructed A', we can trace any ray from A going through the lens. After drawing a few of them, it should be clear that the rays going through the edges of the lens end up at the edges of the spot, and that those two therefore determine the size of the spot. Specifically, decreasing the size of the lens decreases the size of the spot.



The smaller the spot, the less blurry the object. Therefore, a small lens allows to make reasonably clear images of objects that are not quite at the right distance for their image to be exactly on the screen. If we define a maximum spot size above which we call something blurry, decreasing the size of the lens increases the range of distances at which an object can be without being blurry.

On the other hand, a larger lens lets more light through, creating a brighter spot. Therefore, the ideal lens size results from a trade-off and needs to be adjusted to (1) what is beeing observed and (2) the light conditions.

Rather than resizing the lens, it's possible to put a circular mask right in front of it that is smaller than the lens. In cameras, the part that creates the tunable hole is called a diaphragm. In the eye, it's called the iris.


f/1.4
f/2.0
f/2.8
f/4.0
f/5.6
f/8.0