A Classical Physics Review for Modern Physics

This material is written for the student taking modern physics. It is intended as a review of general principles of classical physics, concentrating on topics most important to modern physics, some of which may not have been emphasized in the student's classical physics course.

Contents

Edition 1.0

Conserved Quantities

Some of the most important ideas in physics are conservation laws. In introductory physics we learn of three important conserved quantities: energy, momentum and angular momentum. For these quantities to be conserved, of course, the system must be isolated, so that they are not altered by external influences. It is easily overlooked that these quantities are important *because* they are conserved. The quantity mgy, for instance—celebrated as gravitational potential energy—would be of little interest to physicists if a ball thrown upward could speed up, that is, if energy were not conserved. It has meaning when we say that the ball must slow down—lose kinetic energy—because it is gaining potential energy. We cannot prove that these special quantities are conserved; conservation laws, like all physical laws, are fundamental things we *observe* to hold true, not things we prove or derive from more-basic ideas. We accept the laws then build upon the foundation.

*   *   *   *

Energy

In an isolated system, total energy neither increases nor decreases. While familiarity surely makes the idea seem reasonable, verification requires that we account for energy in all its forms, and this is no simple matter. Energy is a most varied conserved quantity. Indeed, whenever experiments have appeared to show a loss or gain, strong belief in the law has lead to new, previously hidden forms being postulated—*and found*. Still, we can go a long way by considering only a few of the most prevalent forms.

Kinetic

The energy of motion of an object of mass m and speed v, called kinetic energy, is given by:

$$\mathrm{KE} = \tfrac{1}{2}\, m\, v^2 \qquad (1)$$

As is the case for many of the formulas presented in this classical physics review, equation (1) is correct only if the speed is a small fraction of the speed of light.

Objects can have another form of kinetic energy, associated not with motion as a whole in a given direction, but with rotational motion about an axis: rotational kinetic energy. This depends on the object's **angular velocity** ω, in radians per second, and on how its mass is distributed, specified by the object's **moment of inertia** I. The functional dependence of rotational kinetic energy is very similar to that of ordinary, or "translational," kinetic energy (as are all rotational and translational analogs). In place of v is ω and in place of m is I:

$$\mathrm{KE}_{\mathrm{rot}} = \tfrac{1}{2}\, I\, \omega^2 \qquad (2)$$

Moment of inertia is calculated via $I = \Sigma\, m_i\, r_i^2$, where m_i is the mass of a tiny particle of the object, r_i is its distance from the rotation axis, and the sum—or integral, if the mass distribution is continuous—is taken over all the particles that make up the object. Common examples are: a thin, hollow ring, where essentially all the mass is the same distance R from the axis, giving $I = M_{\mathrm{total}}\, R^2$; and a solid sphere of radius R, for which integration shows that $I = \tfrac{2}{5}\, M_{\mathrm{total}}\, R^2$.

Potential

Anytime a conservative force acts there will be associated with it a potential energy, symbol U, and the functional form of the potential energy depends on the functional form of the force. A few of the more common examples follow.

The gravitational force near Earth's surface is given by $\mathbf{F} = -mg\,\hat{y}$, where m is an object's mass, g is the acceleration due to gravity and $\hat{y}$ is a unit vector in the upward direction. Associated with this force is gravitational potential energy:

$$U_{\text{grav}} = mgy$$

An ideal spring exerts a force $F = -\kappa x$, where κ is the spring's force constant and x is its displacement from its equilibrium length. Associated with this force is elastic potential energy:

$$U_{\text{elastic}} = \tfrac{1}{2}\kappa x^2$$

The force exerted on point charge q_2 by point charge q_1 is $F = \frac{1}{4\pi\epsilon_o}\frac{q_1 q_2}{r^2}$, where $\frac{1}{4\pi\epsilon_o}$ is a fundamental constant of electrostatics and r is the separation between the charges. Associated with this force is electrostatic potential energy:

$$U_{\text{point charges}} = \frac{1}{4\pi\epsilon_o}\frac{q_1\,q_2}{r}$$

As the reader may verify, all these potential energies may be found from their forces via the same general relationship.

$$U_2 - U_1 = -\int_1^2 \mathbf{F}\cdot\mathbf{d}\boldsymbol{l} \qquad (3)$$

The differential displacement dl is dy, dx, dr, etc., as the case may be. Equation (3) says that the change in potential energy is the negative of the **work** done by the force (not the work we might have to do "against" the force). Gravity, for example, does *positive* work speeding up a falling ball; meanwhile the change in the gravitational potential energy is *negative*. As we see, only a *difference* in U is determined; we are free to choose U to be zero at a convenient location.

Internal: Thermal and Chemical

One of the less conspicuous forms of energy is internal energy. For most classical purposes (i.e., when "mass energy" does not vary) it is the sum of the microscopic kinetic and potential energies of all the atoms or molecules in an object. An object's *overall* kinetic and potential energies can be converted into increased random internal energy, as when a book is dropped to the floor. But the second law of thermodynamics says that a complete reversal, turning this random thermal energy completely into ordered overall kinetic or potential energy, is impossible. Chemical reactions can convert internal energies from one form to another, often making an object feel hotter or colder. They can also produce more visible results, like the lifting of an eyelid or the explosion of a firecracker. Given its seeming complexity, it isn't surprising that classical physics has no comprehensive formula for internal energy—though modern physics does.

Momentum—and Force, its time rate of change

Energy is a scalar. There is a physical principle that says that for each "symmetry" there will be a conserved quantity. We won't investigate this interesting principle; we simply use it to give some insight into the next conserved quantity. Our observations indicate that the physical universe obeys the same laws—behaves the same way—at one instant in time as at the next. The consequence of this "symmetry in time" is energy conservation. The physical universe also seems to obey the same laws if we stand at one location or at another. The consequence of this "symmetry in space" is another conservation law. But since space is three-dimensional, so is the conserved quantity: Momentum is a vector.

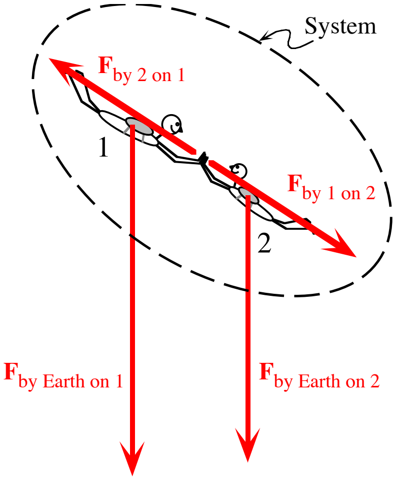

System
$\mathbf{F}_{\text{by 2 on 1}}$
$\mathbf{F}_{\text{by 1 on 2}}$
1
2
$\mathbf{F}_{\text{by Earth on 1}}$
$\mathbf{F}_{\text{by Earth on 2}}$

Figure 1

Of several approaches to finding a conserved quantity besides energy, Figure 1 shows a very basic example: a system consisting of two falling objects—skydivers. Acting on each object is an interparticle force, shared between things internal to the system, and a gravitational force, due to something external the system (i.e., Earth).[1] Newton's second law of motion $\Sigma\mathbf{F} = m\mathbf{a}$ applies to each object, so we can add the two equations.

$$\begin{array}{rl} & \mathbf{F}_{\text{by 2 on 1}} + \mathbf{F}_{\text{by Earth on 1}} = m_1\mathbf{a}_1 \\ + & \mathbf{F}_{\text{by 1 on 2}} + \mathbf{F}_{\text{by Earth on 2}} = m_2\mathbf{a}_2 \\ \hline \end{array}$$

$$\mathbf{F}_{\text{by 2 on 1}} + \mathbf{F}_{\text{by Earth on 1}} + \mathbf{F}_{\text{by 1 on 2}} + \mathbf{F}_{\text{by Earth on 2}} = m_1\mathbf{a}_1 + m_2\mathbf{a}_2$$

By Newton's third law of motion, the forces the objects exert on one another are equal and opposite, so they cancel, leaving only the forces involving things external to the system. Noting that acceleration is the time rate of change of velocity, we have

$$\Sigma\,\mathbf{F}_{\text{external}} = m_1\frac{d}{dt}\mathbf{v}_1 + m_2\frac{d}{dt}\mathbf{v}_2 = \frac{d}{dt}\left[m_1\mathbf{v}_1 + m_2\mathbf{v}_2\right] \qquad (4)$$

Suppose now that external forces are somehow switched off.

$$\frac{d}{dt}\left[m_1\mathbf{v}_1 + m_2\mathbf{v}_2\right] = 0 \qquad \text{No external force}$$

Though $\mathbf{v}_1$ and $\mathbf{v}_2$ might still change due to the forces the objects exert on one another, the total quantity in brackets cannot change with time—it is conserved. The momentum **p** of an object we define as follows:

$$\mathbf{p} = m\mathbf{v} \qquad (5)$$

Thus, equation (4) says that the total momentum—the sum over all the particles—changes due to *external* forces only, and in the absence of external forces the total momentum is conserved. Once again, this conserved quantity is clearly a vector.

Newton's second law of motion is properly written as:

$$\Sigma\mathbf{F} = \frac{d}{dt}\mathbf{p} \qquad (6)$$

This form is more general. It can be applied not only to a single object of unchanging mass, reverting to $\Sigma\mathbf{F} = m\mathbf{a}$; it is also applicable to a *system* of particles, where $\Sigma\mathbf{F}$ is the net *external* force only and **p** is the total momentum. Internal forces may be huge—sending particles off at great speeds, breaking chunks off some particles, fusing them to others, and so on—but if the net external force is zero the total momentum cannot change. Equation (6) is also applicable in cases, such as things moving at light speed or close to it, where the expression m**v** for momentum is simply not correct and the concept of acceleration less clearly applicable.

Angular Momentum—and Torque, its time rate of change

The physical universe seems to behave the same way whether we are facing North or East or any other direction. This rotational symmetry is at the root of another conserved quantity: angular momentum. The usual way of introducing this principle is in connection with a rigid object.

Once we graduate from studying "point objects" to studying "real" objects that occupy space we encounter difficulty in trying to apply only Newton's second law $\Sigma\mathbf{F} = m\mathbf{a}$. It is true that knowing the net external force on a object of mass m gives the acceleration $\mathbf{a}_{\text{cm}}$ of its center of mass, but there is something else a "real" object may do—it may rotate. We can, for instance, exert equal and opposite forces at the edges

[1] We ignore air resistance. Perhaps the skydivers are not yet falling fast enough for it to be a factor.

of a wheel, as shown in Figure 2, and the second law becomes 0 = 0, which isn't very useful. How can we quantify what *is* important—how the object as a whole rotates? Clearly it has something to do with how the forces exerted at the edges are communicated from molecule to molecule throughout the object, but is such analysis feasible? We state without proof that when dealing with a rigid object, the happy alternative to applying $\Sigma\mathbf{F} = m\mathbf{a}$ to every molecule in the object is to apply to the object as a whole $\Sigma\mathbf{F}_{\text{external}} = m_{\text{total}}\,\mathbf{a}_{\text{cm}}$ and one other "law": $\Sigma\boldsymbol{\tau}_{\text{external}} = I_{\text{cm}}\,\boldsymbol{\alpha}_{\text{cm}}$, where I is the object's moment of inertia and the "cm" subscripts mean that the rotation axis passes through the center of mass. The latter is usually called the rotational second law. With some effort it can be derived from the ordinary, or "translational," second law and the assumption of a rigid object, but our interest is in its use. As force $\mathbf{F}$ is the cause and rate of change of velocity, $\mathbf{a} = \frac{d\mathbf{v}}{dt}$, the effect, torque $\boldsymbol{\tau}$ is the cause whose effect is an angular acceleration, a rate of change of angular velocity, $\boldsymbol{\alpha} = \frac{d\boldsymbol{\omega}}{dt}$. The forces shown in Figure 2 will not cause the wheel's center of mass to accelerate; they will cause its angular velocity to change with time.

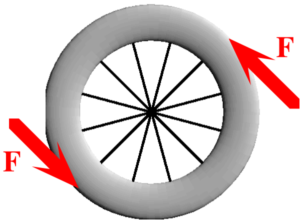


Figure 2

Causing not acceleration, but *angular* acceleration, torque is not force. It does require a force, but the force's point of application and orientation relative to the rotation axis are crucial factors. We don't swing open a door on rusty hinges by pushing at the door's inner edge, along the axis (hinges); we push at the outer edge. Even then, we don't exert the force parallel to the door, directly toward the axis or away, nor upward nor downward; we push perpendicular to the door's surface. The formula for torque takes all these things into account.[2]

$$\boldsymbol{\tau} = \mathbf{r}\times\mathbf{F} \qquad\qquad |\boldsymbol{\tau}| = |\mathbf{r}|\,|\mathbf{F}|\sin\theta \tag{7}$$

The moment arm $\mathbf{r}$ extends from the rotation axis to the point of application of the force $\mathbf{F}$, and θ is the angle between $\mathbf{r}$ and $\mathbf{F}$. If $\mathbf{r}$ is zero (pushing at the axis) there is no ability to change the angular velocity. If $\mathbf{F}$ is opposite or parallel to $\mathbf{r}$ (toward or away from the axis, i.e., θ = 180° or 0°) the torque is also zero. To affect rotation about a given axis there must be a nonzero torque component along that axis (explaining why pushing up or down doesn't open the door—$\mathbf{r}\times\mathbf{F}$ would be perpendicular to the axis).

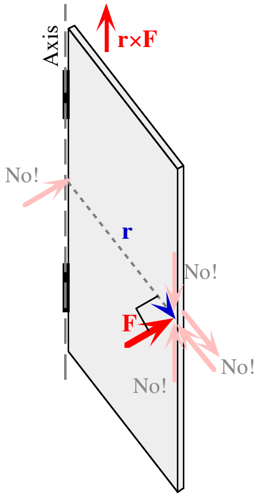


Figure 3

If torque $\boldsymbol{\tau}$ is a vector, so must be angular acceleration. All of the rotational laws we use are in accord with a convention that assigns the direction of the angular *velocity* $\boldsymbol{\omega}$ by a right-hand rule: Curling the fingers of your right hand in the direction of the rotation, your thumb gives the direction of $\boldsymbol{\omega}$, perpendicular to the plane of rotation, as shown in Figure 4a. If the magnitude of a vector *increases*, its change is parallel to itself, so the angular acceleration $\boldsymbol{\alpha}$, rate of change of $\boldsymbol{\omega}$, is parallel to $\boldsymbol{\omega}$ if ω increases. In Figure 4b, both torques $\boldsymbol{\tau} = \mathbf{r}\times\mathbf{F}$ are upward along the axis, and assuming that the wheel is already rotating counterclockwise, this net torque would speed it up, giving an angular acceleration in the direction of $\boldsymbol{\omega}$ and thus in the same direction as the net torque. Were the wheel initially rotating clockwise, $\boldsymbol{\omega}$ would be down along the axis, but the torque would then slow the wheel down, meaning that $\boldsymbol{\alpha}$ would be opposite $\boldsymbol{\omega}$. Again, $\boldsymbol{\alpha}$ would be up along the axis, the same direction as $\boldsymbol{\tau}$, as $\Sigma\boldsymbol{\tau}_{\text{external}} = I_{\text{cm}}\,\boldsymbol{\alpha}_{\text{cm}}$ says.

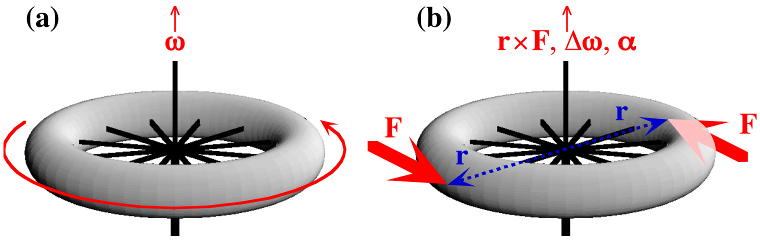


Figure 4

[2] Here we use the standard right-hand rule for cross products. Point the fingers of your right hand in the direction of the first vector, $\mathbf{r}$, then orient your palm so that it is easiest to curl your fingers toward the direction of the second, $\mathbf{F}$. Your thumb gives the direction of the cross product $\boldsymbol{\tau}$

The general statement of the translational second law is $\Sigma\mathbf{F} = \frac{d}{dt}\mathbf{p}$. It becomes the more familiar $\Sigma\mathbf{F} = m\mathbf{a}$ only if the momentum $\mathbf{p}$ is correctly given by $m\mathbf{v}$ and m doesn't vary with time. There is an analogous general statement of the rotational second law:

$$\Sigma\boldsymbol{\tau} = \frac{d}{dt}\mathbf{L} \qquad (8)$$

$\mathbf{L}$ is known fittingly as **angular momentum**. Therefore, it is most general to say that a torque causes not an angular acceleration but a rate of change of angular momentum. In many cases, $\mathbf{L}$ is given by $I\boldsymbol{\omega}$ (analogous to $\mathbf{p} = m\mathbf{v}$). It thus has the same direction as $\boldsymbol{\omega}$, and if I doesn't change with time we recover the formula $\Sigma\boldsymbol{\tau} = I\boldsymbol{\alpha}$. However, it is perhaps even more important to keep in mind the general statement in the rotational case than in the translational, for even if an object's mass doesn't change it can still have a changing moment of inertia. When a spinning ice skater pulls her arms inward to her body, she does not change her mass, but she does change her mass *distribution*. External torques due to air resistance and friction with the ice are minimal, but the ice skater does angularly accelerate; her angular velocity increases because her moment of inertia decreases, and the product, $\mathbf{L} = I\boldsymbol{\omega}$, cannot change without an external torque. That an object may "speed up all by itself" troubles some students, but the system is simply doing what fundamental laws of physics say it should do.[3]

As $\Sigma\mathbf{F} = \frac{d}{dt}\mathbf{p}$ is general, applicable to systems of particles and when $\mathbf{p}$ is not given by m$\mathbf{v}$, the formula $\Sigma\boldsymbol{\tau} = \frac{d}{dt}\mathbf{L}$ is applicable to systems of particles and when $\mathbf{L}$ is not $I\boldsymbol{\omega}$. Many expressions for angular momentum are equivalent to $I\boldsymbol{\omega}$. For example, if the rotation axis passes through the origin, a point mass m whose position vector is $\mathbf{r}$ and whose momentum is $\mathbf{p}$ has an angular momentum $\mathbf{L} = \mathbf{r} \times \mathbf{p}$. This has the same direction as $\boldsymbol{\omega}$ (out of the page in Figure 5) and its magnitude also agrees with $I\boldsymbol{\omega}$: $|\mathbf{r} \times \mathbf{p}| = r\, mv \sin\theta = r\, m\, v_{\text{tangent}} = r\, m\, \omega\, r = (mr^2)\, \omega$, and mr^2 is indeed the moment of inertia of a point mass a distance r from the axis. However, there are cases where angular momentum simply cannot be written as $I\boldsymbol{\omega}$. A good example is the "spin" angular momentum fundamental to the electron. Another is the photon—particle of light—which has no mass(!), yet it too has "spin" angular momentum. When discussing microscopic systems of these particles, we cannot use $\Sigma\boldsymbol{\tau} = I\boldsymbol{\alpha}$. The correct law is $\Sigma\boldsymbol{\tau} = \frac{d}{dt}\mathbf{L}$. In a system isolated from external torques, the total angular momentum, in whatever strange forms it may be found, is conserved.

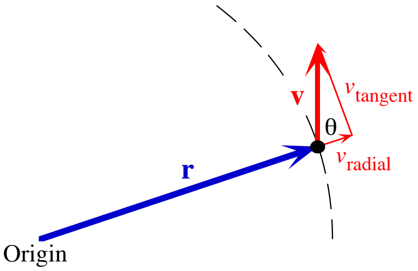


Figure 5

Are They Laws, or Aren't They?

Our discussions of momentum and angular momentum conservation sound like derivations: From the second law of motion and its corollary the rotational second law, we "showed" that in a system isolated from external forces and torques, momentum and angular momentum are conserved. But if one thing—a conservation law—can be derived from another—the second law of motion—then the first is not a law at all. Laws are things we *cannot* derive—simply the way things are. We won't devote much space to the philosophical arguments behind all this. The bottom line is that it is the conservation laws that we consider true laws. But if so, aren't the foregoing discussions backward—shouldn't we be deriving the second "law" of motion from the conservation laws? It isn't quite that straightforward. Accepting on the basis of laborious observation that a particular quantity is conserved doesn't necessarily indicate how it might be exchanged, or at what rate. This is where the second law comes in. Today we view it as a *definition*, in harmony of course with the accepted conservation laws: Momentum is conserved, but it is also quite obvious that it can be exchanged between two things not isolated from each other, and the rate at which it is exchanged, $d\mathbf{p}/dt$, we *define* to be the force between them. Moreover, if it is indeed true that momentum is conserved at every instant, then the rate at which the momentum of one object changes in one direction ($d\mathbf{p}_1/dt$) must be the rate at which the momentum of the other changes in the opposite direction ($d\mathbf{p}_2/dt$), so this definition and momentum conservation combined encompass Newton's *third* law of motion: that the mutual forces (rates of

[3] No violation of *energy* conservation results. The skater's rotational kinetic energy does increase, but this comes from her internal chemical energy, which she must expend to pull her arms inward.

change of momentum) two objects exert on each other are equal and opposite. Similarly, the rate of change of an object's angular momentum is by definition the torque on the object, and if angular momentum is indeed always conserved, mutual torques must be equal and opposite. In this light, the foregoing momentum and angular momentum conservation discussions should be viewed as merely showing the harmony between the conservation laws and more familiar "laws" of motion.

Precession

The behavior of a spinning object can be very complex. However, owing to its importance in modern physics, we close with a simplified discussion of one aspect of this behavior: precession. To begin, consider a wheel *not* spinning, whose axle lies in the *y*-*z* plane, as shown in Figure 6a. Were we to push as indicated on the ends of its axle for a time *dt*, we would exert a torque in the *x*-direction (out of the page, by the right-hand rule) and the axle would rotate counterclockwise, giving the whole assembly a small angular momentum $d\mathbf{L}$ in the *x*-direction. Does this fit with $\Sigma\boldsymbol{\tau} = d\mathbf{L}/dt$? This equation does *not* say that the angular momentum $\mathbf{L}$ is in the direction of the net torque $\Sigma\boldsymbol{\tau}$, but that the *change* in angular momentum $d\mathbf{L}$ is in the direction of $\Sigma\boldsymbol{\tau}$. (Similarly, $\Sigma\mathbf{F} = m\, d\mathbf{v}/dt$ says that $\Sigma\mathbf{F}$ and $\mathbf{a}$ are in the same direction, not $\mathbf{F}$ and $\mathbf{v}$.) Nevertheless, the torque, causing a $d\mathbf{L}$ in the *x*-direction, does cause the new angular momentum $\mathbf{L}_f$ to be in the *x*-direction because *in this case the initial angular momentum* $\mathbf{L}_i$ *was zero*.

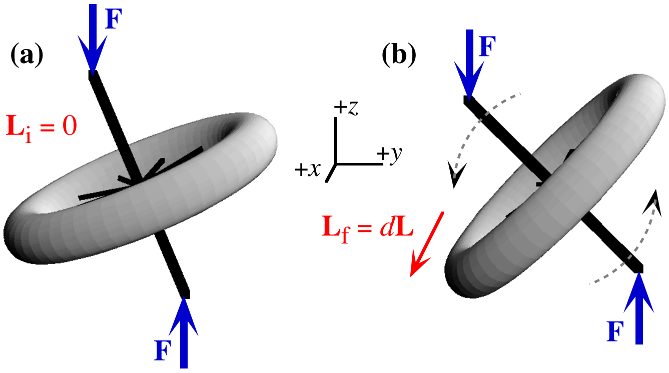
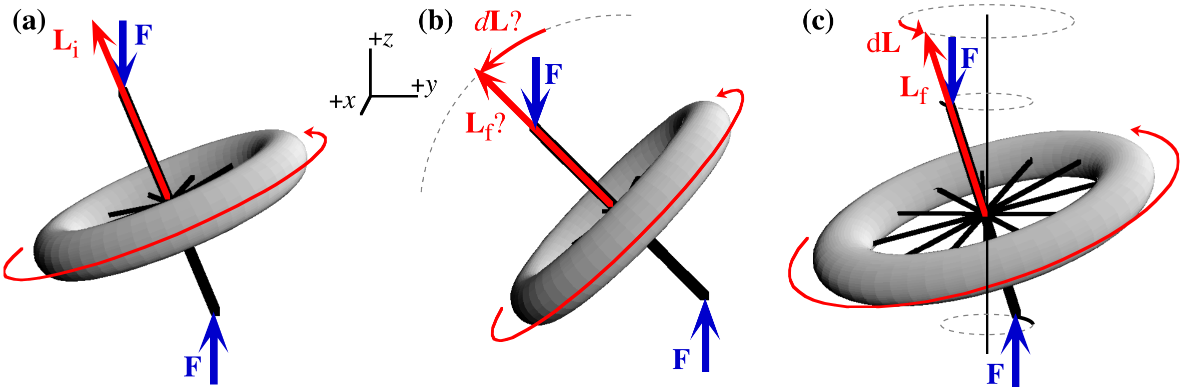


Figure 6

But what if the wheel is already spinning? Suppose it is spinning with a large angular momentum parallel to its axle, as shown in Figure 7a, *with no x-component*. Were its axle to rotate counterclockwise as before, the change in the angular momentum vector would be in the direction labeled $d\mathbf{L}$? in Figure 7b—the *y*- and *z*-components of $\mathbf{L}$ would change. *This cannot be!* $d\mathbf{L}$ must be in the direction of the torque, *its cause*. In the diagram this is the *x*-direction, so the wheel assembly must move in such a way that the new angular momentum vector has the same *y*- and *z*-components, merely acquiring a small component in the perpendicular *x*-direction. By simply redefining our coordinate system, the same could be said at any instant of time: The change $d\mathbf{L}$ is always perpendicular to $\mathbf{L}$. And if the change in a vector is always perpendicular to the vector, the vector's magnitude can never change—it simply changes direction (the velocity vector in circular orbit is analogous). Thus, a torque perpendicular to $\mathbf{L}$ causes $\mathbf{L}$ to maintain a given angle with an axis (here the *z*-axis), preserving its components parallel and perpendicular to that axis while continuously rotating about it. This is what we refer to as precession.



Impossible!

Precession

Figure 7

Thermodynamics

One of the first things we must confront in trying to understand the physical world is systems of countless particles whose microscopic behaviors are beyond our control. We call such systems **thermodynamic systems**, and the number of particles is huge. The number of air molecules in a typical classroom, for example, is about 10^{28}. In a modern computer microchip, even just one of the millions of transistors—a micrometer on a side—contains many millions of silicon atoms. We simply cannot know, much less control, what each of the atoms or molecules is doing. Yet, there are limitations on overall behaviors in thermodynamic systems. The most far-reaching are the fundamental laws of thermodynamics.

* * * *

Energy

Unlike the looming potential energy of a boulder teetering on the edge of a precipice or the obvious kinetic energy it gains if it falls, the random energies internal to a thermodynamic system are neither easily seen nor easily cataloged. As a place for other forms of energy to end up, internal energy seems quite hospitable—all the kinetic energy of a falling book may become invisible internal energy in the book and floor in the blink of an eye. But it is randomly distributed among countless atoms, so it tends to be viewed differently. Indeed, in an important sense it is different, as we see when we discuss the second law of thermodynamics. Still it is energy, and energy conservation has made sense as an accepted law of physics only when we take internal energy into account.

Yet, if it is distributed in microscopic forms that may be beyond our abilities to catalog directly, how indeed can we account for internal energy? We do so by deduction. Perhaps we cannot know the actual value of the internal energy, but if we can measure what goes in and what comes out, and if energy is really conserved, then determining the *change* in the internal energy is simply a matter of arithmetic. A thermodynamic system exchanging no matter with its surroundings can exchange energy in two ways: via work, usually involving a macroscopic displacement of the system in tandem with its surroundings, such as the pushing of a piston; and via heat, conveyed microscopically through collisions or electromagnetic radiation.

Change in internal energy = energy added as heat + energy added as work

Definitions may differ in sign, but if we define Q as heat *added* to the system (which is negative if the system transfers heat to the surroundings) and W as the work done *by* the system (which is negative if the surroundings transfer energy to the system in the form of work), then we have what is usually referred to as the first law of thermodynamics:

$$\Delta E_{\text{internal}} = Q - W \tag{9}$$

For example, if we add heat Q = 70J to a system of gas molecules in a cylinder—energy in—and the gas does W = 20J of work in expanding to lift the weight of a piston—energy out—the internal energy of the gas must have risen by 50J. Of course this is only a *statement* of energy conservation. Nevertheless, despite our inability to see it, countless experiments, adding and subtracting energy in every way imaginable, have confirmed that the internal energy of a system doesn't change all by itself; it changes precisely according to how much energy goes in and how much goes out.

Entropy

Students of introductory physics are often bombarded with seemingly different statements of the second law of thermodynamics, based on seemingly different ideas. There is one unifying idea. Given fixed values of certain overall properties of a system—its total energy, its volume and the number of particles—a disordered arrangement of particles and energy is more probable than an ordered one; the system will thus most likely move toward the more disordered state, and if the number of particles is huge, the probability of it doing otherwise is virtually zero. In short: Isolated thermodynamic systems are extremely unlikely to move toward more-ordered states.

The system shown in Figure 8 is a simple one, twenty gas molecules in a "room" divided by an imaginary line into two halves. The distribution in Figure 8a is very ordered—all the molecules on one side—and, *directly related to this,* it is highly improbable that this is how we would find the particles distributed. Figure 8b is a more probable distribution, 10 on each side, and it is more disordered. Having no knowledge of the system's history, we would certainly guess that if we could glimpse for an instant the distribution of molecules it is much more likely that they would be about evenly distributed than all on one side. Perhaps we could intervene somehow, and start the molecules out all on the right-hand side, but looking in later on a system left to itself (i.e., isolated) we would probably find the molecules spread more or less evenly. Actually, with only twenty particles, it is not out of the question that they might be unevenly distributed. In fact, the probability that the number on one side is between 9 and 11 (inclusive) is only about fifty percent. But with the number of particles increased it becomes exceedingly likely that the particles will be in an evenly distributed—that is, more disordered—state. For "only" 2000 particles, the probability that the number on one side is between 900 and 1100 is 0.9999931. In true thermodynamic systems, with countless particles, ordered states are so improbable that we proclaim it a law: Isolated thermodynamic systems *don't* move toward more-ordered states.

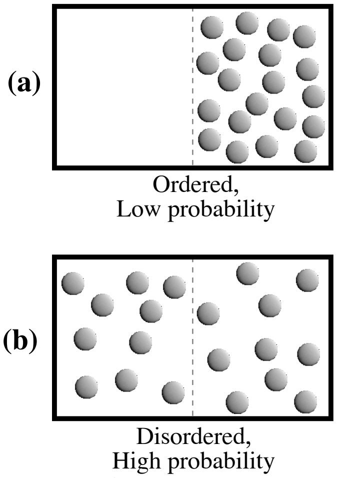


Figure 8

A system's overall kinetic energy is relatively ordered; internal motion aside, all the atoms of a moving object have the same average velocity. A system's overall potential energy is also ordered; all the atoms of a boulder teetering on the edge of a precipice are about the same distance from the bottom. Internal thermal energy is disordered. When a book falls to the floor, initially it has ordered potential energy, then ordered kinetic energy, then—thump!—increased random thermal energy of countless atoms now moving every which way faster than they were before. Although there is a mathematical chance of the atoms in the book and floor colliding in such a way that the book and floor cool down while the book springs upward (conserving energy, at least), it is so extraordinarily unlikely that we are safe in declaring that it will not happen.

Other common statements of the second law follow. A system of randomly mixed fast and slow molecules will not by itself separate into a comparatively ordered state with fast at one place and slow at another. In other words, we cannot expect to see a mass of warm material in a room by itself separate into cooler material inside a box and warmer surroundings outside the box. Such would be an "ideal refrigerator," spontaneously cooling the inside while warming the kitchen around it with the heat removed. No matter how clever or intricate the device, the net result is starting with mixed-up disorder and ending with separated order. It cannot be done! Of course we can build a refrigerator, but it must be plugged in, and the electrical energy coming into the device—a very ordered form of energy—ends up converted to random, disordered thermal energy in the kitchen. This conversion increases the disorder of the entire system, now encompassing the electrical energy source, by at least as much as the separation of hot and cold decreases it. How do we know? We accept the second law. Similarly, we cannot expect to build a device that converts random internal energy completely into useful work, into ordered overall kinetic or potential energy—an "ideal heat engine." A real heat engine does run on random thermal energy, but invariably some of the energy is rejected into cooler surroundings, increasing its disorder necessarily by at least as much as the partial conversion to ordered kinetic or potential energy decreases it.

Quantifying disorder is of course a different question. As already suggested, it is done through probabilities. An ordered state is one of low probability, and a high disorder corresponds to a high probability. All we have said about disorder is borne out by defining the measure of disorder, known as **entropy** S, as:

$$\text{Entropy, the measure of disorder:} \qquad S \equiv k_B \ln \mathrm{W} \tag{10}$$

where k_B is the **Boltzmann constant**, pervasive in thermodynamic relationships, and W is the number of microscopic ways in which the particle/energy distribution can be obtained,[4] which is in turn directly related to the probability of the distribution occurring. With this definition the second law of thermodynamics—disorder doesn't decrease—is usually written as:

$$\Delta S \geq 0 \qquad \text{(Isolated system)} \tag{11}$$

Being related to probability, entropy has meaning only for thermodynamic systems, for there would be no need to speak of *probabilities* were it not for the random internal motion beyond our control.

Temperature

Related to entropy is another physical property known to all: temperature. Everyone has a notion of what temperature means, but in science we demand clear quantitative relationships. We discuss familiar applications soon. Here we simply state that the temperature of a thermodynamic system in equilibrium is defined by the following relationship:

$$dS = \frac{dQ}{T} \tag{12}$$

Adding a slight amount of random energy to a system in the form of heat would raise its disorder/entropy slightly, and the temperature is the ratio of the two.[5] As entropy has meaning only for thermodynamic systems, so too temperature is a property special to systems characterized by random internal motion.

Equipartition Theorem and Average Energy

A characteristic of thermodynamics systems that can be determined fairly easily is the amount of energy to be found in certain internal forms: kinetic, potential, rotational, etc. Foremost is the equipartition theorem, which states that for each independent variable upon which a particle's energy depends quadratically—that is, on the *square* of the variable—that particle will on average have $\frac{1}{2}k_BT$ of energy. Such independent modes of internal energy storage are usually referred to as "degrees of freedom," and the equipartition theorem can thus be stated as follows:

$$\text{Equipartition theorem:} \qquad \overline{E}_{\text{ per particle}} = \frac{1}{2}k_BT \quad \text{per degree of freedom} \tag{13}$$

where the overscore bar indicates an average. Perhaps the theorem's most common application is to average translational kinetic energy in three dimensions, given by $\frac{1}{2}m\ v^2 = \frac{1}{2}m(v_x{}^2 + v_y{}^2 + v_z{}^2)$. This depends quadratically on three independent variables, so the equipartition theorem gives:

$$\overline{\mathrm{KE}}_{\text{ trans, per particle}} = \frac{3}{2}k_BT \tag{14}$$

Proving the equipartition theorem involves calculating the average value of a representative quadratic "piece" of energy (e.g., $\frac{1}{2}mv_y{}^2$) over all the microscopic ways referred to in equation (10), making use also of equation (12). This we leave to a higher level course.

[4] Unfortunately, the symbol W is a standard one for both work and number of ways, entirely different quantities.

[5] Strictly speaking, temperature is defined by $T \equiv (\partial S/\partial E)^{-1}$. A system's entropy is a function of its internal energy E, the number of particles N and its volume V, and the *partial* derivative with respect to the system's energy E keeps constant both N and V—allowing no energy exchange with the surroundings via work.

Kinetic Theory and the Ideal Gas

Although cataloging internal energies in general is problematic, there are systems simple enough that we can fairly easily quantify important properties related to the internal energy, still without knowing exactly what each particle is doing. One of the simplest systems imaginable is the so-called **ideal gas**, comprising identical particles so small that the volume they actually occupy is only a minute fraction of the volume of the container in which they are held, with the rare collision occurring elastically in so short a time that there is never significant potential energy stored in "bumpers." The particles bounce around randomly in mostly empty space. Of course this is idealized, but real gases often behave very much as simple calculations say this system should behave. The system merits study also because it is one of the best for showing how averages crop up in studies of thermodynamic systems.

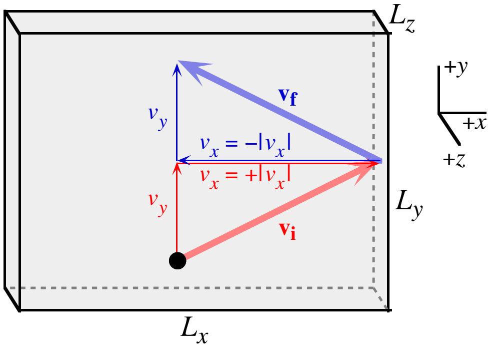


Figure 9

Important and familiar connections arise when we calculate the average force particles in an ideal gas exert on the walls of the container in which they are held. This we do by finding the equal-magnitude force the wall must exert on them to change their momentum, easily calculated via the general form of the second law of motion $\mathbf{F} = d\mathbf{p}/dt$. We assume that the box has sides of length L_x , L_y , and L_z , and we calculate the force at the "right-hand wall," which is parallel to the y-z plane, as shown in Figure 9. For now we consider only those particles whose x-component of velocity has *magnitude* $|v_x|$ (their x-components of velocity are either $+|v_x|$ or $-|v_x|$). There are N_{v_x} such particles. If we assume that they collide elastically and that the force is perpendicular to the wall, the sole result of a "reflection" off a wall parallel to the y-z plane is to reverse their *x-components* of velocity. To change a momentum component from $+m|v_x|$ to $-m|v_x|$ does indeed require a force, and Δp_x is $-\,2m|v_x|$. It is convenient to let the time interval be the time required for a particle of this $|v_x|$ to traverse the entire width L_x of the box:

$$\Delta t = \frac{L_x}{|v_x|} \tag{15}$$

In this time, all the particles of this $|v_x|$ would strike either the left or the right wall.[6] On average, half, those with $v_x = +|v_x|$, will strike the right wall, so this wall causes a total change in the momentum of these particular particles of

$$\Delta p_x = \frac{1}{2} N_{v_x} (-\,2m|v_x|) \tag{16}$$

Combining equations (15) and (16), the force exerted by the wall on only those particles whose x-component of velocity has magnitude $|v_x|$ is

$$\text{On particles of } |v_x|: \qquad F_x = \frac{\Delta p_x}{\Delta t} = -\frac{N_{v_x}}{L_x} m\, v_x^2$$

The volume V of the box is $L_xL_yL_z$ and the area A of the right-hand wall is L_yL_z, so we may write L_x as V/A.

$$\text{On particles of } |v_x|: \qquad F_x = -\frac{A}{V} N_{v_x}\, m\, v_x^2$$

[6] Of course any collision a particle suffered along the way would change its velocity. But collisions are rare, and in any case, if the system is indeed in equilibrium, N_{v_x} remains constant—as many particles must have their x-components changed *to* a given v_x as *from* that v_x.

To find the total force exerted on particles of all magnitudes $|v_x|$ we sum over all such values. For reasons soon to be clear, we also multiply and divide by the total number of particles $\Sigma_{v_x} N_{v_x} = N$.

$$\text{On all particles:} \qquad F_x = -\frac{NA}{V}\,\frac{\Sigma_{v_x} N_{v_x}\, m\, v_x^2}{\Sigma_{v_x} N_{v_x}}$$

The second fraction on the right-hand side of the equation is an average. Anytime we multiply a particular value of a given quantity (here, $m\, v_x^2$) by the number of times that value is counted (here, N_{v_x}), sum over all the allowed values, then divide by the total number of values, we have an average of the given quantity. Thus,

$$F_x = -\frac{NA}{V}\,\overline{mv_x^2}$$

Now, if the particles in this system do indeed move randomly, without a preferred direction of motion, then the average motion along one axis must be the same as the average along another. Accordingly, it must be true that $\overline{mv_x^2} = \overline{mv_y^2} = \overline{mv_z^2}$, and if this is true then $\overline{mv_x^2} = \frac{1}{3}\left(\overline{mv_x^2} + \overline{mv_y^2} + \overline{mv_z^2}\right)$. This allows us to put our result in a more general form.

$$F_x = -\frac{1}{3}\frac{NA}{V}\left(\overline{mv_x^2} + \overline{mv_y^2} + \overline{mv_z^2}\right) = -\frac{2}{3}\frac{NA}{V}\,\overline{\tfrac{1}{2}mv^2} = -\frac{2}{3}\frac{NA}{V}\,\overline{\text{KE}}_{\text{trans}}$$

Quite reasonably, the force the wall must exert on the particles is proportional to their average kinetic energy. Finally, the force these particles exert on the wall is equal and opposite, and this force divided by the area is the pressure on the wall.

$$P = \frac{2}{3}\frac{N}{V}\,\overline{\text{KE}}_{\text{trans}} \qquad \text{(Ideal Gas)} \qquad (17)$$

This is an important result, relating a tangible and easily measured macroscopic property, pressure, to an average microscopic property of particles in a system, kinetic energy.

If we combine equation (17) with equation (14), relating average kinetic energy and temperature, we have:

$$P = \frac{2}{3}\frac{N}{V}\,\frac{3}{2}k_B\,T \qquad \text{or} \qquad P\,V = N\,k_B\,T \qquad (18)$$

This is the famous **ideal gas law**. It is often one of the first things presented in an introductory class on thermodynamics, appearing as a “law of physics,” based on observation. But it follows quantitatively from ideal-gas assumptions combined, via the equipartition theorem, with more truly fundamental thermodynamic relationships such as equations (10) and (12).[7]

[7] The Boltzmann constant k_B, 1.38×10^{-23}J/K, appearing in so many thermodynamic relationships, is chosen so that in the ideal gas law pressure and volume are in standard units and temperature in the convenient but arbitrarily chosen units of Kelvin.

Adding Heat to a System

An important thing to know is how much heat must be added to a system to change its temperature or its phase (e.g., from solid to liquid). For very simple systems this can be determined fairly easily from the "first principles" already discussed. Suppose our system is an ideal gas in which the particles have only translational kinetic energy (which is the case for helium at most temperatures but not usually for polyatomic molecules). According to the equipartition theorem, the internal energy of N such particles is

$$E = \frac{3}{2} N\, k_B\, T$$

Consequently, if the internal energy is augmented by heat Q from the surroundings, the system's temperature will rise and will be related to the heat as follows:

$$Q = \Delta E = \frac{3}{2} N\, k_B\, \Delta T$$

The quantity $\frac{3}{2} N\, k_B$ is a proportionality factor determining the amount of heat needed to give a certain temperature rise. Suppose instead that we have particles capable of storing additional energy in rotation (such as diatomic molecules), with two additional "internal degrees of freedom." According to the equipartition theorem, each particle would have on average $2 \times \frac{1}{2} k_B T$ additional energy, a total of $\frac{5}{2} k_B T$, and thus

$$Q = \Delta E = \frac{5}{2} N\, k_B\, \Delta T$$

Now the proportionality factor giving the heat needed per unit temperature rise is $\frac{5}{2} N\, k_B$. We see that it takes more energy to raise the temperature of this gas than the previous one.

Calculating these proportionality factors becomes very complicated if not impossible for real materials. In fact, they are in general not constants, but tend to vary with temperature. Accordingly, quoted values are always the result of experimental observations in a restricted temperature range. The quantity often quoted is the **specific heat** c, defined as follows:

$$Q = m\, c\, \Delta T \qquad (19)$$

As we see from the earlier simple cases, the heat needed to increase the temperature is directly proportional to the amount of the material (i.e., $\propto N$). In equation (19), m is the mass of the material, proportional to the amount, leaving c dependent only on the *kind* of material. For the earlier strictly translational case, it is easily seen that $c = \frac{3}{2} k_B N/m$, or $\frac{3}{2} k_B/m_o$, where m_o is the mass *per particle*.

If we continue to add heat to or extract heat from a real material, we may well run into a **phase transition**—a solid will become a liquid, or a liquid a gas, or vice versa. Generally speaking, in a phase transition from a solid to a liquid or a liquid to a gas intermolecular bonds must be broken. This takes energy. (The same amount of heat must be extracted to turn liquid to solid as must be added to turn the same quantity from solid to liquid.) When the temperature characteristic of the transition is reached, additional energy goes not into increasing the average speed of the molecules—not into increased temperature—but into simply breaking more bonds. Thus, phase transitions occur at approximately constant temperature, and the amount of heat needed is proportional to the number of bonds that must be broken—that is, to the amount of the material. We express this as follows:

$$Q = m\, L \qquad (20)$$

Here m is the mass of the material whose phase is changed, and L, called the **latent heat**, is an energy per unit mass characteristic of the material and the transition, i.e., whether solid-liquid, or liquid-gas.

Electromagnetism

The more closely we investigate things, the closer we get to fundamental interactions. The frictional and normal forces at play when a book slides along a table are very complicated on the microscopic scale. The forces that an electric or magnetic field exerts on a charge or a fairly simple collection of charges, such as a dipole or an electric current, are comparatively simple, and it is these that interest us most in modern physics.

*   *   *   *

Maxwell's Equations

Maxwell's equations are the definitive statements about the character of electric and magnetic fields: what the fields can and cannot do and what things produce them. They express mathematically what our observations have revealed.

$$\oint \mathbf{E}\cdot\mathbf{dA} = \frac{1}{\epsilon_o}\int \rho \, dV \qquad (21)$$

$$\oint \mathbf{B}\cdot\mathbf{dA} = 0 \qquad (22)$$

$$\oint \mathbf{B}\cdot\mathbf{d}\boldsymbol{l} = \mu_o \int \mathbf{J}\cdot\mathbf{dA} + \epsilon_o\mu_o \frac{d}{dt}\int \mathbf{E}\cdot\mathbf{dA} \qquad (23)$$

$$\oint \mathbf{E}\cdot\mathbf{d}\boldsymbol{l} = -\frac{d}{dt}\int \mathbf{B}\cdot\mathbf{dA} \qquad (24)$$

Equation (21), known as Gauss' law, says that there will be a positive net **flux** of electric field lines, a net "flow" outward through a surface enclosing a volume, if the net charge enclosed in that volume is positive. The enclosed charge is the charge density ρ, charge per unit volume, summed over the entire enclosed volume. A positive flux is characterized by a net positive dot product between the electric field **E** and the outwardly-directed area vectors **dA** at the surface, as is the case on the spherical surface enclosing the positive point charge in Figure 10. Conversely, equation (21) says that there will be a negative/inward net flux for a negative net enclosed charge. Thus, *charge produces electric field*, directed out from positive charge and in toward negative. Equation (22) says that there simply cannot be a nonzero net flux of magnetic field lines through a surface enclosing a volume. Could we isolate a magnetic north pole, analogous to a positive charge, we could have a net outward flux—and an isolated magnetic south pole would give a net inward flux. But as yet there has been no verifiable observation of such **magnetic monopoles**. Then what produces magnetic fields? The answer is in equation (23), generally referred to as Ampere's law. The symbol **J** stands for current density, current (charge per unit time) per unit area. If there is a flow of current through an area—the first term on the right in equation (23)—there will be a net **circulation** of magnetic field lines around a path enclosing that area—the term on the left. In Figure 11, for example, a current flowing through a wire produces a magnetic field with a consistent circulation direction along a path circling the wire, so that the dot product between the magnetic field **B** and the vectors **d*l*** along the path is always of the same sign (negative in the figure, since **B** and **d*l*** are everywhere opposite). Thus, *current produces magnetic field*. Equation (23) says that there is another source of magnetic fields: an electric field changing with time. Similarly, the thrust of equation (24), Faraday's law of electromagnetic induction, is that another source of electric fields is a magnetic field changing with time. We will not investigate the latter two relationships, but simply note that they tie electric and magnetic fields together inextricably and explain the propagation of an electromagnetic wave through space, in the complete absence of charges and currents.

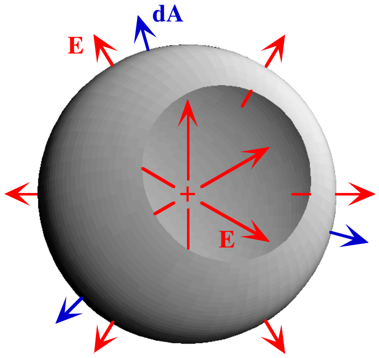


Figure 10

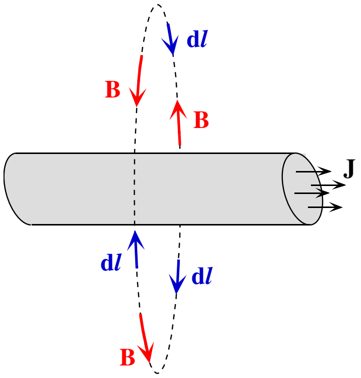


Figure 11

Basic Fields

A point positive charge q produces a rather simple electric field (see Fig. 10) given by $\mathbf{E} = \frac{1}{4\pi\epsilon_o}\frac{q}{r^2}\hat{\mathbf{r}}$, where r is the distance from the charge and $\hat{\mathbf{r}}$ is a unit vector directed away from the charge. Being always outward, it does give a positive net flux, as equation (21) demands.[8] The $1/r^2$ dependence is like that of the *gravitational* field around a spherical heavenly body such as Earth.

Another basic case is the **uniform field**, which is constant in both direction and magnitude. One way to produce a uniform electric field is via a very large plate on which charge is distributed uniformly. If the charge is positive, as in Figure 12a, the field points perpendicularly away from the plate (a positive/outward flux through a surface enclosing any portion of the plate); if negative, as in Figure 12b, the field points toward the plate. Often two such plates—one positive, one negative—are placed parallel to each other, as in Figure 12c. Between the plates, where the two fields are the same in both magnitude and direction, the resulting uniform net field points from the positive to the negative. Outside the plates, the individual fields being the same in magnitude but *opposite* in direction, the net field is *zero*! We refer to this configuration as a **parallel-plate capacitor**.

A basic magnetic field pattern is that due to a long, straight current-carrying wire, shown in Figure 11. The field's direction is tangent to circles centered on the wire, consistently clockwise or counterclockwise depending on the direction of the current.

Whereas a uniform electric field is found inside a parallel-plate capacitor, a uniform magnetic field is found inside a solenoid, Figure 13—current-carrying wire coiled tightly in a cylindrical form.[9]

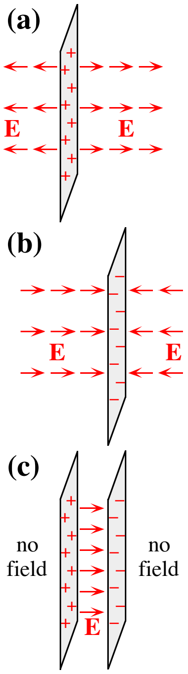


Figure 12

For all kinds of fields, if the region of interest is "small," there will be little variation of the field over the region, so it can often be treated as uniform. Earth's gravitational field, for example, is essentially uniform—downward at 9.8m/s^2—in a "small" region near its surface.

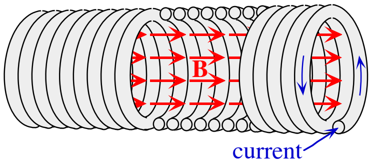


Figure 13

Fields and Forces

Among our primary interests are *effects* of electric and magnetic fields upon charges and currents. Charges and current *produce* fields, but they also respond to them. We often regard the field as the intermediary conveying a force between one charge or current, which produces the field, and another, which responds to it. Experimental observation has verified that the response may be expressed follows: A charge q_o in a region of space in which there is an electric field $\mathbf{E}$ and/or magnetic field $\mathbf{B}$ (produced by something other than the charge itself) will experience a force given by

$$\mathbf{F} = \mathbf{F}_E + \mathbf{F}_B = q_o\mathbf{E} + q_o\mathbf{v} \times \mathbf{B} \tag{25}$$

A stationary positive charge experiences an electric force $\mathbf{F}_E = q_o\mathbf{E}$ in the direction of $\mathbf{E}$. We measure charge is coulombs, C. Accordingly, common units for electric field are newtons per coulomb, N/C. The most familiar charged particles are the proton, of charge $+e$, where e is the **fundamental charge**, 1.6×10^{-19}C, and the electron, of charge $-e$. When q_o is negative, the force $\mathbf{F}_E = q_o\mathbf{E}$ is opposite the electric field $\mathbf{E}$. If a charge is moving it also experiences a magnetic force $\mathbf{F}_B = q_o\mathbf{v} \times \mathbf{B}$. We defer further discussion of magnetic fields and forces till later, concentrating now on electrostatics.

In the parallel-plate capacitor of Figure 12c, there is a field at all points in the region between the plates, produced by the charges on them. If *another* charge q_o were put somewhere in this region, there would be a force $\mathbf{F} = q_o\mathbf{E}$ on it—but the field $\mathbf{E}$ is there already. There is no force before q_o is introduced and it makes

[8] Because we are generally more interested in the effects of fields than in how they are produced, and because of the time that such derivations would entail, we will not go into the details of showing how a given field results from a given charge or current distribution and Maxwell's equations. We refer the reader to his or her comprehensive introductory text.

[9] The field outside is virtually zero, so any path enclosing wire has a net, if inconsistent, circulation of field lines around it—clockwise if enclosing wire along the bottom, for example.

no sense that there would be. Given a preexisting field, the force on a charge q_o put at a given point could be of any magnitude, depending on the magnitude of q_o , and either direction, depending on its sign. Indeed we often regard a “test charge” q_o as a way of probing a preexisting electric field. The charge q_o is introduced and experiences a force, the magnitude of the field is given by $E = F/|q_o|$, and its direction is the same as that of the force if q_o is positive and opposite if q_o is negative. The relationship between field and force, between a quantity preexisting at all points in space and different possible values of a second quantity when a charge is introduced, recurs in the relationship between two other important quantities, as we now see.

Potential and Potential Energy

Anytime a conservative force is at play there will be associated with it a potential energy, for which we use the symbol U. If an object experiences such a force and we have to push on it to move it somewhere, we are storing/increasing this potential energy. Lifting a weight, subject to the gravitational force, increases gravitational potential energy; compressing a spring increases elastic potential energy. Conversely, allowing the object to move in the other direction, the direction in which the conservative force acts, such as “holding back” a weight while lowering it, decreases the potential energy.[10]

Now consider the parallel-plate capacitor of Figure 12c. Moving a charge q_o in the region between the plates certainly involves a change in potential energy—but is it an increase or a decrease? We would have to push a positive charge to move it to the left, and would thus be increasing potential energy. But if a *negative* charge of the same magnitude were used, the force would at all points be opposite, we would have to hold back on it in allowing it to move to the left, and the potential energy would *decrease* by the same amount.[11] Clearly we cannot assign a value to the potential energy at any point in space—it depends on the charge placed there. But there is quantity whose value is well-defined without/before placing a charge somewhere: the **potential** V. Just as force can have vastly different values when a charge is placed at a point where there is an electric field, the potential energy can have different values when a charge is placed at a point where there is a potential.[12]

Force and field are related via $\mathbf{F} = q_o\mathbf{E}$. Potential energy and potential are analogously related:

$$U = q_o V \tag{26}$$

As field is measured in newtons per coulomb, potential is measured in joules per coulomb J/C, defined to be a volt V.[13] An obvious difference between the relationships is that force and field are vectors, while potential energy and potential are scalars.

Nevertheless, the analogies are very helpful in understanding potential. If we accept that a charge produces an electric field everywhere around it, which will exert a force on any other charge introduced, we should be willing to accept that it also produces a scalar property everywhere, potential, which determines the potential energy that results if another charge is introduced. Potential is very important for several reasons: First, useful things such as batteries and generators by nature establish a potential difference between one terminal and the other. Second, as noted previously, potential does have a well-defined value at all points in space, as potential energy does not. But these things are related! If region R is our region of interest, and charges external to region R are responsible for conditions in the region, these external charges do not establish *forces* in region R—the region may contain no charges at all! They establish a field at all points. Similarly, they do not establish potential energies; they establish a potential at all points. For example, a 12 *volt* battery establishes a 12V difference in potential between its terminals and nothing can be said of potential *energy* until we know something about what charges might be around to respond. Whenever we speak of fields, rather than forces, the related scalar quantity will naturally be potential, rather than potential energy.

[10] The best example of a *non*conservative force is friction. No potential energy is stored in pushing a book along a table. We don’t get back the energy we expend; it becomes “lost” thermal energy. If this thermal energy *could* somehow return to the book as kinetic energy—violating the second law of thermodynamics—we would have to hold books after sliding them along tables, to keep them from darting back where they were, meanwhile cooling the table. But this we are never required to do.

[11] Alternatively we could simply let it go. Its potential energy would decrease as its kinetic energy increases.

[12] One of the most irksome hurdles in learning electrostatics is the similarity of the terms “potential” and “potential energy.” It is made worse by the fact that even before studying electromagnetism students have probably already become accustomed to the habit of shortening the term “potential energy” to “potential.” In electrostatics we must be very careful never to do this.

[13] By unfortunate convention, the same symbol is used for the quantity and the units in which it is measured. $V = 12\text{V}$, for example, while looking like an incorrect equation, means that the potential is twelve volts.

This is borne out by two important relationships between the four quantities, which also make clear that potential is not something “new”; it is inextricably related to the field—where there is a field there will be a potential. Introductory mechanics tells us how to find potential energy from force:

$$U_2 - U_1 = -\int_1^2 \mathbf{F}\cdot\mathbf{d}\boldsymbol{l} \tag{27}$$

Suppose we divide this equation by q_o, the charge that might respond to the field but is unknown beforehand. Inside the integral we will have $\mathbf{F}/q_o$ which is the electric field $\mathbf{E}$. On the left of the equal sign we will have $U_2/q_o - U_1/q_o$. This “potential energy per unit charge” is related to field, force per unit charge, exactly as potential energy is to force. Indeed, it is just what was introduced in equation (6), $U/q_o = V$. Thus[14]

$$V_2 - V_1 = -\int_1^2 \mathbf{E}\cdot\mathbf{d}\boldsymbol{l} \tag{28}$$

If nothing has been said about charges that might be present to experience a force (i.e., q_o), the discussion must revolve around **E** and V, not **F** and U.

Equation (28) shows that potential may be calculated at arbitrary point 2 knowing its value at point 1 and the electric field. In common with potential energy, it is the *differences* in potential that are important and we are free to choose the zero value at any point we find convenient. Since V can be found from **E**, it might seem that potential is redundant. But just as there are times in introductory mechanics when it is easier to work with the scalar potential energy than the vector force, the are many occasions when it is easier to work with potential than with field.

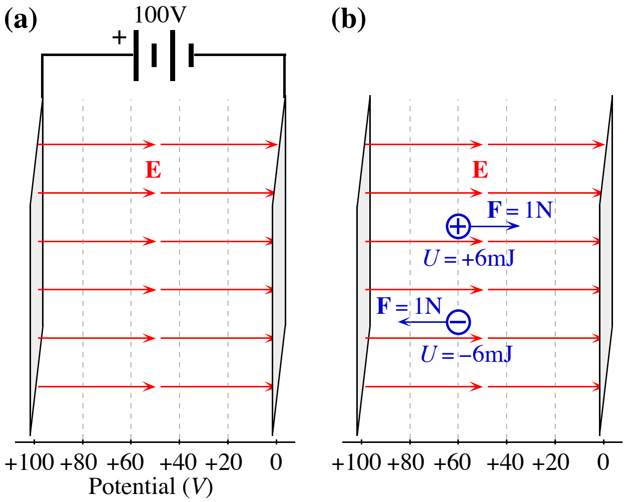


Figure 14

Figure 14 somewhat summarizes the relationships. In diagram (a) a battery establishes a potential difference, 100 volts higher on the positive/left plate of a parallel-plate capacitor than on the negative/right plate. We arbitrarily define the negative plate to be zero potential, so the left plate is +100V. The labeling of the potential shows a constant rate of change with x. This must be so because equation (28) says that $|dV| = |E|\,|dx|$ for a small displacement parallel to the electric field and $|E|$ is constant inside a parallel-plate capacitor. The magnitude $|E|$ of the field has been chosen as 10^4N/C.[15] Thus far, there are no forces and no potential energies.

In diagram (b) we add a charge q_o of *magnitude* 100μC. For a positive q_o, the force, $\mathbf{F} = q_o\mathbf{E}$, is to the right, magnitude $(10^{-4}\text{C})(10^4\text{N/C}) = 1\text{N}$, and the potential energy, $U = q_oV$, is positive at all points. The charge is placed where the potential is +60V, giving a potential energy of $(+10^{-4}\text{C})(60\text{J/C}) = +6\text{mJ}$. The potential decreases to the right, so the potential energy decreases to the right. When released, a charge should move in the direction of decreasing potential energy (increasing kinetic), and everything here seems to fit. For a negative q_o, the force is to the left and the potential energies are all negative. For the $q_o = -100\mu\text{C}$ shown in the diagram, the force is again of 1N magnitude and $U = q_oV = (-10^{-4}\text{C})(60\text{J/C}) = -6\text{mJ}$. The potential energy now decreases to the *left*, so we see that a negative charge too, though moving toward increased *potential*, is pushed in the direction of decreasing potential *energy*.

[14] In this equation we see the reason electric field is sometimes expressed in units other than newtons per coulomb, N/C. If a potential is a field times a distance then a field is a potential per unit distance, volts per meter, or V/m.

[15] Across the entire distance between the plates, we have $|\Delta V| = |E|\,|\Delta x|$. With $|\Delta V|$ already chosen to be 100V, in fixing $|E|$ at 10^4N/C, or 10^4V/m (see preceding footnote), we are fixing the plate separation at 1cm.

Accelerating a Charge Through a Potential Difference

One of the most basic applications of these ideas is simply accelerating a charge. If a potential difference ΔV is established between two plates, a charge q_o starting at one plate and accelerating to the other would experience a potential energy decrease of magnitude $|\Delta U| = |q_o|\,|\Delta V|$. It gains in kinetic energy what it loses in potential energy, so its kinetic energy gain would be the same magnitude.

$$\Delta \mathrm{KE} = |q_o|\,|\Delta V| \tag{29}$$

In practice, the plate toward which the charge accelerates usually has a hole at the proper place to allow the particle to exit (the hole being too small to affect the fields). Once outside, the field is zero, so the force on the particle is zero and it moves at constant speed.

It is helpful to plot field, potential and potential energy versus position for the situation depicted in Figure 14. In Figure 15, the field between the plates is constant and points in the positive direction (to the right), while outside the plates it is zero. As equation (28) demands, the plot of potential is the negative integral of the electric field plot (as the E-plot is the negative derivative of the V-plot) and so is linear between the plates. There are two potential energy plots because U depends on the q_o chosen. Both are plots of $U = q_oV$, but charges of opposite sign have been chosen, so one plot is the negative of the other. To accelerate a particle we must start it where the potential energy is high. Clearly this is the positive/left plate for a particle of positive charge, and the negative/right plate for a negative particle. Accelerating toward the other plate, the particle gains kinetic energy as it loses potential energy. After it exits, the force on it is zero, so its potential energy doesn't change. Note that a negative particle after exiting would have a constant negative potential energy. This value is set by our original assignment of zero potential at the right-hand plate; a particle that has been accelerated to a high kinetic energy must end up with a potential energy lower than when it started, lower than zero.

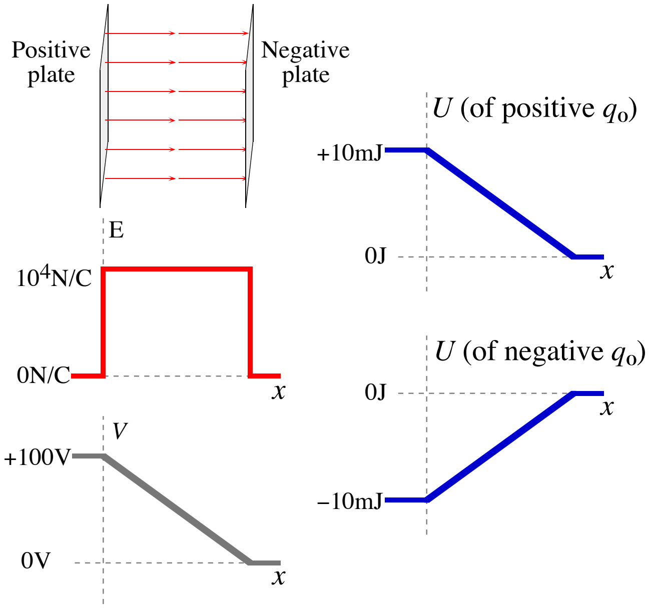


Figure 15

Before moving on, we note an important definition. According to equation (29), if a particle whose charge is of magnitude e, such as a proton or an electron, were accelerated through a 1V potential difference, it would gain a kinetic energy of $(1.6\times10^{-19}\text{C})(1\text{V}) = 1.6\times10^{-19}\text{J}$. This amount of energy we define to be one **electron volt**, symbol eV. Thus, $1.6\times10^{-19}\text{J} \equiv 1\text{eV}$. This is a very convenient definition; an electron accelerated through 200 volts acquires 200eV of kinetic energy as its potential energy decreases by 200eV. When dealing with subatomic particles whose charges are specified in multiples of e rather than in coulombs, electron volts is usually preferable to joules.

Magnetic Forces

Equation (25) gives the force a magnetic field exerts on a moving point charge: $\mathbf{F}_B = q_o\mathbf{v} \times \mathbf{B}$. It is always perpendicular to both the charge's velocity **v** and the magnetic field **B**, the direction given by the right-hand rule.[16]

Often our concern is the behavior of a steady stream of charges, that is, of a current. Current I is measured in coulombs per second, or amps A, and is understood by convention to be in the direction of positive charge flow. This is the idea of **conventional current**. If the moving charges are actually negatively charged, as they are in common conductors such as copper, the current's direction is by this convention opposite the charges' velocity. In determining forces on currents and in many other applications, it doesn't matter whether it is negative charges moving one way through a wire or positive charges moving the other—and it is easier to assume positive charge flow.

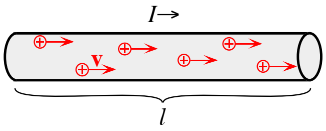


Figure 16

Figure 16 shows a current flowing in a segment of wire of length l. At speed v, all the moving[17] charge Q in the segment will pass through the end in time $t = l/v$. The current is thus given by $I = Q/t = vQ/l$, so that $Qv = Il$. For a straight segment, all this charge experiences a force in the same direction as does a single charge, so we may simply replace q_o by Q in the magnetic force equation. Qv then becomes Il, and if we define a vector $\boldsymbol{l}$ to have magnitude l and the same direction as the charge velocity **v** we arrive at:

$$\mathbf{F}_B = I\,\boldsymbol{l} \times \mathbf{B} \tag{30}$$

Dipoles

The physical world abounds in dipoles. A polar molecule, a neutral molecule with more positive on one side and more negative on the other, is an electric dipole; an electron orbiting in an atom, forming a loop of current, is a magnetic dipole. A dipole is characterized by a "dipole moment," a property that we define. But what is it and why do we define it as we do? Without the definition it is impossible to quantify important things that dipoles do. To give the most basic argument, in the presence of a uniform electric field, a polar molecule experiences no net force, because it is electrically neutral. But it may experience a torque; it may rotate. How can we quantify this if all we have is the net charge—which is zero?! Even knowing the amount of separated positive and negative charge isn't enough, for torque depends not just on force, but on the length of the moment arm—how *far* the charges are separated. At the very least, then, we need to specify—define—a property that gives some idea of how much charge is separated and by how far.

[16] Point the fingers of your right hand in the direction of the first vector, **v**, then orient your palm so that it is easiest to curl your fingers toward the direction of the second, **B**. Your thumb gives the direction of the cross product $\mathbf{v} \times \mathbf{B}$. If q_o is positive, then the force is in this direction; if negative, **F** is opposite $\mathbf{v} \times \mathbf{B}$.

[17] Usually wires are electrically neutral. Charges of one sign are stationary while an equal density of charges of the other sign carry the current.

Electric Dipole

Figure 17 shows an electric dipole, a charge of $+q_o$ separated from a charge of $-q_o$ by a distance a. (Here q_o is understood to be positive; the signs of the charges will be explicitly shown.) It is in a region permeated by a uniform electric field **E** (produced by charges elsewhere). The force on the $+q_o$, in the direction of **E**, and on the $-q_o$, opposite **E**, cancel, so the net force is zero. But there is a torque. From classical mechanics we know that torque is given by $\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}$, its magnitude $|r|\ |F| \sin\theta$.[18] Choosing our origin at the center about which the dipole rotates, we have two torques whose moments arms r are both $a/2$. Each has magnitude $|\tau| = |r|\ |F| \sin\theta = (a/2)\ \big(q_o\ |E|\big) \sin\theta$, and both are in the same direction, tending to rotate the dipole clockwise, so we may add their magnitudes, giving a total magnitude twice this value.

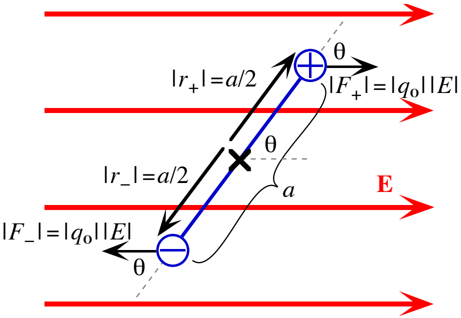


Figure 17

$$|\tau| = a\ q_o\ |E| \sin\theta \qquad (31)$$

We see, then, that to determine the torque we need to know (besides the field, of course) a, q_o and θ—properties of the dipole and its orientation relative to the field. But we must also be able to specify the torque's direction. In Figure 17, **r × F** is into the page. Instead of specifying all these things separately each time we wish to find torque, it would be nice to come up with a definition encompassing everything in a compact form. It is for this purpose that we define electric dipole moment:

$$\text{Electric Dipole Moment Vector:} \qquad \mathbf{p} \equiv \begin{cases} \text{magnitude:} & q_o\ a \\ \text{direction:} & \text{from } -q_o \text{ toward } +q_o \end{cases} \qquad (32)$$

Why is this convenient? The magnitude of the torque is now $|\tau| = |p|\ |E| \sin\theta$, which, since it is the same θ between **p** and **E** as between **r** and **F**, is the magnitude of **p × E**. Moreover, **p × E** has the same direction as **r × F**, so it serves to specify the torque vector in both its aspects.

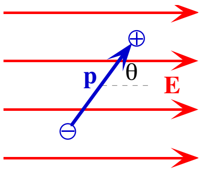


Figure 18

$$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E} \qquad (33)$$

This says that for a given field and angle, the larger the value of $|p|$, the larger will be the torque. It also says that the torque is zero when $\sin\theta$ is zero. In other words, the tendency to rotate ceases when **p** is aligned with **E**. This vector is defined wisely, for it allows us to express in an equation an important principle: *Dipoles tend to align with fields*!

[18] In this discussion of dipoles we will often use absolute value signs where, strictly speaking, they are not needed--with vectors. According to convention, boldface type is used for a vector (e.g., **v** for the velocity vector). If the symbol appears *not* in boldface (e.g., v) it is understood to mean the magnitude, or absolute value, of the vector. However, at certain points here it is so important to distinguish vectors from their magnitudes that we often include the redundant absolute value signs.

Magnetic Dipole

If we could concentrate magnetic north pole at one place and magnetic south pole at another, as positive and negative charges can be separated, the idea of a magnetic dipole would be *completely* analogous to an electric dipole. Following equation (32), the magnetic dipole moment would be an amount of “magnetic charge” times a separation. It might seem that a “bar magnet,” with its N on one end and S on the other, would fill the role. From the electrostatic analog, we should expect a magnetic dipole to align with a magnetic field, and we have all seen a bar magnet do this, when a compass aligns with Earth’s magnetic field. Also, though we haven’t discussed the electric field that an electric dipole itself produces, we should expect a magnetic dipole to produce a magnetic field of a similar “shape,” and indeed a bar magnet does produce such a field, known fittingly as a “dipole field.” But the bar magnet is deceptive. It has no point containing more N than S, or vice versa—magnetic monopoles have never been isolated! Yet, if we decide to steer clear of it (the “mystery” of the bar magnet is discussed afterward), what else fulfills our expectations?

They are fulfilled by a current loop. Figure 19 shows a rectangular loop of current in a region of magnetic field. The field exerts forces on each segment, given by equation (30). Those on the front-most and back-most segments are equal in magnitude, opposite in direction and they act along a line directly through the center of the loop, so they exert no torque. The forces on the top and bottom segments, though also equal and opposite (yielding zero net force), both act so as to rotate the loop clockwise. Due to the symmetry of the situation, both torques are of the same magnitude, $|\tau| = |r|\,|F| \sin\theta = (a/2)\,\big(I\,b\,|B|\big) \sin\theta$, so the net torque is twice this magnitude.

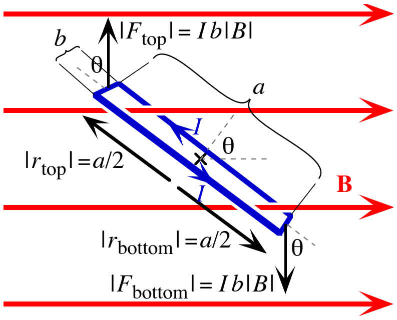

b
$|F_{\text{top}}| = I\,b|B|$
θ
a
I
θ
$|r_{\text{top}}| = a/2$
I
B
$|r_{\text{bottom}}| = a/2$
θ
$|F_{\text{bottom}}| = I\,b|B|$

Figure 19

$$|\tau| = I\,ab\,|B| \sin\theta \qquad (34)$$

With the proper definition we should again be able to write a compact expression for $\boldsymbol{\tau}$. Noting that the product *ab* is simply the loop’s area *A*, this definition is:

$$\text{Magnetic Dipole Moment Vector:}\quad \boldsymbol{\mu} \equiv \begin{cases} \text{magnitude: } I\,A \\ \text{direction: } \perp \text{ to plane of loop, by right-hand rule} \end{cases} \qquad (35)$$

Here the right-hand rule is to wrap the fingers of your right hand in the direction of the current, and the dipole moment vector is the direction of your thumb (see Fig. 20). Since the angle θ between $\boldsymbol{\mu}$ and $\mathbf{B}$ is the same as between $\mathbf{r}$ and $\mathbf{F}$, we see that the magnitude of $\boldsymbol{\mu} \times \mathbf{B}$, $(IA)\,|B| \sin\theta$, is the same as that of the torque in equation (34), and the direction is also the same as $\mathbf{r} \times \mathbf{F}$. Thus, our definition yields

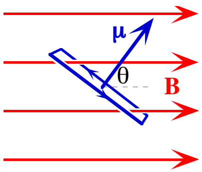


Figure 20

$$\boldsymbol{\tau} = \boldsymbol{\mu} \times \mathbf{B} \qquad (36)$$

The torque is zero, the dipole is aligned, when $\boldsymbol{\mu}$ is parallel to $\mathbf{B}$, which is when the plane of the loop is *perpendicular* to $\mathbf{B}$. In this sense *a current loop behaves just as would a bar magnet oriented perpendicular to the plane of the loop*, as depicted in Figure 21. (It should be noted that while we have assumed a rectangular loop, definition (35) gives the correct torque regardless of the loop’s shape.) Moreover, a current loop meets our other expectation for a dipole: It produces a magnetic field of the characteristic dipole “shape” (at sufficient distance), and it is oriented, again, just as would be the field of the bar magnet shown in the figure. The definitions of electric and magnetic dipole are not perfectly analogous—$|p|$ is a charge times a distance while $|\mu|$ is a charge per unit time times an area—but a current loop exhibits the important behaviors we expect of a dipole.

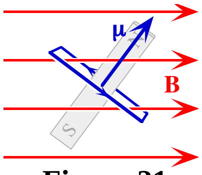


Figure 21

Finally we note that if we view the electric monopole (i.e., point charge) as the fundamental source of electric fields, then, there being no magnetic monopoles, we must view the magnetic dipole as the fundamental source of magnetic fields. A bar magnet, for instance, does not contain N and S monopoles; it is countless microscopic magnetic dipoles—whose directions are largely aligned—in the atoms spread throughout the magnet. This explains why breaking a bar magnet apart just makes two smaller ones.

Dipole Orientation Energy

If something has a tendency to rotate one way, then to rotate it the other way we must do work, store potential energy. One way to calculate the potential energy associated with the orientation of a dipole in a field is to integrate the torque. As work is $\int F\,dx$ for linear motion, it is $\int \tau\,d\theta$ for angular motion. Our torques are of magnitude $|p|\,|E| \sin\theta$ and $|\mu|\,|B| \sin\theta$, are the corresponding potential energies are:

$$U = -\,|p|\,|E| \cos\theta = -\,\mathbf{p}\cdot\mathbf{E} \qquad\qquad U = -\,|\mu|\,|B| \cos\theta = -\,\boldsymbol{\mu}\cdot\mathbf{B} \qquad (37)$$

As always, the location of zero potential energy is a matter of choice. By inspection, it is chosen so that U is zero when $\cos\theta$ is zero, when the dipole vector is perpendicular to the field. In either Figure 18 or 20, if θ were initially 90º, the dipoles would tend to rotate clockwise to align with the field, which means that being aligned with a field is a lower potential energy state. Equations (37) agree; when $\theta = 0$, the energy is negative. Conversely, we would have to do work to rotate either dipole to $\theta = 180º$, and would thus increase the potential energy. Again equations (37) agree; U is positive when $\cos\theta = -1$. A dipole opposite a field is a highest energy state. It is worth reiterating: *Dipoles tend to align with fields*.

Force on a Dipole

In Figures 17 and 19, the forces responsible for the nonzero net torque act, of course, at different points. Pairs of *forces* were equal and opposite—the net force was zero—because the *field* was assumed uniform, of the same magnitude and direction at all points. In a *non*uniform field, the forces might either be of different magnitudes or different directions or both, and the net force would not be zero. The upshot is that a dipole in a *nonuniform* field will in general experience a force as well as a torque. Perhaps the most common example is the force between two bar magnets. Each produces a nonuniform field and each experiences a net force in the presence of the field produced by the other. The force may be found via $\mathbf{F} = -\,\nabla U$, the gradient operator ∇ involving, as always, derivatives with respect to the coordinates x, y, and z.[19] From the potential energy formulas (37) we see that, since neither $|p|$ nor $|\mu|$ depend on location (they aren't functions of x, y, z), the force on a dipole at a given orientation angle can be nonzero only if the *field* varies with position.

Dipoles in the Real World

We have introduced electric and magnetic dipoles fairly rigidly: two point charges and a well-defined loop of current. In practice, things are often more complicated. A polar molecule is a charge *distribution*. There is no place to which we could point and say "There is the $+q_0$!" Similarly, an electron orbiting in an atom does not follow a well-defined loop. Nevertheless, these things have dipole moments. Indeed, we often "work backward," measuring the torque in a known field to find p or μ, which might then tell us something about the charge or current distribution.

[19] The relationship $\mathbf{F} = -\,\nabla U$ follows directly from equation (27), a general relationship between force and potential energy. Loosely speaking, if U is the negative integral of F, then F is the negative derivative of U. Still, simple though the expression may appear, the force can be fairly complicated to analyze, particularly in the magnetic case.

The Wave Equation—Waves on a String

In the final analysis, the one thing governing how a wave behaves is the **wave equation** that it obeys. Each kind of wave is different, and so is its wave equation, but all are differential equations, that is, equations involving derivatives of the function describing the wave. The "matter waves" we discuss in quantum mechanics obey the Schrödinger (wave) equation, and, though of course different in some ways, many of the potential hurdles in understanding matter waves can be removed by first gaining a good grasp of the simplest case of a wave equation and the waves that obey it—waves on a string. The topic is a good bridge between introductory mechanics and the more advanced applications of modern physics.

*   *   *   *

A string stretches between two fixed ends. We pluck it. What will it do? Naturally we expect some sort of disturbance to propagate along the string, but what is often overlooked is that it is simply obeying the fundamental (classical) law governing the movement of things—Newton's second law of motion. It is overlooked because the nature of the application dictates that we put the second law in an unusual form.

We assume that the string lies along the horizontal x-axis and that we are interested in transverse displacements, that is, displacements in the y-direction, as depicted in Figure 22a. Now we face a question: We can't analyze the entire string's motion all at once, since as a whole it doesn't go anywhere, so where does our analysis start? We consider a small segment, and apply the second law to it.

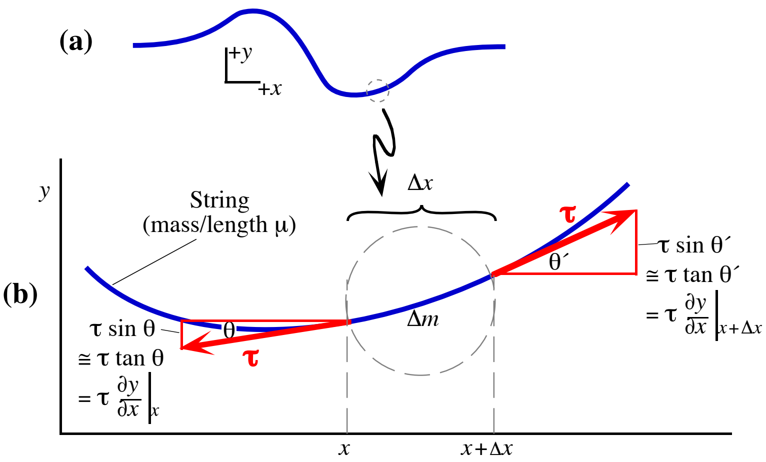

(a)
+y
+x
(b)
y
String
(mass/length μ)
Δx
τ
θ´
τ sin θ´
≅ τ tan θ´
$= \tau \left.\frac{\partial y}{\partial x}\right|_{x+\Delta x}$
Δm
τ sin θ
≅ τ tan θ
$= \tau \left.\frac{\partial y}{\partial x}\right|_{x}$
θ
τ
x
x+Δx

Figure 22

Let us focus on a segment of mass Δm and length Δx, extending from x to $x+\Delta x$, as shown in Figure 22b. Its acceleration is due to the net force acting on it. Ignoring gravity, there are two forces, exerted by the parts of the string *outside* the segment: one pulling left (more or less) and the other pulling right. Unless these are of equal magnitude and opposite direction, the segment will accelerate. Because things can get very complicated otherwise, we assume that the displacements are small—that the string, though deformed, is still nearly flat (unlike the exaggerated figure). This means that the angles θ and θ´ are small. We also assume that the tension τ in the string, while perhaps varying to some extent, is still effectively the same value everywhere.[20] Accordingly, the forces' x-components—nearly equal to τ itself, since the string is nearly flat—cancel, while their y-components vary quite a bit from one place to another. (For small α, cos $\alpha \cong 1$, while sin $\alpha \cong \alpha$, which can double, triple, become negative, etc.) Thus, there can be acceleration along the y-direction, and if the two forces' *magnitudes* are indeed equal it will be due solely to pulls at the segment's edges in slightly different *directions*. Figure 22b shows the forces' vertical components and how, for small angles θ and θ´, they are related to the tension τ and the string's slopes at the edges of the segment. Any acceleration is due to the varying slope along the string, and the slope is the rate of change—the derivative—of y with respect to x. This is a *partial* derivative with respect to x because the string's displacement y in general varies not only with x (i.e., along the length of the string), but also with time t—the displacement is really $y(x,t)$. We now have all we need to apply Newton's second law.

$$\text{Newton's Second Law: } m\,a_y = \Sigma F_y \quad \rightarrow \quad \Delta m\,a_y = \tau \left.\frac{\partial y(x,t)}{\partial x}\right|_{x+\Delta x} - \tau \left.\frac{\partial y(x,t)}{\partial x}\right|_{x}$$

[20] These assumptions are borne out well in many situations, but fail if the wave amplitude is too large.

We have considered a “small” segment, but where do we draw the line? If we allow Δx to be zero, then Δm would zero and the slopes would be equal—and $0 = 0$ doesn’t tell us much. The trick is to first divide by Δx, *then* let Δx approach zero.

$$\frac{\Delta m}{\Delta x}\, a_y = \tau\, \frac{\left.\frac{\partial y(x,t)}{\partial x}\right|_{x+\Delta x} - \left.\frac{\partial y(x,t)}{\partial x}\right|_{x}}{\Delta x}$$

The ratio $\Delta m/\Delta x$ is simply the mass per unit length μ, which is a constant and therefore finite as Δx goes to zero. But if the left side is now finite, so must be the right side. What is it? By the *definition* of the derivative, $\frac{dg(x)}{dx} \equiv \lim_{\Delta x\to 0} \frac{g(x+\Delta x) - g(x)}{\Delta x}$, it is the *second* derivative of $y(x,t)$ with respect to x, a rate of change of a rate of change (the function g representing the first derivative $\frac{\partial y}{\partial x}$ of the function y). Thus,

$$\mu\, a_y = \tau\, \frac{\partial^2 y(x,t)}{\partial x^2}$$

Finally, noting that the segment’s (vertical) acceleration is by definition the second derivative of its (vertical) displacement with respect to *time* (not x), we have the wave equation

$$\mu\, \frac{\partial^2 y(x,t)}{\partial t^2} = \tau\, \frac{\partial^2 y(x,t)}{\partial x^2} \qquad \text{Wave Equation} \qquad (38)$$

Clearly, this does not look much like $F = m\,a$, but we have shown that it follows directly from it. How is it more useful? Again, it is no use to consider the string as a whole; a wave gives a string curves, and the wave equation speaks of curvature—$\partial^2 y/\partial x^2$. Where this second derivative is positive, concave up, we should expect the string to be pulled upward (toward flat), with an upward acceleration, meaning that the other second derivative, $\partial^2 y/\partial t^2$, should also be positive. This is exactly what the wave equation says.

Solving differential equations like (38) is a topic in itself. We will be content to say that the displacement $y(x,t)$ must be a solution of this unusual statement of Newton’s second law, and one solution is *any* function whose argument is $x \pm \sqrt{\tau/\mu}\, t$, that is, any function of the form $y(x \pm \sqrt{\tau/\mu}\, t)$. To verify that it is a solution, we plug it in, but to streamline the task, we use the symbol A for the argument; that is, $A \equiv x \pm \sqrt{\tau/\mu}\, t$, so that $y(x,t) = y(x \pm \sqrt{\tau/\mu}\, t)$ becomes $y(A)$. Now, noting that $\frac{\partial A}{\partial t} = \pm\sqrt{\tau/\mu}$ and $\frac{\partial A}{\partial x} = 1$, and making liberal use of the chain rule,

$$\text{Left-hand side of (38): } \mu \frac{\partial^2 y(A)}{\partial t^2} = \mu \frac{\partial}{\partial t}\left(\frac{\partial y(A)}{\partial t}\right) = \mu \frac{\partial}{\partial t}\left(\frac{\partial y(A)}{\partial A}\, \frac{\partial A}{\partial t}\right) = \mu \frac{\partial}{\partial t}\left(\frac{\partial y(A)}{\partial A}\, \left(\pm\sqrt{\tau/\mu}\,\right)\right)$$

$$= \pm\sqrt{\mu\,\tau}\, \frac{\partial}{\partial t}\left(\frac{\partial y(A)}{\partial A}\right) = \pm\sqrt{\mu\,\tau}\, \frac{\partial^2 y(A)}{\partial^2 A}\, \frac{\partial A}{\partial t} = \pm\sqrt{\mu\,\tau}\, \frac{\partial^2 y(A)}{\partial^2 A}\, \left(\pm\sqrt{\tau/\mu}\,\right) = \tau\, \frac{\partial^2 y(A)}{\partial^2 A}$$

$$\text{Right-hand side of (38): } \tau \frac{\partial^2 y(A)}{\partial x^2} = \tau \frac{\partial}{\partial x}\left(\frac{\partial y(A)}{\partial x}\right) = \tau \frac{\partial}{\partial x}\left(\frac{\partial y(A)}{\partial A}\, \frac{\partial A}{\partial x}\right) = \tau \frac{\partial}{\partial x}\left(\frac{\partial y(A)}{\partial A}\, 1\right)$$

$$= \tau \frac{\partial}{\partial x}\left(\frac{\partial y(A)}{\partial A}\right) = \tau\, \frac{\partial^2 y(A)}{\partial^2 A}\, \frac{\partial A}{\partial x} = \tau\, \frac{\partial^2 y(A)}{\partial^2 A}\, 1 = \tau\, \frac{\partial^2 y(A)}{\partial^2 A}$$

In general, a function $y(x \pm \sqrt{\tau/\mu}\, t)$ may be thought of as a plot of $y(x)$ displaced to the right (– sign) or left (+ sign) by an amount $\sqrt{\tau/\mu}\, t$. As time t increases by one second, the function moves a distance of $\sqrt{\tau/\mu}$. Thus, the shape merely “slides” undeformed right or left along the string at the speed $\sqrt{\tau/\mu}$. We might have worried that requiring $y(x,t)$ to obey equation (38) would somehow conflict with our freedom to initially deform the string as we please, but we have shown that the equation holds no matter what the function y might be (sin, cos, sech), just so long as its argument is $x \pm \sqrt{\tau/\mu}\, t$. It merely requires that, to obey Newton’s second law of motion, the shape we choose propagate undeformed at constant speed $v = \sqrt{\tau/\mu}$.

Interference and Diffraction

The propagation of a wave can be very complicated. Invariably it is governed by an underlying wave equation, a differential equation that is different for each kind of wave (e.g., sound, light). However, though we will refer here to light, certain behaviors are common to all kinds of waves and many of these can be explained on the basis of one plausible and fairly simple assumption: that each point of a propagating wave *alone* acts as a source of wavelike disturbances spreading uniformly in all directions and the net result is the algebraic sum of all these disturbances, or **wavelets**. This we know as Huygens' principle.[21] Behaviors that it can explain and that are very important to understanding quantum mechanics are interference and diffraction.

* * * *

Single Point Source

Figure 23 shows light waves spreading uniformly in all directions from a single point source. At a given instant of time, different points in space "see" different parts of the cycle. At some points the oncoming wave is at its maximum, at others its minimum. But so long as there is just one source, the important result—what *we* would see—is very simple. The wave here is an oscillating electric field,[22] but we don't see electric fields; we see the average intensity of the oncoming wave, and if we average over time, the part of the cycle we see at a particular instant is unimportant.

Let us show this, by calculating the average intensity. We assume that at our chosen "observation point" the electric field is given by

$$E_0 \sin(\omega t - \phi) \tag{39}$$

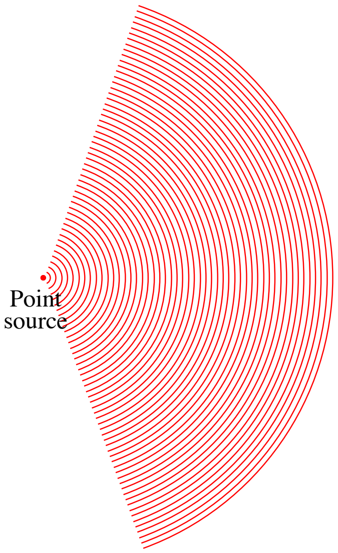


Figure 23

where E_0 is the amplitude of the oscillating field, ω is its angular frequency ($\omega = 2\pi f = 2\pi/T$, where T is the period), and ϕ is a phase factor determining what part of the cycle occurs at the observation point at the instant $t = 0$. Intensity is proportional to the square of the oscillating wave.[23] For our purposes, we will simply use k for the proportionality constant. Thus, we have

$$\mathrm{I}(t) = k\,E_0^2 \sin^2(\omega t - \phi) \tag{40}$$

Averaging over one period T then yields

$$\bar{I} = \frac{1}{T}\int_0^T k\,E_0^2 \sin^2(\omega t - \phi)\,dt \;= k\,E_0^2\,\frac{1}{T}\left(\frac{t}{2} - \frac{\sin(2\omega t - 2\phi)}{4\omega}\right)\Bigg|_0^T = \frac{1}{2}\,k\,E_0^2 \equiv I_{\text{1-source}} \tag{41}$$

Regardless of the value of ϕ, the sine function, being periodic, is the same value at $t = 0$ and $t = T$. Thus, it drops out when evaluating the antiderivative at its limits (remember: $T = 2\pi/\omega$), leaving the average intensity independent of time, as expected.

[21] We won't discuss the principle in detail. Still, it is certainly reasonable that each individual point of an oscillating wavefront, *were it the only thing oscillating*, would cause a disturbance everywhere else around it at some later time, that this disturbance would propagate outward in all directions at the wave's characteristic speed, and that the overall behavior would be the sum of behaviors caused by all these individual disturbances.

[22] Of course Maxwell's equations tell us that there is also a magnetic field part, but it propagates along in phase with the electric field part, so to understand the interference of one is to understand the interference of both.

[23] Intensity is power per unit area, or equivalently, energy density times speed, and energies simply tend to depend on squares of things—kinetic energy on the square of the speed, elastic (spring) potential energy on the square of the displacement, electromagnetic field energy on the square of the field, etc.

Two Point Sources—Double-slit Interference

If we now consider two point sources, we must add the two waves reaching our observation point. While it is true that the amplitude and thus intensity of waves naturally diminish as waves spread from a point, we need not be concerned with this, for we will consider only cases where all observation points are about the same distance from the sources. Of course they cannot be *exactly* the same distance from different sources, but they are close enough to ignore any variation in the amplitude. For instance, suppose we consider light of 500nm wavelength spreading from two point sources $\frac{1}{100}$ cm apart (a rather large separation as physical optics goes). As depicted in Figure 24, there is an observation point precisely 1 meter from both. Wavelets from the two sources would each have 2,000,000.0 wavelengths (1m/500nm) to travel to this point, would thus have precisely the same amplitude, and would interfere constructively. Another observation point is 0.999950m from one source and 1.000050m from the other. Traveling distances differing by only $\frac{1}{100}$%, the two wavelets will still have essentially equal amplitudes. However, these distances are 2,000,100.0 and 1,999,900.0 wavelengths, respectively. While these also give constructive interference—differing by an *integral* number of wavelengths—it is clear that at observation points in other directions the interference might be constructive, destructive or anything in between. In the cases we consider, it is not any variation in the amplitude of the interfering waves that causes the complexity of the overall intensity pattern; rather, it is the stark sensitivity, when *summing* multiple waves, to whether those waves are in phase or out of phase.

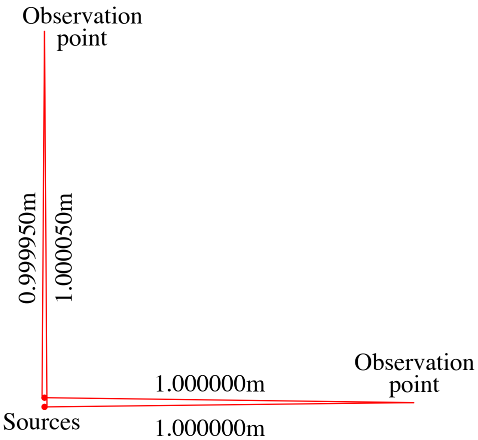


Figure 24

Let us determine the average intensity. As shown in Figure 25, the distance between the sources is *d*, and point C is halfway between the sources. The distance from point C to the observation point is *L*, and the observation point is at an angle θ from the perpendicular bisector to the line passing through the sources. We assume that *L* is so much larger than *d* that lines from the sources to the observation point are nearly parallel. Thus we see that the length l_2 of the line from source 2 to the observation point is longer than the length l_1 from source 1 by *d* sin θ, the so-called "path difference." The distances from the two sources to the observation point are therefore either longer or shorter than *L* by *one half* of *d* sin θ.

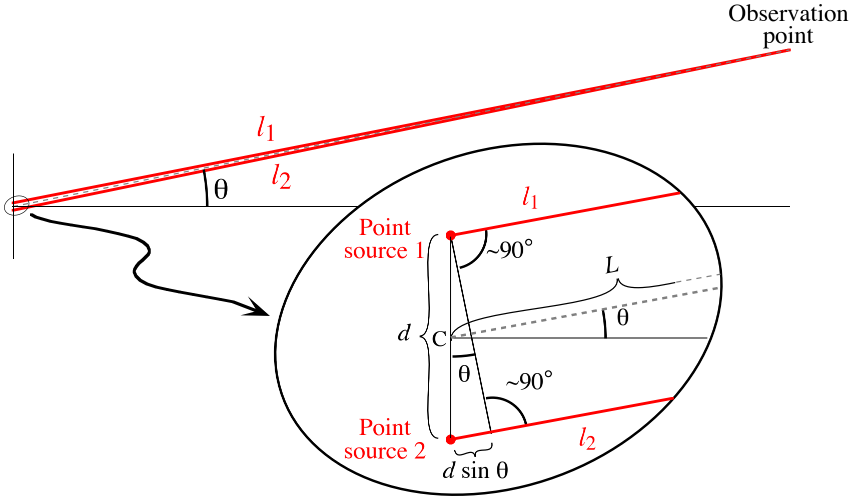


Figure 25

$$l_1 = L - \tfrac{1}{2}\, d \sin\theta \qquad\qquad l_2 = L + \tfrac{1}{2}\, d \sin\theta$$

It is best to set things up symmetrically about point C. Suppose that a source at point C (there isn't one!) would produce an electric field at the observation point given by $E_0 \sin(\omega t - \phi)$. A wave traveling a distance to the observation point *differing by Δl* would be displaced $\frac{\Delta l}{\lambda}$ cycles relative to this one, and at 2π radians per cycle, its electric field would be $E_0 \sin\!\left(\omega t - \phi - 2\pi \frac{\Delta l}{\lambda}\right)$. We have two such waves, with $\Delta l = \pm\frac{1}{2} d \sin\theta$, and the intensity, proportional to the square the *net* electric field, includes the contributions from both.

$$E_{\text{total}}(t) = E_0 \sin\!\left(\omega t - \phi + 2\pi \frac{{}^1\!/_2\, d \sin\theta}{\lambda}\right) + E_0 \sin\!\left(\omega t - \phi - 2\pi \frac{{}^1\!/_2\, d \sin\theta}{\lambda}\right) \tag{42}$$

Using the identity sin(a+b) + sin(a–b) = 2 sin a cos b, this becomes,

$$E_{\text{total}}(t) = 2\, E_0 \sin(\omega t - \phi) \cos\!\left(\frac{\pi}{\lambda} d \sin\theta\right)$$

Note that this is simply the one-source electric field (39) multiplied by the factor $2\cos\!\left(\frac{\pi}{\lambda} d \sin\theta\right)$. No matter what time t we consider, the net field and thus the intensity will be zero when this factor is zero. To find the average intensity, we square $E_{\text{total}}(t)$ then average over time; but since $2\cos\!\left(\frac{\pi}{\lambda} d \sin\theta\right)$ doesn't depend on t, its square comes out of the average-intensity integral, in the end simply multiplying the one-source intensity of equation (41).

$$\bar{I} = 4\cos^2\!\left(\frac{\pi\, d \sin\theta}{\lambda}\right) I_{\text{1-source}} \tag{43}$$

Cosine squared varies between zero and a maximum of unity, which occurs when its argument is any integral multiple of π. We see, then, that the intensity is maximum, four times the one-source intensity, when the following condition holds:

$$\text{Constructive interference:} \frac{\pi\, d \sin\theta}{\lambda} = m\,\pi \quad \text{or} \quad d \sin\theta = m\,\lambda \quad m = 0, 1, 2, \ldots \tag{44}$$

This makes sense. At angles θ where the path difference $d \sin\theta$ is an integral number of wavelengths, there should be constructive interference; the electric field waves from the two sources add, and an electric field twice as large means an intensity, proportional to the field's square, four times as large. At the other extreme are the angles θ where $\cos^2([\pi\, d \sin\theta]/\lambda)$ is zero—destructive interference.

Figure 26 plots intensity (43) versus angle θ along a screen. The value of m gives the **order**. The constructive interference corresponding to $m = 0$, where θ is zero, we refer to as the "zeroeth order maximum." At the angle θ corresponding to $m = 1$ we find the "first order maximum," and so on. Logically, the pattern repeats symmetrically on both sides of center. Moreover, the maxima move to angles farther from the center as d/λ decreases, which we see in the labeling of the θ-axis and from equation (44); to obtain a path difference of a given number of whole wavelengths with a smaller d and/or larger λ would require a more pronounced imbalance of path lengths—a larger angle.

The question might be asked: Why do we spend so much time studying two-source (or two-slit) interference? The answer is that interference effects involving waves are everywhere. Often they are due to essentially only two sources, but in any case, two-source interference is the simplest to analyze, and it is always best to start simple.

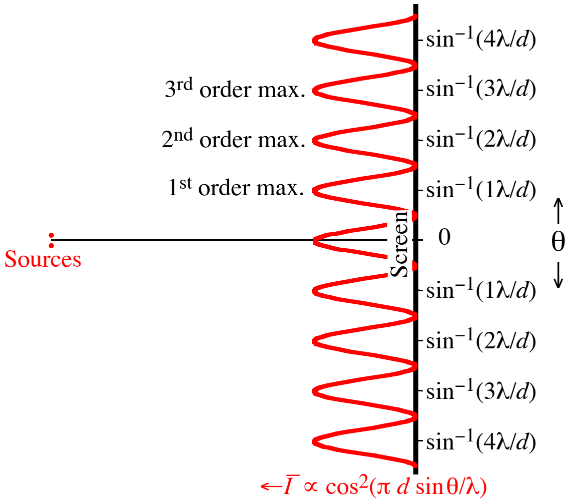


Figure 26

Finite-width Source—Single-slit Diffraction

It certainly sounds as though "the single slit" should be easier to analyze than "the double slit." Well, yes and no. If by single slit we meant a single *point* source, and by double slit two point sources, the single would be simpler. In fact, these are the two cases we have already considered, and the two-source was a more complex "problem." However, a single point source involves no interference at all and so is somewhat uninteresting. Accordingly, when in physical optics we refer to a "single slit," it is usually understood to mean a *finite-width* slit—not a single point source, but an infinite number of them spread over the finite width of the slit. As might be imagined, with an infinite number of sources, the interference is more complicated than for just one or two point sources, and it goes by a different name: **diffraction**. Usually, when we refer to the double slit we are referring to a case where the two sources may indeed be treated as point sources, and this is why "the double slit" is easier to analyze than "the single slit." The double slit *would* be more complicated than the single slit if by double slit we meant two finite-width sources. This case, less often referred to, at least in modern physics, we discuss briefly later.

Consider a barrier in which there is a single slit of width w, as shown in Figure 27. Applying Huygens' principle, we treat the wave passing through the slit as an infinite number of point sources, all in phase with each other, from which wavelets spread in all directions. Accepting that the interference of so many sources might become complicated, let us consider just two cases where we can argue easily whether it is constructive or destructive. First, to reach an observation point at $\theta = 0$, wavelets traveling from every point in the finite width of the slit would have the same distance to travel—they should interfere constructively.[24] Second, consider a point where the angle θ is such that $w \sin\theta$ is exactly one wavelength, as depicted in Figure 28. Were we to have simply *two* point sources at the *edges* of the slit, the one-wavelength path difference would lead to constructive interference. However, there are now intervening sources. In fact, the wavelet emanating from the point at the slit's exact center would travel a path to the observation point whose length differs from either of the previous two by *one-half* of a wavelength. Clearly it would destructively interfere in some way. The most common approach is to assert that the center wavelet cancels one of those from the edge; let us assume that it cancels the wavelet from the bottom edge. Now, the wavelet from the source a distance dy above the slit's center cancels that from the source a distance dy above the bottom edge, because these two wavelets must also have a path difference of $\frac{1}{2}\lambda$. Moving up dy each time, we may cancel wavelets, in pairs, from *every source in the slit*. In other words, at this angle, where $w \sin\theta = \lambda$, we should have completely destructive interference.

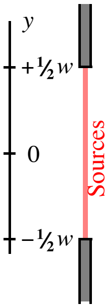


Figure 27

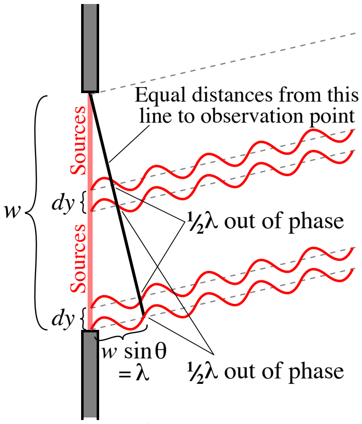


Figure 28

[24] They don't all have *exactly* the same distance; the distance from the slit's outside edge to the observation point at $\theta = 0$ is $\sqrt{L^2 + (w/2)^2}$, but in typical cases L (~1m) is so much larger than w (~1µm or less) that this distance is effectively just L.

It is possible to find other, larger angles at which the interference is completely destructive—such as having a $\frac{1}{2}\lambda$ path difference between sources $\frac{1}{4}w$, rather than $\frac{1}{2}w$, apart[25]—but there is a better way: Let us calculate the intensity. Before squaring to find intensity, we must add the electric fields reaching a given observation point from the infinite number of sources, which we assume stretch from $y = -\frac{w}{2}$, the slit's bottom, to $y = +\frac{w}{2}$, its top. Of course this is an integral. However, adding waves from an infinite number of sources, each of finite intensity, would give an infinite total intensity, so we must first scale down the electric field produced at the observation point from each source. Let us choose it so that the total when summed (integrated) over the wavelets from all the sources along the slit would be the same E_o as before.

$$dE = E_o \frac{dy}{w}$$

Ignoring any phase differences, integrating this from $y = -\frac{w}{2}$ to $y = +\frac{w}{2}$ would give E_o. Now to the path differences. In the two-point-sources case, the distance from the point *halfway between* the sources to the observation point was L and the distance from a *source* to the observation point was longer or shorter by $\frac{1}{2}d \sin\theta$. Observing that we now have a slit of width w rather than two sources separated by d, we may express the distance to the observation point from an arbitrary source as $L - y \sin\theta$. Note that this has the proper limits: L for the source at the very center of the slit ($y = 0$), longer by $\frac{w}{2}\sin\theta$ for the source at the slit's bottom, $y = -\frac{w}{2}$, and shorter by $\frac{w}{2}\sin\theta$ for the source at its top. Now we may write down, as we did in equation (42) for the two-source case, the total electric field at the observation point corresponding to the angle θ.

$$E_{\text{total}}(t) = \int_{-\frac{w}{2}}^{+\frac{w}{2}} E_o \frac{dy}{w} \sin\left(\omega t - \phi + \frac{2\pi}{\lambda} y \sin\theta\right)$$

Carrying out the integration,

$$E_{\text{total}}(t) = -\frac{E_o}{w} \left. \frac{\cos\left(\omega t - \phi + \frac{2\pi}{\lambda} y \sin\theta\right)}{\frac{2\pi \sin\theta}{\lambda}} \right|_{-\frac{w}{2}}^{+\frac{w}{2}}$$

$$= -\frac{E_o}{w} \frac{\cos\left(\omega t - \phi + \frac{2\pi}{\lambda}\frac{w}{2}\sin\theta\right) - \cos\left(\omega t - \phi - \frac{2\pi}{\lambda}\frac{w}{2}\sin\theta\right)}{\frac{2\pi \sin\theta}{\lambda}} \tag{45}$$

Using the identity $\cos(a+b) - \cos(a-b) = -2 \sin a \sin b$, this becomes

$$E_{\text{total}}(t) = \frac{E_o}{w} 2 \frac{\sin(\omega t - \phi)\sin\left(\frac{2\pi}{\lambda}\frac{w}{2}\sin\theta\right)}{\frac{2\pi \sin\theta}{\lambda}} = E_o \sin(\omega t - \phi) \frac{\sin\left(\frac{\pi\, w \sin\theta}{\lambda}\right)}{\frac{\pi\, w \sin\theta}{\lambda}} \tag{46}$$

[25] The approach doesn't work for *constructive* interference. It would seem that if the bottom-edge-of-slit source and center-of-slit source have a path difference of one *whole* wavelength, so would each other pair. True, each pair would be in a constructive interference condition. But one pair would be slightly out of phase with the next, so the interference would not be completely constructive for all sources. Only at the very center of the pattern is this the case.

Equation (46) is just the one-source electric field (39) multiplied by the factor $\sin\left(\frac{\pi\, w \sin\theta}{\lambda}\right) / \frac{\pi\, w \sin\theta}{\lambda}$. As in the two-source case, this factor does not depend on t, so when the average-intensity integral is carried through, its square will simply multiply the one-source intensity.

$$\bar{I} = \left(\frac{\sin\left(\frac{\pi\, w \sin\theta}{\lambda}\right)}{\frac{\pi\, w \sin\theta}{\lambda}} \right)^2 I_{\text{1-source}} \qquad (47)$$

At the center of the pattern, where $\theta = 0$, the fraction becomes zero over zero, but L'Hopital's rule shows it to be unity. In fact, the ratio $(\sin\alpha)/\alpha$ is never larger than unity, so we see that the maximum intensity occurs at $\theta = 0$; as we expected, all sources constructively interfere at the pattern's center. *Destructive* interference, zero intensity, occurs when the numerator $\sin\left(\frac{\pi\, w \sin\theta}{\lambda}\right)$ is zero but $\sin\theta$ itself (i.e., the denominator) is not zero, which occurs when $\frac{\pi\, w \sin\theta}{\lambda}$ is any *nonzero* multiple of π. Thus we arrive at the condition:

$$\text{Destructive interference:} \frac{\pi\, w \sin\theta}{\lambda} = m\,\pi \quad \text{or} \quad w \sin\theta = m\,\lambda \quad m = 1, 2, 3, \ldots \qquad (48)$$

This agrees with our earlier argument that destructive interference should occur where $w \sin\theta$ is exactly one wavelength.

Sadly, expressions (44) and (48) look the same—but they aren't! The interference of two points sources is entirely different from the interference of an infinite number of sources, so there is no reason one equation should not give points of constructive interference while a similar-looking equation gives points of destructive interference.

Intensity (47) is plotted in Figure 29. The angle corresponding to $m = 1$ is the "first order diffraction minimum," to $m = 2$ the "second order diffraction minimum," and so on. (Remember: $m = 0$ is a *maximum*.) From equation (48) and the labeling of the θ-axis in the figure we see that a smaller value of w/λ would cause the important features, the diffraction minima, to be found at ever larger angles θ—the pattern spreads out. In fact, in the special case $w = \lambda$, the first-order minimum would be found at 90°. In other words, it wouldn't be found at all! The sources are not far enough apart, in terms of wavelengths, to completely cancel one another; there would be nonzero intensity emanating at all angles within 90° of center. In the extreme case, where $w \ll \lambda$, the fraction in equation (47) is $(\sin\alpha)/\alpha$ with α approaching zero, which is unity *regardless of θ*. In other words, the intensity is *the same* in all directions, just as for a point source. We see then that *a single slit will behave as a point source so long as the slit width is much less than the wavelength.* At the other extreme, $w \gg \lambda$, the intensity is still necessarily maximum at $\theta = 0$, but at nonzero angles the numerator of equation (47) is bound between 0 and 1 while the denominator becomes very large. That is, the intensity goes to zero for any angle θ other than zero. The wave is not diffracted to the sides, but continues moving essentially in a straight line.

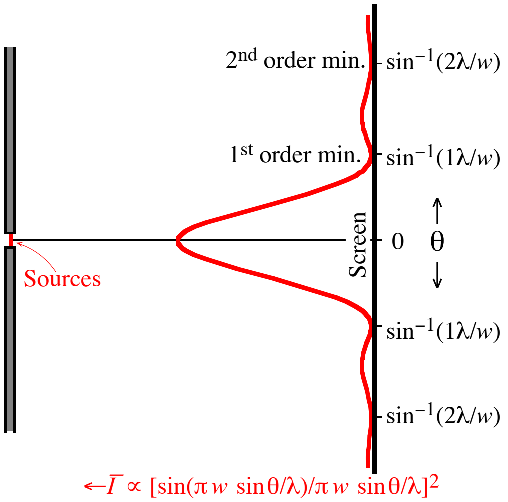


Figure 29

While 2-slit interference is important as the simplest case of multisource interference, the single slit is important because waves passing through apertures happens all the time (e.g., sound through doorways, light through pupils), and it is important to know whether the wave will continue in a straight line, as it will if $w \gg \lambda$, spread out in all directions, as it will if $w \ll \lambda$, or exhibit some interesting angle-dependent variation in its intensity, as it will if w is comparable to λ.

Two Finite Slits

When light passes through two finite-width slits, the result is a combination of single-slit diffraction and double-slit interference. Single-slit diffraction governs how much light would actually get to observation points at various angles from *each* slit, then "simple" two-source interference takes place between the light that does get there from the two. To best quantify it, we would integrate the electric field contributions as we did for a single finite-width slit, but over two separated regions (i.e., the slit openings). The resulting intensity is plotted in Figure 30. The angles at which the two-source maxima and the diffraction minima are found can be varied independently of each other because they depend on d/λ and w/λ, respectively, which are of course independent. For instance, keeping d/λ fixed but allowing w/λ to become very small, thus spreading out the diffraction pattern only, would recover the two-point-source pattern of Figure 26.

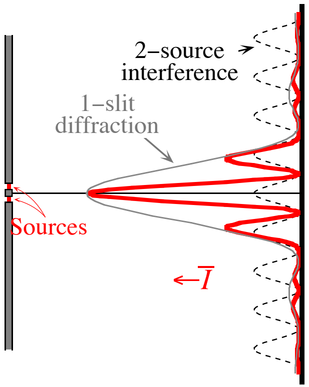


Figure 30

Many Sources

In conclusion it is worth mentioning that, roughly speaking, whenever there are *many* sources interfering, it is hard to obtain constructive interference and easy to obtain destructive. If sources are regularly spaced, and waves from adjacent ones are out of phase by *exactly* an integral number of wavelengths, then *all* will constructively interfere. However, if waves from adjacent sources are out of phase by even a small fraction of a wavelength, then the wave from one source and another one dozens, hundreds, even thousands of sources "down the line" will be a half wavelength out of phase, as will the two sources next to these, and so on, leading to essentially complete destructive interference. This is why "diffraction gratings," often employing thousands of narrow slits, produce almost complete darkness between the tiny, isolated bright spots. It is why diffraction from a crystal (i.e., from many, regularly-spaced atoms) similarly produces bright, well-defined spots. And it is why a wide single slit, with many sources, produces a narrow, straight-line beam surrounded by darkness, while a narrow slit, with "only a few" sources, allows the wave to spread in all directions.