Physics Notes

Jeremy Kelly
www.anthemion.org
October 31, 2017

Contents

1 Newton's laws of motion

1.1 Linear motion

A **trajectory** is a path along which some object moves. Motion within a trajectory is called **translational motion**.

Speed is a scalar quantity equal to the magnitude of the velocity. Positive acceleration entails increasing speed only if the velocity is currently positive in direction. A **turning point** is a place where some quantity, such as velocity, changes sign.

In a single dimension, vector quantities like $\vec{v}$ can be represented with scalars, like v. When this is done, the position after a displacement at velocity v:

$$x_1 = x_0 + \int_{t_0}^{t_1} v \, \mathrm{d}t$$

In **uniform motion**, v is constant, so that:

$$\begin{aligned} x_1 &= x_0 + vt\big|_{t_0}^{t_1} \\ &= x_0 + v(t_1 - t_0) \\ &= x_0 + v\Delta t \end{aligned}$$

In general, the delta symbol is understood to reference the time spanned by an event, so that $\Delta t = t_1 - t_0$ and $\Delta x = x_1 - x_0$.

Given a graph of position over time, the slope of the line *connecting* two points gives the **average velocity** between them, equal to $v = \Delta x/\Delta t$. The slope of the *tangent* to any point gives the **instantaneous velocity**:

$$v \equiv \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{\mathrm{d}x}{\mathrm{d}t}$$

Given a graph of velocity over time, the slope of the line connecting two points gives the **average acceleration** between them, equal to $a = \Delta v/\Delta t$. The slope of the tangent to any point gives the **instantaneous acceleration**:

$$a \equiv \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}^2x}{\mathrm{d}t^2}$$

The velocity after a period of acceleration:

$$v_1 = v_0 + \int_{t_0}^{t_1} a \, \mathrm{d}t$$

Assuming *uniformly accelerated* motion:

$$v_1 = v_0 + a\Delta t$$

so that:

$$\begin{aligned} x_1 &= x_0 + \int_{t_0}^{t_1} v \, \mathrm{d}t \\ &= x_0 + \int_{t_0}^{t_1} v_0 + a(t - t_0) \, \mathrm{d}t \end{aligned}$$

Notice that the velocity function is $v_0 + a(t - t_0)$ rather than $v_0 + at$. By definition, the time is t_0 when the velocity is v_0, so any change relative to v_0 must relate to a change *relative to the time at the same point.* This produces:

$$x_1 = x_0 + \left[v_0 t + \frac{1}{2}at^2 - at_0 t\right]_{t_0}^{t_1}$$

$$= x_0 + v_0\Delta t + \frac{1}{2}at_1^2 - \frac{1}{2}at_0^2 - at_0t_1 + at_0^2$$
$$= x_0 + v_0\Delta t + \frac{1}{2}a\Big(t_1^2 - 2t_0t_1 + t_0^2\Big)$$

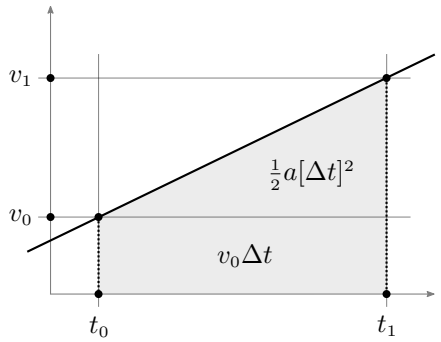
Finally, since:

$$(t_1 - t_0)^2 = t_1^2 - 2t_0t_1 + t_0^2$$

the displacement after a period of uniform acceleration:

$$x_1 = x_0 + v_0\Delta t + \frac{1}{2}a[\Delta t]^2$$

This is confirmed geometrically:



Because $\Delta t = (v_1 - v_0)/a$:

$$x_1 = x_0 + v_0\left(\frac{v_1 - v_0}{a}\right) + \frac{1}{2}a\left(\frac{v_1 - v_0}{a}\right)^2$$
$$= x_0 + \left(\frac{2v_0v_1}{2a} - \frac{2v_0^2}{2a}\right) + \left(\frac{v_1^2}{2a} - \frac{2v_0v_1}{2a} + \frac{v_0^2}{2a}\right)$$
$$= x_0 + \frac{v_1^2 - v_0^2}{2a}$$

This allows the velocity change to be expressed relative to the displacement, rather than the time elapsed:

$$v_1^2 = v_0^2 + 2a\Delta x$$

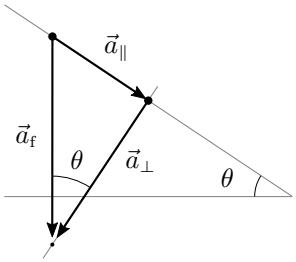
1.2 Motion on an inclined plane

The **acceleration due to gravity** g is always positive; it represents the magnitude of the true acceleration, not the direction. g varies at different points on the earth's surface, but the standard value is approximately 9.81m/s^2.

An object is in **free fall** when gravity is the only force acting upon it; such an object accelerates at rate $\vec{a}_\text{f}$, and this vector always points down. If the object is sliding down a frictionless inclined plane, $\vec{a}_\text{f}$ can be decomposed into two vectors, $\vec{a}_\parallel$ that is parallel to the plane, and $\vec{a}_\perp$ that is perpendicular to it:

$$\vec{a}_\text{f} = \vec{a}_\parallel + \vec{a}_\perp$$

If the angle between the plane and the planet's surface is θ, then the angle between $\vec{a}_\text{f}$ and $\vec{a}_\perp$ is also θ:



Therefore, the magnitude of the parallel acceleration:

$$a_\parallel = g\sin\theta$$

This is the component that accelerates the object, since $\vec{a}_\perp$ is opposed by the plane.

1.3 Force

Weight $\vec{w}$ describes the force exerted on an object by gravity, with $w = mg$.

When one object *pulls* another, it exerts **tension force** $\vec{T}$. The tension force exerted by a cable has the same direction as the cable itself. When one object *presses* against another, the second object exerts a **normal force** $\vec{n}$ against the first that is perpendicular to its own surface. Tension and normal forces are the result of molecular bonds, which behave like springs with pulled or pressed.

Inertia describes the innate tendency of objects to resist changes in their velocity. The **inertial mass** of an object:

$$m = \frac{F}{a}$$

A **superposition of forces** is a summation of multiple forces to form a *net* or *resultant* force. **Newton's first law**, known as the 'law of inertia', holds that the velocity of an object will remain constant if and only if the *net* force acting upon it is zero. **Newton's second law** holds that an object with mass m, subjected to *net* force $\vec{F}$, will experience acceleration:

$$\vec{a} = \frac{\vec{F}}{m}$$

When net force remains constant, acceleration is constant as well.

The SI unit of force is the **newton**:

$$N = \mathrm{kg} \cdot \mathrm{m/s^2}$$

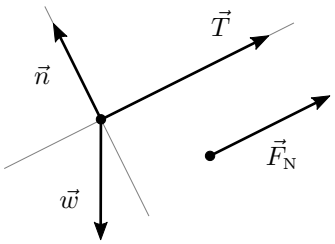
The **pound** is a force unit in the English system, approximately equal to 4.45N.

An object that is pushed or pulled to produce uniform acceleration a in the vertical axis is subject to two forces: the normal or tension force $\vec{F}_\mathrm{n}$ that creates the sensation of weight, and the actual weight, $w = mg$. Since $F_\mathrm{n} - mg$ must equal ma, the **apparent weight**:

$$\begin{aligned} F_\mathrm{n} &= m(g + a) \\ &= w\left(1 + \frac{a}{g}\right) \end{aligned}$$

For an object in free fall, $a = -g$, and the apparent weight is zero. The *actual* weight is still mg.

A **free-body diagram** places some object at the origin of a coordinate system, and shows all the forces acting upon it, along with the net force vector:



An object is in **mechanical equilibrium** when the net force acting upon it – and thus its acceleration – is zero. When at rest, such an object is in **static equilibrium**, and when in motion, it is in **dynamic equilibrium**, though either can be observed by selecting an appropriate reference frame.

1.4 Resistive forces

Static friction $\vec{f}_\mathrm{s}$ acts on objects that *rest* on some surface; its direction is opposite the surface-relative motion that would result if there were no static friction. **Kinetic friction** $\vec{f}_\mathrm{k}$ acts on objects that *slide* on some surface; its direction is opposite the velocity of the motion. **Drag** $\vec{D}$ acts on objects that move through fluids; its direction is also opposite the velocity of the motion. **Resistive forces** are those (like friction and drag) that *always oppose* the direction of motion.

For the most part, static friction is caused by molecular bonding between surfaces, although, due to their roughness, only 0.01% of these areas may actually touch. Kinetic friction is produced by weaker attractive forces between molecules.

The static friction force has no fixed magnitude. An object held in place by static friction exerts a force $\vec{f}_\mathrm{s}$ that is equal to and opposite the motive force acting upon it. There is a maximum force beyond which static friction fails to operate; in the simplest model:

$$f_\mathrm{max:s} \approx \mu_\mathrm{s} n$$

where μ_s is the **coefficient of static friction**, and n the magnitude of the normal force exerted by the surface. In practice, the condition of the surface also affects this calculation. The angle at which a resting object will slip from an inclined surface is called the **angle of repose**. This angle is a function of static friction.

Objects that slide are subject to kinetic friction force $\vec{f}_\mathrm{k}$ that opposes the direction of motion. Experimentally, f_k is nearly constant, and is less than $f_\mathrm{max:s}$. More generally:

$$f_\mathrm{k} \approx \mu_\mathrm{k} n$$

where μ_k is the **coefficient of kinetic friction**. In practice, surface area and speed can also contribute.

Rolling motion is opposed by **rolling friction** $\vec{f}_\mathrm{r}$. This acts like kinetic friction, with:

$$f_\mathrm{r} \approx \mu_\mathrm{r} n$$

The **coefficient of rolling friction** μ_r is generally much less than μ_k.

Drag is too complex to be easily generalized, but when an object moves fast enough to produce turbulence behind it:

$$D \approx \frac{1}{2} C_\mathrm{d}\, \rho A v^2$$

with ρ being the density of the fluid, A the object's cross-sectional area, and C_d the **drag coefficient**, which is often between 0.1 and 1.5.

An object falling straight down will accelerate until $D = w$. Equating $\frac{1}{2} C_\mathrm{d}\, \rho A v^2$ with mg gives the object's **terminal speed**:

$$v_\mathrm{t} \approx \sqrt{\frac{2mg}{C_\mathrm{d}\, \rho A}}$$

At lower speeds, the relationship between drag and speed is approximately linear.

1.5 Planar motion

A **position vector** is drawn from the origin to some position in space:

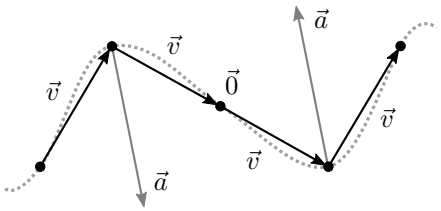
$$\vec{r} = x\hat{\imath} + y\hat{\jmath}$$

A change in position is shown by the **displacement**, which is drawn from the start position to the end position:

$$\Delta\vec{r} = \vec{r_1} - \vec{r_0} = \Delta x\hat{\imath} + \Delta y\hat{\jmath}$$

A displacement is *not* a distance; it is a vector quantity, like velocity or acceleration, and its direction is significant.

A **motion diagram** shows the position of an object at equally spaced points in time, with velocity vectors connecting each point with the next. Because they cover *intervals* of time, the velocity vectors are necessarily averages. Where possible, the acceleration vector representing the difference between $\vec{v}_n$ and $\vec{v}_{n+1}$ is shown to emanate from the point joining $\vec{v}_n$ and $\vec{v}_{n+1}$. Points showing no acceleration are labeled with the **zero vector** $\vec{0}$:



The average velocity $\Delta\vec{r}/\Delta t$ necessarily points in the same direction as $\Delta\vec{r}$. Similarly, the average acceleration $\Delta\vec{v}/\Delta t$ points in the same direction as $\Delta\vec{v}$.

The instantaneous velocity:

$$\begin{aligned}\vec{v} &\equiv \lim_{\Delta t\to 0}\frac{\Delta\vec{r}}{\Delta t} = \frac{\mathrm{d}\vec{r}}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t}\hat{\imath} + \frac{\mathrm{d}y}{\mathrm{d}t}\hat{\jmath} \\ &= v_x\hat{\imath} + v_y\hat{\jmath}\end{aligned}$$

As $\Delta t \to 0$, $\Delta\vec{r}$ becomes tangent with the trajectory. If $\vec{v}$ has angle θ relative to the positive x-axis, then $v_x = v\cos\theta$, $v_y = v\sin\theta$, and $\theta = \arctan(v_y/v_x)$.

Similarly:

$$\begin{aligned}\vec{a} &\equiv \lim_{\Delta t\to 0}\frac{\Delta\vec{v}}{\Delta t} = \frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = \frac{\mathrm{d}v_x}{\mathrm{d}t}\hat{\imath} + \frac{\mathrm{d}v_y}{\mathrm{d}t}\hat{\jmath} \\ &= a_x\hat{\imath} + a_y\hat{\jmath}\end{aligned}$$

Acceleration can also be decomposed such that $\vec{a} = \vec{a}_\parallel + \vec{a}_\perp$, with component $\vec{a}_\parallel$ parallel to $\vec{v}$, and $\vec{a}_\perp$ perpendicular to it; $\vec{a}_\parallel$ then gives the change in speed, and $\vec{a}_\perp$ the change in direction. This causes the coordinate system to change as $\vec{v}$ changes direction.

Planar motion can be modeled by decomposing $\vec{a}$ into $\vec{a}_x$ and $\vec{a}_y$, and then applying the linear motion model in both dimensions.

1.6 Projectile motion

When an object moves in the horizontal and vertical axes while subject to no force but gravity, **projectile motion** is produced. Any object with constant, non-zero acceleration along one axis and none along the other will follow a parabolic trajectory.

Given projectile motion with initial velocity $\vec{v}_0$ and launch angle θ:

$$\begin{aligned}v_{x:0} &= v_0\cos\theta \\ v_{y:0} &= v_0\sin\theta\end{aligned}$$

The only acceleration $a_y = -g$. Therefore:

$$\begin{aligned}\Delta x &= (v_0\cos\theta)\Delta t \\ \Delta y &= (v_0\sin\theta)\Delta t - \frac{1}{2}g[\Delta t]^2\end{aligned}$$

If the projectile lands at the same height it was launched, $\Delta y = 0$, and:

$$0 = \Delta t(v_0\sin\theta - \frac{1}{2}g\Delta t)$$

This equation has two roots, with the zero root representing the launch time displacement, and the other that of the landing time. Solving for the second of these:

$$\Delta t = \frac{2v_0}{g}\sin\theta$$

Multiplying by $v_{x:0} = v_0\cos\theta$ gives the horizontal displacement. Because $2\sin\theta\cos\theta = \sin 2\theta$:

$$\Delta x = \frac{v_0^2}{g}\sin 2\theta$$

The distance is maximized when θ is 45°. Because $\sin(180^\circ - 2\theta) = \sin 2\theta$, launch angles of θ and $(90^\circ - \theta)$ produce the same distances for $0 \le \theta \le 90^\circ$.

The trajectory is found by calculating the component positions with respect to time, and then substituting one solution into the other to eliminate the time variable:

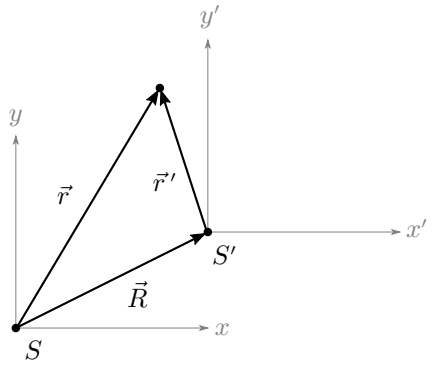
$$\Delta y = (\tan\theta)\Delta x - \Big(\frac{g}{2v_0^2\cos^2\theta}\Big)[\Delta x]^2$$

1.7 Relative motion

An **inertial reference frame** is a coordinate system within which Newton's first and second laws hold. A reference frame defined relative to a point that is accelerating is not inertial, as free objects will accelerate spontaneously in the opposite direction. In this sense, the earth is *not* a true inertial frame, since it accelerates around the sun.

If inertial frames S and S' have comparable coordinate systems, and if $\vec{R}$ is the position of the second origin relative to the first, displacement vectors referencing the same point are related by:

$$\vec{r} = \vec{r}\,' + \vec{R}$$



If S' moves relative to S at velocity $\vec{V}$, and if they meet when $t = 0$, then $\vec{R} = \vec{V}t$, and:

$$\vec{r} = \vec{r}\,' + \vec{V}t$$

This is known as the **Galilean transformation of position**. It follows that:

$$x = x' + V_x t$$
$$y = y' + V_y t$$

Since the horizontal velocity component in projectile motion is constant, projectile motion can be understood as free fall motion viewed from a different reference frame, and vice versa.

If the point referenced by $\vec{r}$ and $\vec{r}\,'$ also moves:

$$\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} = \frac{\mathrm{d}\vec{r}\,'}{\mathrm{d}t} + \frac{\mathrm{d}\vec{R}}{\mathrm{d}t}$$

which gives the **Galilean transformation of velocity**:

$$\vec{v} = \vec{v}\,' + \vec{V}$$

Similarly:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = \frac{\mathrm{d}\vec{v}\,'}{\mathrm{d}t} + \frac{\mathrm{d}\vec{V}}{\mathrm{d}t}$$

However, the relative acceleration of inertial reference frames is defined to be zero:

$$\vec{A} = \frac{\mathrm{d}\vec{V}}{\mathrm{d}t} = 0$$

This gives the **Galilean transformation of acceleration**:

$$\vec{a} = \vec{a}\,'$$

Neither the mass of an object nor the force exerted upon it change when observed from different frames. Thus the **Galilean principle of relativity**, which states that Newton's laws, applying to phenomena viewed from one inertial reference frame, still hold when the same phenomena are viewed from any such frame. By contrast, the speed of a given ray of light is *identical* in all reference frames, no matter what their relative velocity. This produces the **principles of special relativity**.

1.8 Uniform circular motion

Constant-speed motion in a circular path is called **uniform circular motion**. The **period** T of a circular motion is the time required to complete one revolution. Given radius r, dividing the circumference by the period produces the speed of the motion:

$$v = \frac{2\pi r}{T}$$

The **angular position** θ of some point is the angle between the positive x-axis and the line segment connecting that point with the origin. The difference between two such angles is the **angular displacement**.

In a circle with radius r, each radian of angular displacement spans an arc of length r. More generally, when θ is measured in radians, the **arc length**:

$$s = \theta r$$

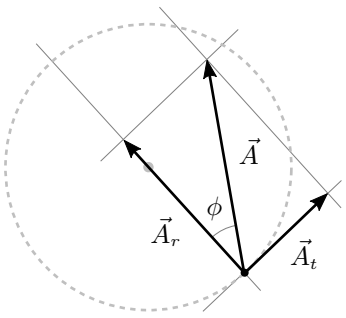
The **angular velocity**:

$$\omega \equiv \lim_{\Delta t \to 0} \frac{\Delta\theta}{\Delta t} = \frac{\mathrm{d}\theta}{\mathrm{d}t}$$

gives the rate at which the angle changes, in radians per unit of time, with positive values representing counterclockwise motion. Because there are 2π radians in each revolution, and because the period cannot be negative:

$$|\omega| = \frac{2\pi}{T} \qquad T = \frac{2\pi}{|\omega|}$$

The properties of an object in circular motion can be described with the rtz coordinate system, with the **radial axis** r projecting from the object *toward* the circle's center, the **tangential axis** t tangent to the circle and projecting in the counterclockwise direction, and the **perpendicular** axis z perpendicular to the plane of motion. Given vector $\vec{A}$ in the plane of motion, with angle ϕ between the vector and the r-axis:



$\vec{A}_r$ and $\vec{A}_t$ form a right angle at the vector's starting point, and $\vec{A}$ divides that angle, so that the vector's radial and tangential components:

$$A_r = A\cos\phi$$
$$A_t = A\sin\phi$$

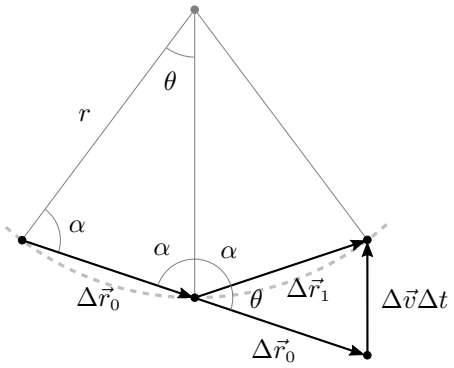
Conversely:

$$A = \sqrt{A_r^2 + A_t^2}$$
$$\phi = \arctan\left(\frac{A_t}{A_r}\right)$$

The velocity components v_r and v_z of an object in circular motion are zero, while the tangential velocity v_t shows the rate at which the object travels the circle. Given arc length s and angular displacement θ, in radians:

$$v_t = \frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}\theta}{\mathrm{d}t}r = \omega r$$

Though its speed never changes, an object in uniform circular motion experiences constant **centripetal acceleration**. With each interval Δt, assuming average velocity $\vec{v}$ for that interval, the circle's center combines with the endpoints of displacement $\Delta\vec{r} = \vec{v}\Delta t$ to form an isosceles triangle. The equal sides of this triangle have length r and interior angle θ. When two such intervals pass, and the beginning of $\Delta\vec{r}_0$ is aligned with the beginning of $\Delta\vec{r}_1$, a similar triangle is formed, the base of which gives the difference between $\Delta\vec{r}_0$ and $\Delta\vec{r}_1$:



so that:

$$\Delta\vec{r}_1 - \Delta\vec{r}_0 = \vec{v}_1\Delta t - \vec{v}_0\Delta t = \Delta\vec{v}\Delta t$$

Because the sides of similar triangles have equal ratios:

$$\frac{|\Delta\vec{v}\Delta t|}{v\Delta t} = \frac{v\Delta t}{r}$$

which allows:

$$a_r = \frac{|\Delta\vec{v}|}{\Delta t} = \frac{v^2}{r} = \omega^2 r$$

with $\vec{a}_r$ pointing at all times toward the center. For a particular v, a_r appears to decrease with r when expressed as v^2/r, and to increase with r when expressed as $\omega^2 r$, but in fact it always *increases*. This is because $v = \omega r$.

To maintain the uniform circular motion, the forces on the r-axis must sum to ma_r, while those on the other axes must sum to zero.

Centrifugal force is the *fictitious force* that seems to pull objects away from the center of motion. In fact, it is simply a manifestation of inertia, and a demonstration that accelerating points cannot be used to define inertial reference frames.

When solving circular motion problems, it is necessary to remember forces like gravity and friction that may not seem relevant at first. Be sure that the elements creating the centripetal force are actually in the plane of motion.

1.9 Circular orbits

When the initial velocity of a projectile is sufficiently large, the earth's curvature can no longer be ignored, as the surface will curve away from the projectile as it moves laterally. The centripetal acceleration of an object in a perfectly circular orbit is proportional to the force of gravity at the

orbital radius. At the earth's surface, given orbital speed v_o:

$$a_r = \frac{v_o^2}{r} = \frac{w}{m} = g$$

This necessitates that:

$$v_o = \sqrt{rg}$$

Though v_o appears to increase as r increases, g decreases with the *square* of r, so that v_o decreases as well. Because the earth's radius is almost 6400 km, and because low Earth orbits can be as near as 160 km, the surface value of g can sometimes provide a reasonable approximation.

Since $T = 2\pi r/v$, the period of a circular orbit:

$$T_o = 2\pi\sqrt{r/g}$$

A satellite in orbit at 160 km has a speed near 28,800 km/h, and a period of approximately 88 minutes. Like any falling object, an orbiting body experiences free fall.

1.10 Non-uniform circular motion

An object coasting or swinging through a circular motion in the vertical plane is subject to two forces: the force of gravity, and a normal or tension force F_r that pushes or pulls it toward the center. This normal force depends on the object's velocity. Since the motion is circular, it is always true that $F_r = mv_t^2/r$, and, at the bottom, it also true that $F_r = n - w$. n must exceed w if the motion is to continue, so the apparent weight at the bottom is greater than w.

At the top of the circle, $F_r = n+w$. Both n and F_r increase with v, becoming arbitrarily large as long as the track or cable does not break. Since w never changes, the minimum F_r (and therefore the minimum v consistent with circular motion at the top) is that where $n = 0$. At this point, the centripetal acceleration is entirely due to the object's weight. This gives $mv^2/r = w$, which yields the **critical speed**, the least speed at the top that will complete the circular motion:

$$v_c = \sqrt{rg}$$

Since $v = \omega r$, this can also be expressed as the **critical angular velocity**:

$$\omega_c = \sqrt{g/r}$$

As long as $v \geq v_c$ at the top, the normal force will be zero or more, the apparent weight will be zero or more *away* from the center, and the circular motion will continue.

For an object in non-uniform circular motion:

$$\vec{a} = \vec{a}_r + \vec{a}_t$$

Whereas centripetal acceleration changes an object's *direction*, **tangential acceleration**:

$$a_t = \frac{\mathrm{d}v_t}{\mathrm{d}t} = \frac{\mathrm{d}\omega}{\mathrm{d}t}r = \frac{\mathrm{d}^2 s}{\mathrm{d}t^2} = \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}r$$

changes its speed. If a_t is constant:

$$\Delta v_t = a_t \Delta t$$

while the arc displacement:

$$\Delta s = v_t \Delta t + \frac{1}{2}a_t[\Delta t]^2$$

Since $\omega = v_t/r$ and $\theta = s/r$, dividing by r produces:

$$\Delta\omega = \frac{a_t}{r}\Delta t$$
$$\Delta\theta = \omega\Delta t + \frac{a_t}{2r}[\Delta t]^2$$

1.11 Action and reaction

Newton's third law holds that every force on some object is matched by another force affecting another object, with the forces being equal in magnitude and opposite in direction. Together, the objects form an **action/reaction pair**. The forces must affect *different objects*; two forces affecting the same object can produce an *action* – if they combine to generate a net force – but they cannot themselves produce a *reaction*. In particular, no pair is formed even if the forces are equal and opposite.

Because no action is possible without a complementary reaction, no interaction can be completely understood without studying all the objects that participate. For convenience, some forces in such interactions are ignored, such as the gravitational attraction exerted by a small falling object on the earth. When both forces are included in the system, they are called **internal forces**. When one force is ignored, the included force is called an **external force**, and is said to be part of the **environment**. As will be seen, internal forces conserve system momentum, but external forces do not.

A motive force produced by an internal energy source is called **propulsion**. When walking, the foot exerts a static friction force against the floor that pushes the floor back,

while the floor exerts an opposing friction force against the foot that prevents it from sliding. The force exerted on the foot is the propulsive force.

Assume an object of mass m is suspended by a cable. If the object is still or moving at a constant velocity, the forces on it must be in equilibrium, so the tension at the end of the cable must equal the object's weight, mg. If the object is accelerating up or down, the net force must be non-zero, and the tension must be greater or less than the object's weight. If two such objects are connected by a cable, and if the cable is suspended by a pulley with the objects hanging at either side, the tension on the cable is still mg. Any tension greater than mg would cause both objects to rise.

Assume objects A and B are connected by a cable, and the objects and the cable are at rest. If A is accelerated away from B so that the cable and B follow it, then the ends of the cable form action/reaction pairs with A and B. If gravity is ignored, the only forces acting on the cable are the tension forces at its ends. The cable has mass m. Since it is accelerating, the tension at the end near A must be ma greater than the tension near B, this being the additional force necessary to accelerate the cable itself. If the cable had zero mass, the tension would be consistent throughout its length, and A and B could be treated as a single action/reaction pair.

When diagramming interactions, create a free-body diagram for each object, being sure that forces are attached to the objects *upon which they act*, rather than those from which they originate. Then draw a dotted line between each force and its corresponding counterforce. Except for external forces, every force should join a force on a different object. The net force on each object should produce the expected motion for that object.

2 Momentum

The **momentum** of an object:

$$\vec{p} = m\vec{v}$$

Assuming m is constant over time, this allows force to be defined as the *rate of momentum change* over time, which is how Newton originally presented his second law:

$$\vec{F} = m\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = \frac{\mathrm{d}\vec{p}}{\mathrm{d}t}$$

Momentum has units $\mathrm{kg} \cdot \mathrm{m/s}$.

Alternatively, since $F_x \,\mathrm{d}t = m\,\mathrm{d}v_x$, and since v_x varies from $v_{x:0}$ to $v_{x:1}$ as t varies from t_0 to t_1, summing over these ranges gives:

$$\int_{t_0}^{t_1} F_x \,\mathrm{d}t = m\int_{v_{x:0}}^{v_{x:1}} \mathrm{d}v_x = m\Big(v_{x:1} - v_{x:0}\Big) = \Delta p_x$$

An **impulsive force** is one that occurs over a short period. More generally, an **impulse**:

$$\vec{J} \equiv \int \vec{F} \,\mathrm{d}t = \Delta\vec{p}$$

The statement that $J = \Delta p$ is called the **impulse-momentum theorem**. J has units $\mathrm{N} \cdot \mathrm{s}$, equivalent to $\mathrm{kg} \cdot \mathrm{m/s}$. If m is constant:

$$\Delta\vec{v} = \frac{\vec{J}}{m}$$

As will be demonstrated, the **angular momentum** of an object in circular motion:

$$L = mrv_t = mr^2\omega$$

with r being the object's distance from the rotation axis. Unlike translational momentum, L has the unit $\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{s}$.

2.1 Conservation of momentum

A **system** is a group of objects that interact with each other. An **isolated system** is one that does not allow matter or energy to enter or exit; where momentum is concerned, this is one for which the net external force on the objects is zero. In particular, external gravitational forces are excluded from isolated systems. A **closed system** allows energy to enter or exit, but prevents matter from doing so. An **open system** allows either.

If two objects interact so that the magnitude of the force on the first object is $F_{\mathrm{s:A}}$, and that on the second is $F_{\mathrm{s:B}}$, and if the forces occur along the same axis:

$$\frac{\mathrm{d}p_{\mathrm{s:A}}}{\mathrm{d}t} = F_{\mathrm{s:A}} \qquad \frac{\mathrm{d}p_{\mathrm{s:B}}}{\mathrm{d}t} = F_{\mathrm{s:B}}$$

Newton's third law requires that $F_{\mathrm{s:A}} = -F_{\mathrm{s:B}}$. Adding these equations:

$$\frac{\mathrm{d}p_{\mathrm{s:A}}}{\mathrm{d}t} + \frac{\mathrm{d}p_{\mathrm{s:B}}}{\mathrm{d}t} = 0$$

shows that total momentum is constant in the absence of an external force. This gives the **law of conservation of momentum**, which states that the total momentum in an *isolated* system is constant. All interactions must be

examined before a system can be considered isolated; for instance, momentum is *not* conserved when a ball bounces against the ground unless the earth and its momentum is considered as well.

An object's velocity constantly changes as it follows a circular path, so its translational momentum (exclusive of the system that contains it) is not conserved. However, the **law of conservation of angular momentum** states that, when the net tangential force is zero, the angular momentum of an object *does* remain constant. As a result, if r changes, v_t will change to hold L constant.

2.2 Rocket propulsion

If m is allowed to vary, force must be related to momentum in a more general way:

$$F = \frac{\mathrm{d}p}{\mathrm{d}t} = m\frac{\mathrm{d}v}{\mathrm{d}t} + v\frac{\mathrm{d}m}{\mathrm{d}t}$$

Rockets propel themselves by expelling *reaction mass* at high velocities. The momentum of the system remains constant as this is done, while that of the *rocket* changes to offset the momentum of the reaction mass.

If m and v are the rocket's starting mass and velocity, then the system's starting momentum:

$$p_0 = mv$$

If $\mathrm{d}m$ is the change in total mass as reaction mass is exhausted, and if v_e is the velocity of the exhausted mass in the v reference frame, then the momentum after this incremental acceleration:

$$p_1 = (m + \mathrm{d}m)(v + \mathrm{d}v) - \mathrm{d}m\, v_\mathrm{e}$$

Notice that $\mathrm{d}m$ is added once and subtracted once. As the rocket's mass *increases* by *negative* quantity $\mathrm{d}m$, the exhaust mass *decreases* by the same negative quantity, leaving the mass of the entire system constant.

If v'_e is the exit velocity of the exhaust relative to the rocket:

$$v_\mathrm{e} = v'_\mathrm{e} + v + \mathrm{d}v$$

with $v + \mathrm{d}v$ being the rocket's velocity after acceleration. This allows:

$$\begin{aligned} p_1 &= (m + \mathrm{d}m)(v + \mathrm{d}v) - \mathrm{d}m(v'_\mathrm{e} + v + \mathrm{d}v) \\ &= m(v + \mathrm{d}v) - \mathrm{d}m\, v'_\mathrm{e} \end{aligned}$$

The momentum difference:

$$\begin{aligned} p_1 - p_0 &= m(v + \mathrm{d}v) - \mathrm{d}m\, v'_\mathrm{e} - mv \\ &= m\,\mathrm{d}v - \mathrm{d}m\, v'_\mathrm{e} \end{aligned}$$

Any change in momentum must be produced by an impulse:

$$\int_{t_0}^{t_1} F\ \mathrm{d}t = p_1 - p_0$$

In the absence of gravity or drag, the net force on the system is zero, so that $\mathrm{d}m\, v'_\mathrm{e} = m\,\mathrm{d}v$. If $u = -v'_\mathrm{e}$ is the *positive* speed at which reaction mass is ejected:

$$-\mathrm{d}m\, u = m\,\mathrm{d}v$$

If $R = -\mathrm{d}m/\mathrm{d}t$ is the rate at which reaction mass is consumed, this allows:

$$\begin{aligned} -\frac{\mathrm{d}m}{\mathrm{d}t}\, u &= m\frac{\mathrm{d}v}{\mathrm{d}t} \\ Ru &= ma \end{aligned}$$

This is the **first rocket equation**. Ru gives the rate of change in momentum, equivalent to *force*. In this case, the force is called **thrust**:

$$T = Ru$$

and it is related by the first equation to the rocket's acceleration, as Newton's second law requires. If $-\mathrm{d}m\, u = m\,\mathrm{d}v$ is instead solved for velocity:

$$\mathrm{d}v = -\frac{\mathrm{d}m}{m}\, u$$

the rocket's *acceleration*:

$$\int_{v_0}^{v_1} \mathrm{d}v = -u\int_{m_0}^{m_1} \frac{\mathrm{d}m}{m}$$

so that:

$$\Delta v = u \ln \frac{m_0}{m_1}$$

This is the **second rocket equation**, which relates acceleration to the consumption of reaction mass.

3 Energy

3.1 Gravitational potential energy

As already shown:

$$v_{y:1}^2 = v_{y:0}^2 + 2a_y(y_1 - y_0)$$

For a falling object *near the earth's surface*, acceleration will be approximately constant, so that $a_y = -g$. Therefore:

$$v_{y:1}^2 + 2gy_1 = v_{y:0}^2 + 2gy_0$$

This shows that $v_y^2 + 2gy$ remains constant over time. Alternatively, because $F_y = ma_y$:

$$F_y = m\frac{\mathrm{d}v_y}{\mathrm{d}t} = m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}v_y}{\mathrm{d}y} = mv_y\frac{\mathrm{d}v_y}{\mathrm{d}y}$$

The ratio $\mathrm{d}v_y/\,\mathrm{d}y$ allows kinetic energy (which varies with v_y) to be related to gravitational potential energy (which varies with y). Because it is also true that $F_y = -mg$:

$$mv_y\,\mathrm{d}v_y = -mg\,\mathrm{d}y$$

v_y varies from $v_{y:0}$ to $v_{y:1}$ as y varies from y_0 to y_1, so summing over these ranges:

$$\int_{v_{y:0}}^{v_{y:1}} mv_y\ \mathrm{d}v_y = \int_{y_0}^{y_1} -mg\ \mathrm{d}y$$

produces:

$$\frac{1}{2}m(v_{y:1}^2 - v_{y:0}^2) = -mg(y_1 - y_0)$$
$$\frac{1}{2}mv_{y:1}^2 + mgy_1 = \frac{1}{2}mv_{y:0}^2 + mgy_0$$

The expression:

$$K = \frac{1}{2}mv^2$$

gives the **kinetic energy** of the object, measured in **joules**:

$$\mathrm{J} = \mathrm{N}\cdot\mathrm{m} = \mathrm{kg}\cdot\mathrm{m}^2/\mathrm{s}^2$$

Because it varies with v^2, K can never be negative. The expression:

$$U_\mathrm{g} = mgy$$

gives the object's **gravitational potential energy**, if g is constant. It is also measured in joules. As shown, the change in kinetic energy for an object in free fall is matched by an opposite change in potential energy, and vice versa. This can be generalized to all forms of potential energy:

$$\Delta K = -\Delta U$$

U_g can be negative, depending on where the origin is placed, but ΔU_g will be the same in all reference frames. Similarly, K will vary when measured from different reference frames, but ΔK will not.

In the absence of friction, the same results are produced for an object sliding on an inclined surface. Given axis s that is parallel to the surface at the object's position, the acceleration along s:

$$F_s = ma_s = m\frac{\mathrm{d}v_s}{\mathrm{d}t} = m\frac{\mathrm{d}s}{\mathrm{d}t}\frac{\mathrm{d}v_s}{\mathrm{d}s} = mv_s\frac{\mathrm{d}v_s}{\mathrm{d}s}$$

The object's weight can be decomposed into two components, one perpendicular to the surface that is opposed by an equivalent normal force, and one parallel. If the s-axis has angle θ relative to the earth's surface, then the parallel component:

$$F_s = -w\sin\theta = -mg\sin\theta$$

so that:

$$mv_s\,\mathrm{d}v_s = -mg\sin\theta\,\mathrm{d}s$$

However, a unit change in s produces a change in y equal to $\sin\theta$, so that $\sin\theta\,\mathrm{d}s = \mathrm{d}y$. This relates the change in velocity along s to the change in height, just as before:

$$mv_s\,\mathrm{d}v_s = -mg\,\mathrm{d}y$$

This holds whether the surface is flat or curved. The only difference is that, as θ decreases, a_s decreases, so that more time is needed to convert a given height into kinetic energy.

An object's **mechanical energy**:

$$E_\mathrm{m} = K + U$$

with U representing all types of potential energy. The **law of conservation of mechanical energy** holds this value to be constant in the absence of resistive forces like friction. This allows different system states to be quantified without understanding the motions that transform one state to another.

3.2 Restoring forces

A **restoring force** is one that returns a system to an equilibrium state. **Elastic systems** are those that contain restoring forces.

If the end of a spring has position s_e along the axis when the spring is in equilibrium, its displacement from equilibrium:

$$\Delta s = s - s_\mathrm{e}$$

Hooke's law states that the spring's force along the axis varies linearly with Δs:

$$F_\mathrm{e} = -k\Delta s$$

The **spring constant** k has units N/m, and is specific to each spring. This is not a true 'law' but it models many springs adequately if they are not over-compressed or over-stretched. In its most general form, the law is written without the negative sign, leaving the direction of the force unstated. Including the sign shows that F_e opposes the displacement, and is thus a restoring force.

3.3 Elastic potential energy

If an object of mass m is connected to the end of a frictionless, massless spring, the net force along the spring's axis will equal $-k(s - s_\mathrm{e})$. The force needed to accelerate the object:

$$ma_s = m\frac{\mathrm{d}v_s}{\mathrm{d}t} = m\frac{\mathrm{d}s}{\mathrm{d}t}\frac{\mathrm{d}v_s}{\mathrm{d}s} = mv_s\frac{\mathrm{d}v_s}{\mathrm{d}s}$$

As before, equating with the restoring force in this system allows:

$$mv_s\,\mathrm{d}v_s = -k(s - s_\mathrm{e})\,\mathrm{d}s$$

This could be integrated directly, with v varying from v_0 to v_1 as s varies from s_0 to s_1. However, substituting $u = s - s_\mathrm{e}$ changes the integration limits to the spring displacements $s_0 - s_\mathrm{e}$ and $s_1 - s_\mathrm{e}$. Because s_e is constant, $\mathrm{d}u = \mathrm{d}(s - s_\mathrm{e}) = \mathrm{d}s$, so that:

$$\int_{v_0}^{v_1} mv_s\ \mathrm{d}v_s = \int_{u_0=(\Delta s)_0}^{u_1=(\Delta s)_1} -ku\ \mathrm{d}u$$
$$\frac{1}{2}m(v_1^2 - v_0^2) = -\frac{1}{2}k\Big[(\Delta s)_1^2 - (\Delta s)_0^2\Big]$$

This shows that the spring's **elastic potential energy**:

$$U_\mathrm{e} = \frac{1}{2}k(\Delta s)^2$$

3.4 Elastic collisions

Following a **perfectly inelastic collision**, objects stick together, after which they necessarily share a common velocity. During an **elastic collision**, objects are compressed, converting kinetic energy into elastic potential energy. The normal forces between the objects increase until they are maximally compressed, then the objects expand, converting the potential energy back to kinetic energy. The normal forces drop to zero as this happens, and the collision ends. The duration of such a collision depends on the construction of the objects, but one to ten milliseconds is common. In a **perfectly elastic collision** this process is perfectly efficient, and all mechanical energy is conserved. Harder materials produce shorter and more perfectly elastic collisions.

Momentum is conserved during all interactions; this follows from Newton's third law, which guarantees that a force producing a momentum change on one object is matched by a force producing an opposite change on some other object. As a result, in the absence of any external force, a system's center of mass undergoes constant, steady motion, even as its components collide or otherwise interact. Though momentum is conserved, kinetic energy is lost if the objects are imperfectly elastic. Because a difference in kinetic energy represents a difference in velocity, which in turn suggests a difference in momentum, this seems to imply that momentum is *not* conserved. However, for any two or more objects, there is a *range* of individual velocities that combine to produce the same total momentum, and different points in this range yield different amounts of kinetic energy. This is why inelastic collisions can convert kinetic energy to thermal or other types of energy without changing the momentum of the system as a whole.

In a perfectly elastic collision between objects A and B, both momentum and mechanical energy will be conserved. If the motion is limited to one dimension, and if A is in motion when the objects meet, and B at rest, the total momentum:

$$m_\mathrm{A}v_{A:1} + m_\mathrm{B}v_{\mathrm{B}:1} = m_\mathrm{A}v_{\mathrm{A}:0}$$

If there is no restoring force, the total energy:

$$\frac{1}{2}m_\mathrm{A}v_{\mathrm{A}:1}^2 + \frac{1}{2}m_\mathrm{B}v_{\mathrm{B}:1}^2 = \frac{1}{2}m_\mathrm{A}v_{\mathrm{A}:0}^2$$

Solving the first equation for $v_{\mathrm{A}:1}$ and substituting into the second eventually produces:

$$v_{\mathrm{B}:1}\left[\left(1 + \frac{m_\mathrm{B}}{m_\mathrm{A}}\right)v_{\mathrm{B}:1} - 2v_{\mathrm{A}:0}\right] = 0$$

This yields two solutions. The first, $v_{\mathrm{B}:1} = 0$, describes the case where the objects do not meet. In the second:

$$v_{\mathrm{B}:1} = \frac{2m_\mathrm{A}}{m_\mathrm{A} + m_\mathrm{B}}v_{\mathrm{A}:0}$$

Returning this to the momentum equation gives:

$$v_{\mathrm{A}:1} = \frac{m_\mathrm{A} - m_\mathrm{B}}{m_\mathrm{A} + m_\mathrm{B}}v_{\mathrm{A}:0}$$

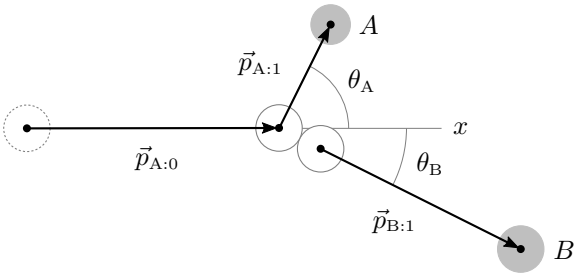
This produces five possible outcomes for a perfectly elastic collision:

- If $m_\mathrm{A} \ll m_\mathrm{B}$, A bounces backward at most of its original speed, and B moves forward slowly;

- If $m_\mathrm{A} < m_\mathrm{B}$, A bounces backward, and B moves forward;
- If $m_\mathrm{A} = m_\mathrm{B}$, A stops, and B moves forward at A's original velocity;
- If $m_\mathrm{A} > m_\mathrm{B}$, A continues forward at a slower rate, and B moves ahead of it at a rate greater than A's original speed;
- If $m_\mathrm{A} \gg m_\mathrm{B}$, A continues forward at nearly its original velocity, and B moves ahead of it at almost twice that rate.

The Galilean transformation of velocity allows these results to be used even when both objects are in motion: simply chose a new frame with velocity V equal to one of the velocities in the original frame. Within the new frame, each object has velocity $v' = v - V$. If they have the same mass, the objects will exchange velocities just as they would if one were at rest.

If the objects are allowed to move in two dimensions, and if the collision is not direct, some energy will be transferred to the axis perpendicular to the original velocity:



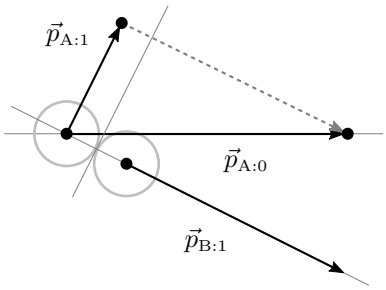
Momentum will be conserved along both axes, so if the original motion follows the x-axis, and if θ_A and θ_B are the counterclockwise angles between that motion and the new paths:

$$m_\mathrm{A} v_{\mathrm{A:0}} = m_\mathrm{A} v_{\mathrm{A:1}} \cos\theta_\mathrm{A} + m_\mathrm{B} v_{\mathrm{B:1}} \cos\theta_\mathrm{B}$$

while on the y-axis:

$$0 = m_\mathrm{A} v_{\mathrm{A:1}} \sin\theta_\mathrm{A} + m_\mathrm{B} v_{\mathrm{B:1}} \sin\theta_\mathrm{B}$$

θ_A and θ_B are determined by the geometry of the impact. As object A strikes B, the normal force accelerates each in opposite directions. The force is perpendicular to the surfaces at the point of contact:



Because momentum is conserved along both axes, any change in $p_{\mathrm{A:}x}$ or $p_{\mathrm{A:}y}$ will be matched by an offsetting change in $p_{\mathrm{B:}x}$ or $p_{\mathrm{B:}y}$, so that the three momentum vectors combine to form a closed triangle. If the masses are equal, the velocity vectors also form such a triangle.

If the collision is perfectly elastic, kinetic energy will be conserved as well, so that:

$$\frac{1}{2} m_\mathrm{A} v_{\mathrm{A:0}}^2 = \frac{1}{2} m_\mathrm{A} v_{\mathrm{A:1}}^2 + \frac{1}{2} m_\mathrm{B} v_{\mathrm{B:1}}^2$$

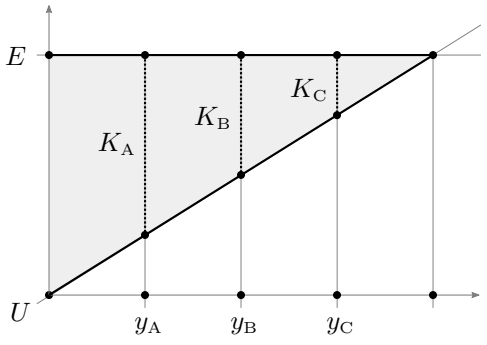
If the masses are equal, this allows:

$$v_{\mathrm{A:0}}^2 = v_{\mathrm{A:1}}^2 + v_{\mathrm{B:1}}^2$$

This recalls the Pythagorean theorem. Since the masses are equal, the velocity vectors produce a closed triangle. If the collision is perfectly elastic, $\vec{v}_{\mathrm{A:1}}$ and $\vec{v}_{\mathrm{B:1}}$ also form a right angle.

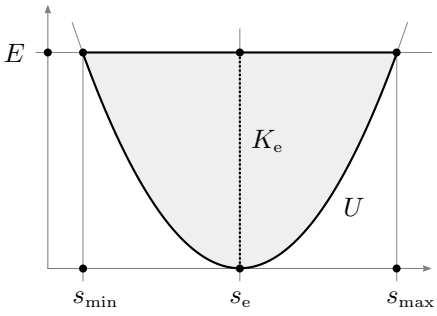
3.5 Energy diagrams

An **energy diagram** graphs an object's energy on the vertical axis against its position on the horizontal. A horizontal line E gives the total energy, a function U shows the potential energy at each position, and $E - U$ produces the kinetic energy K. Since $K > 0$ implies motion, the object has a non-zero velocity wherever $U < E$, and it continues to move until $U = E$. Since K is never negative, the object cannot reach any position where $U > E$:

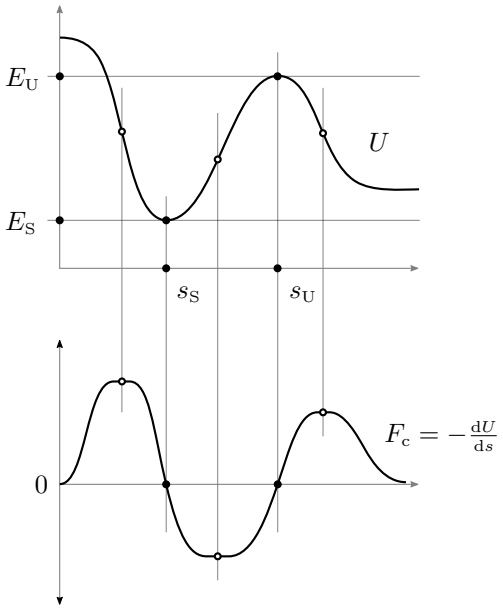


For a bouncing object, the horizontal axis gives the vertical position y, and U is mgy, producing a straight line that intersects the origin. The object has its maximum kinetic energy when y is zero; as it moves upward, U increases and K decreases until U and E meet at the object's maximum height. The object then falls, and U is converted to K until y is zero again, where another bounce occurs. This sequence then repeats.

For an object connected to a spring, the horizontal axis shows the object's axial position s, and U is $-\frac{1}{2}k(s-s_\mathrm{e})^2$, producing a parabola with its vertex where energy is zero and where the displacement is s_e. K cannot be negative, so E must intersect the parabola at one or two points. If it intersects at the vertex, the system contains no energy, and no motion will result. If E is increased, the intersections will show the points at which the spring is most compressed and most stretched, and the system will oscillate between those displacements:



As will be seen, negating the slope of U gives the net *conservative force* acting on the object, so if the slope is zero when $U = E$ (so that $K = 0$) the object will stop; otherwise, it will turn and resume its motion in the opposite direction:



Local minima and maxima in U are *equilibrium positions* where it is possible for the object to rest. Maxima are **unstable equilibria**, since even small increases in E represent motion that will move the object into regions of force that *reinforce* that motion. Minima are **stable equilibria**, since small increases will move the object into regions that *oppose* the motion, leaving the object to oscillate between nearby points. Geometrically, the result is determined by the sign of the slope of the conservative force function ($-\mathrm{d}^2U/\mathrm{d}s^2$) at the equilibrium position. At unstable equilibra, the sign is *positive*, so that forward motion produces a positive force, and backwards motion produces a negative force. At stable equilibria, the sign is *negative*, so that the force is reversed relative to the motion. Regions where U is flat are known as **neutral equilibria**.

4 Work

An object's **thermal energy** E_t is the total kinetic energy of the molecules *within* it, along with the potential energy represented by stretched or compressed molecular bonds. The **system energy** of one or more objects:

$$\begin{aligned} E_\mathrm{s} &\equiv E_\mathrm{m} + E_\mathrm{t} \\ &= K + U + E_\mathrm{t} \end{aligned}$$

The conversion of one energy type to another is called **energy transformation**. The exchange of energy between a system and its environment is called **energy transfer**.

The mechanical transfer of energy to or from a system is called **work**; as will be seen, this is produced by the

application of some force over a displacement. The non-mechanical transfer of energy is called **heat**. Both work and heat are measured in joules. *In the absence of heat*, the work performed on a system:

$$\begin{aligned} W &= \Delta E_\mathrm{s} \\ &= \Delta E_\mathrm{m} + \Delta E_\mathrm{t} \\ &= \Delta K + \Delta U + \Delta E_\mathrm{t} \end{aligned}$$

A process might transfer energy *between* K, U, and E_t, but if their sum increases, W is positive, and energy has also been transferred *into* the system. If W is negative, energy has been transferred *out* of the system.

4.1 Kinetic energy and work

Given a force along axis s:

$$\begin{aligned} F_s &= ma_s = mv_s \frac{\mathrm{d}v_s}{\mathrm{d}s} \\ F_s\,\mathrm{d}s &= mv_s\,\mathrm{d}v_s \end{aligned}$$

Summing over the displacement from s_0 to s_1:

$$\int_{s_0}^{s_1} F_s\ \mathrm{d}s = \frac{1}{2}mv_{s:1}^2 - \frac{1}{2}mv_{s:0}^2$$

As shown earlier, summing force over time produces the change in momentum; now it is seen that work, which sums force over a *displacement*, produces the change in *kinetic energy*:

$$J \equiv \int_{t_0}^{t_1} F_s\ \mathrm{d}t = \Delta p \qquad W \equiv \int_{s_0}^{s_1} F_s\ \mathrm{d}s = \Delta K$$

The statement that $W = \Delta K$ is called the **work-energy theorem**.

Because $p = mv$, K can be expressed in terms of p:

$$K = \frac{p^2}{2m}$$

Since work is performed only by force components that are *parallel* to the displacement, the work performed by a constant force $\vec{F}$ over displacement $\Delta\vec{r}$ is the dot product:

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

A force that coincides with the direction of motion performs *positive* work that *increases* K, while a force that opposes it performs *negative* work that *decreases* K. This is consistent with the idea that positive work transfers energy *into* the system, and negative work transfers it *out*.

4.2 Potential energy and work

A **conservative force** performs the same amount of work over a given displacement, regardless of the shape or length of the path that produces that displacement. Resistive forces are not conservative, since longer paths inevitably produce larger amounts of work. Some form of potential energy can be associated with *any* conservative force, so that the work performed by the force as an object is displaced through it:

$$W_\mathrm{c} = -\Delta U$$

The sign is negative because the potential energy associated with the force *decreases* when the displacement coincides with the direction of the force; this accords with the observation that, for an object in free fall, $\Delta K = -\Delta U$. A single point always represents the same amount of potential energy within a given reference frame. Because nonconservative forces allow different amounts of work to be performed while reaching such a point, they cannot associate a fixed amount of energy with that point without allowing energy to be created or destroyed. In this sense, conservative forces *conserve* mechanical energy.

Because $W_\mathrm{c} = F_\mathrm{c}\Delta s$, it must be that:

$$F_\mathrm{c} = -\frac{\Delta U}{\Delta s}$$

Therefore, the instantaneous conservative force:

$$F_\mathrm{c} = \lim_{\Delta s \to 0} -\frac{\Delta U}{\Delta s} = -\frac{\mathrm{d}U}{\mathrm{d}s}$$

This is the negative of the slope of the U function in an energy diagram.

If W_n gives the work of nonconservative forces, then total work on an object:

$$W = W_\mathrm{c} + W_\mathrm{n}$$

Because mechanical energy is conserved by W_c, any change in E_m must be produced by W_n:

$$\begin{aligned} W_\mathrm{n} &= \Delta E_\mathrm{m} \\ &= \Delta K + \Delta U \end{aligned}$$

4.3 Thermal energy and work

Resistive forces are also known as *dissipative forces*. Such forces are nonconservative, and because they always oppose the direction of motion, they perform negative work that removes E_m from the system. Since they do not contribute to U, the work done by dissipative forces:

$$W_\mathrm{d} = \Delta K$$

The kinetic energy lost this way is converted to thermal energy. If the system is not heated or cooled from the outside:

$$\Delta E_\mathrm{t} = -W_\mathrm{d}$$

Because W_d is always negative, dissipative forces always increase E_t. Since friction and drag affect both the object in motion and the surface or fluid that resists that motion, both object *and* environment must be examined when calculating ΔE_t

4.4 Conservation of energy

Nonconservative forces can be divided into *dissipative* and *external* forces. If W_e is the work performed by nonconservative external forces:

$$W_\mathrm{n} = W_\mathrm{d} + W_\mathrm{e}$$

Therefore:

$$\begin{aligned} W &= W_\mathrm{c} + W_\mathrm{d} + W_\mathrm{e} \\ \Delta K &= -\Delta U - \Delta E_\mathrm{t} + W_\mathrm{e} \end{aligned}$$

so that:

$$W_\mathrm{e} = \Delta K + \Delta U + \Delta E_\mathrm{t}$$

The **law of conservation of energy** holds that the system energy of an *isolated system*, where there is no thermal transfer, and where $W_\mathrm{e} = 0$, is constant. As a result:

$$\begin{aligned} \Delta K + \Delta U + \Delta E_\mathrm{t} &= 0 \\ \Delta E_\mathrm{m} + \Delta E_\mathrm{t} &= 0 \\ \Delta E_\mathrm{s} &= 0 \end{aligned}$$

To solve work problems, it is necessary to understand which forces perform work over a given displacement, whether the work of each force is positive or negative, and which type of energy transfer is represented by the work. In general:

$$K_0 + U_0 + E_{\mathrm{t}:0} + W_\mathrm{e} = K_1 + U_1 + E_{\mathrm{t}:1}$$

with every manifestation of work representing a transfer between two terms on the left. An equation that relates the total energy at one point to that at another is called an **energy equation**.

4.5 Power

Power is the rate at which energy is transferred or transformed:

$$P \equiv \frac{dW}{\mathrm{d}t}$$

Power is measured in **watts**, with W = J/s.

Given constant force $\vec{F}$, $dW = \vec{F} \cdot \mathrm{d}\vec{r}$. Dividing by $\mathrm{d}t$ gives:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \vec{F} \cdot \frac{\mathrm{d}\vec{r}}{\mathrm{d}t}$$

If the angle between $\vec{F}$ and $\vec{v}$ is θ:

$$P = \vec{F} \cdot \vec{v} = Fv\cos\theta$$

In particular, when $\vec{v}$ is constant and directly opposed to a conservative force like gravity, $\vec{F} \cdot \vec{v}$ gives the rate at which potential energy is created. This follows from the fact that K is not changing, and v is the rate of displacement within the field defining that energy.

5 Newton's theory of gravity

Kepler's first law states that planets traverse elliptical orbits with their sun at one focus. His **second law** states that a line drawn between the sun and an orbiting planet covers equal areas over equal time intervals. His **third law** states that, for a given sun, the square of each planet's orbital period varies linearly with the cube of half the orbit's major axis.

Newton's law of gravity states that, for particles of mass m_A and m_B, separated by distance r, the magnitude of the **gravitational force** affecting each of them:

$$F_\mathrm{g} = G\frac{m_\mathrm{A} m_\mathrm{B}}{r^2}$$

with the **gravitational constant**:

$$G \approx 6.67 \times 10^{-11}\mathrm{Nm}^2/\mathrm{kg}^2$$

This can be extended to include spherical objects, or those shaped as spherical shells, in which case r gives the distance between the centers. However, for a particle *inside* a spherical shell, no net gravitational force is produced.

As already seen, *inertial* mass defines an object's mass relative to the acceleration produced by an arbitrary force:

$$m = \frac{F}{a}$$

The **principle of equivalence** holds this value to be identical to the object's **gravitational mass**, defined relative to the gravitational force produced by another object of mass M. Following from Newton's law:

$$m = \frac{r^2}{GM} F_\mathrm{g}$$

Since $F_\mathrm{g} = mg$, the *acceleration due to gravity*:

$$g = \frac{GM}{r^2}$$

As calculated, this produces a value of 9.83m/s², rather than the standard 9.81m/s² for g at sea level, though a scale will measure values lower than 9.83m/s² at most latitudes. Because the earth rotates, a centripetal force is required to maintain an object's position relative to the center of rotation. The scale measures a normal force on the object that partially opposes the gravitational force, and the sum of the normal and gravitational forces must produce the required centripetal force. Near the poles, a scale measures g at 9.83m/s²; near the equator, it measures 9.78m/s², though this value is also affected by the height of the earth's equatorial bulge.

5.1 Gravitational potential energy

ΔU has been equated with the negative work performed by a conservative force, but this provides no *absolute* measure of potential energy; for that, it is necessary to define a zero point for U. In simple gravitation problems, U is defined to be zero where $y = 0$, at the planet's surface, but this is valid only where y is much less than the planet's radius.

If the objects were infinitely distant, the gravitational attraction would be zero. $\Delta U = -W_\mathrm{c}$, so if the objects are moved from center distance r to this infinite distance, the increase in potential energy:

$$\Delta U_\mathrm{g} = -\int_r^\infty F_\mathrm{c} \, \mathrm{d}y$$

The gravitational force is directed toward the center of the opposing object, so $F_\mathrm{c} = -Gm_\mathrm{A}m_\mathrm{B}/y^2$ and:

$$\begin{aligned}\Delta U_\mathrm{g} &= -\int_r^\infty -G\frac{m_\mathrm{A}m_\mathrm{B}}{y^2} \, \mathrm{d}y \\ &= -G\frac{m_\mathrm{A}m_\mathrm{B}}{r}\Big|_r^\infty \\ &= G\frac{m_\mathrm{A}m_\mathrm{B}}{r}\end{aligned}$$

This allows potential energy values to be defined relative to this infinite point, where U_g reaches its maximum value. If the maximum is given a value of zero in absolute terms, the absolute potential energy at any distance r will be the difference between the potential energy at r and that at zero:

$$U_\mathrm{g} = -G\frac{m_\mathrm{A}m_\mathrm{B}}{r}$$

This value can be negative because only *changes* in U_g are significant. The value can be used in energy equations just as $U_\mathrm{g} = mgy$ is, and it remains accurate at any distance. U_g is properly the potential energy of the *system*, not that of one object or the other. If one object is much less massive, this distinction can largely be ignored, since the kinetic energy of the more massive object will change only slightly.

In a system with more than two objects, the gravitational potential energy is the sum of the energies between each pair in the whole. Given a system with three objects, A, B, and C, the total:

$$U_\mathrm{g} = -G\Big(\frac{m_\mathrm{A}m_\mathrm{B}}{r_\mathrm{AB}} + \frac{m_\mathrm{A}m_\mathrm{C}}{r_\mathrm{AC}} + \frac{m_\mathrm{B}m_\mathrm{C}}{r_\mathrm{BC}}\Big)$$

Over time, gravity performs work that changes K. An object's **escape speed** is the minimum starting speed sufficient to prevent the object from returning to some attractive body. To ensure this, the object's velocity must be zero or greater *away* from the body after potential energy has been maximized; this entails that the starting kinetic energy equal or exceed the difference between the starting potential energy and the maximum. For an object to escape the surface of a non-rotating body of mass M and radius R, if there are no drag effects, its speed must equal or exceed:

$$v_\mathrm{e} = \sqrt{\frac{2GM}{R}}$$

5.2 Satellite orbits

a_r equals v^2/r during uniform circular motion. If a satellite with mass m follows a circular orbit around a larger body

of mass M, at distance r from that body's center, it must be the case that:

$$\frac{GMm}{r^2} = ma_r = \frac{mv^2}{r}$$

Therefore, the satellite's speed:

$$v = \sqrt{\frac{GM}{r}}$$

is independent of its mass. The **orbital period**:

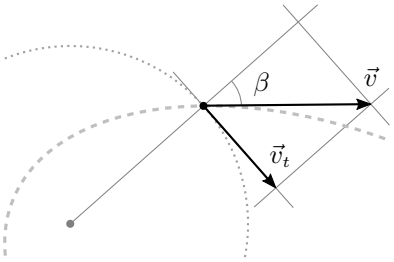
$$T = \frac{2\pi r}{v}$$

Setting T to a planet's rotational period produces a **geosynchronous orbit**. Substituting the result for v produces:

$$T^2 = \frac{4\pi^2}{GM} r^3$$

which is Kepler's third law, fit to a circular orbit.

The t-axis is tangent to a *circle* in the rtz coordinate system, so a tangential component is not necessarily tangent to a non-circular path. The angular momentum of an object in circular motion $L = mrv_t$, and this value remains constant as long as the net tangential force is zero. If a satellite follows an elliptical orbit with instantaneous velocity $\vec{v}$, and if the angle between $\vec{v}$ and the r-axis is β:

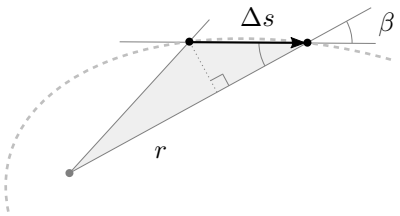


then the magnitude of the tangential velocity $v_t = v \sin\beta$. Therefore:

$$L = mrv \sin\beta$$

Because the gravitational force follows the r-axis, which is always perpendicular to the t-axis, and because no other force affects the object, the angular momentum remains constant, even as $\vec{v}$ changes direction and magnitude.

The satellite experiences displacement $\Delta\vec{s} = \vec{v}_{\text{avg}}\Delta t$ as it orbits during interval Δt. Joining the end points of $\Delta\vec{s}$ with the focus of the orbit produces a triangle. Because the angle between $\Delta\vec{s}$ and the second side is β, the triangle's height is $v_{\text{avg}}\Delta t \sin\beta$:



Bisecting a triangle this way produces two right triangles of the same height, with adjacent sides that sum to r. The total area:

$$\Delta A = \frac{1}{2} r v_{\text{avg}} \Delta t \sin\beta$$

As Δt approaches zero, $\vec{v}_{\text{avg}}$ approaches the instantaneous velocity $\vec{v}$. $rv\sin\beta = L/m$, so the rate at which the area is covered:

$$\lim_{\Delta t \to 0} \frac{\Delta A}{\Delta t} = \frac{1}{2} rv \sin\beta = \frac{L}{2m}$$

Because L is constant, this rate is also constant, thus affirming Kepler's second law.

5.3 Orbital energy

For a satellite in a *circular* orbit, $v = \sqrt{GM/r}$. Therefore, the satellite's kinetic energy:

$$K = \frac{1}{2}mv^2 = \frac{GMm}{2r}$$

Because $U_{\text{g}} = -GMm/r$, it is seen that:

$$K = -\frac{1}{2}U_{\text{g}}$$

The magnitude of $-GMm/r$ decreases as r increases, but the sign is negative, so U_{g} increases with r. The relation between K and U_{g} negates the sign again, so that K decreases with r, as expected. If this ratio between K and U_{g} is not held, the orbit will not be circular.

For circular orbits, any change in energies can be related with:

$$\Delta U_{\text{g}} = -2\Delta K$$

Although K decreases as r increases, U_{g} increases by twice the energy that is lost. The satellite's total mechanical energy throughout this orbit:

$$E_{\text{m}} = K + U_{\text{g}} = \frac{1}{2}U_{\text{g}} = -G\frac{m_{\text{A}}m_{\text{B}}}{2r}$$

The same result holds for elliptical orbits if r is replaced with the length of the semimajor axis.

The zero energy point was associated with the distance at which the attractive force reaches zero. Because E_m is negative, this is seen to be a **bound system**, which is one where a satellite is tied to another body. For the satellite to escape, it would need enough kinetic energy to reach the zero point, and this would require that $K \geq -U_\mathrm{g}$. This equation also defines the energy change necessary to transfer from one orbital radius to another.

5.4 Gravitational fields

Instead of attributing gravitation to 'action at a distance', it is more correct to say that one object's mass produces a spacetime distortion that changes the trajectory of other objects *as if* they were affected by a force. This distortion is called the **gravitational field**.

Fields are represented by **vector fields** that map directions and magnitudes to points in space. If a gravitational field is created by one object, and if another object enters that field, multiplying the mass of the second object by the field strength at its position gives the force affecting it.

Given objects of mass M and m, the magnitude of the gravitational force $F = GMm/r^2$, so the magnitude of the field produced by M is $g = /r^2$. The *spherical* unit vector $\hat{r}$ points *away* from the origin, so placing M at the origin allows the field to be expressed as:

$$\vec{g} = -G\frac{M}{r^2}\hat{r}$$

The magnitude or *strength* of each gravitational field vector is measured in N/kg, equivalent to m/s².

6 Rotation of rigid bodies

Angular acceleration:

$$\alpha \equiv \frac{\mathrm{d}\omega}{\mathrm{d}t}$$

Because $a_t = \mathrm{d}v_t/\mathrm{d}t$ and $v_t = r\omega$, tangential acceleration, for constant r:

$$a_t = r\frac{\mathrm{d}\omega}{\mathrm{d}t}$$

Therefore, just as $v_t = r\omega$:

$$a_t = r\alpha$$

The kinematic equations for translational motion are straightforwardly adapted to rotational motion:

$$\omega_1 = \omega_0 + \alpha\Delta t$$
$$\theta_1 = \theta_0 + \omega_0\Delta t + \frac{1}{2}\alpha[\Delta t]^2$$
$$\omega_1^2 = \omega_0^2 + 2\alpha\Delta\theta$$

Different points on a rotating body will have different tangential speeds and accelerations if they vary in distance from the axis, but they will always have the same angular velocity and angular acceleration.

6.1 Center of mass

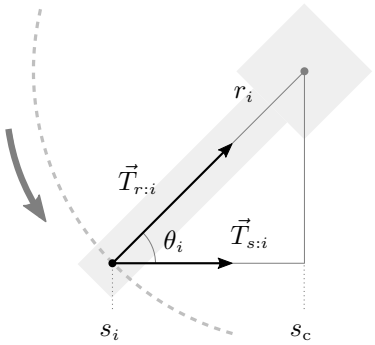
When not constrained by an axle or pivot, the particles in some object will rotate about the center of mass. If the object has mass M, and if the particles have mass m_i and position s_i, the **center of mass** along axis s will be the position-weighted average of the particle masses:

$$s_\mathrm{c} = \frac{1}{M}\sum_i m_i s_i$$

This is seen from the fact that, if particle i is to rotate around the center, it must be subject to a centripetal force:

$$T_{r:i} = m_i a_{r:i} = m_i r_i \omega^2$$

directed toward that point. If the center's angular position relative to the particle is θ_i:



then the s component of the particle's centripetal force:

$$T_{s:i} = T_{r:i}\cos\theta_i = m_i r_i \omega^2 \cos\theta_i$$

Angle θ_i also relates the center's particle-relative horizontal position to its distance from the particle, so that $\cos\theta_i = (s_\mathrm{c} - s_i)/r_i$. As a result:

$$\sum_i T_{s:i} = \sum_i m_i r_i \omega^2 \Big(\frac{s_\mathrm{c} - s_i}{r_i}\Big)$$

$$= \omega^2 \Big(\sum_i m_i s_\mathrm{c} - \sum_i m_i s_i \Big)$$
$$= \omega^2 \Big(M s_\mathrm{c} - \sum_i m_i s_i \Big)$$

The particle forms an action/reaction pair with the center, so the center is affected by an equal force that pulls it toward the particle. If the center is to maintain its position, it must also be subject to a force directed *away* from the particle, so that the tension forces affecting the center along any axis sum to zero:

$$\sum_i T_{s:i} = 0$$

Equating the previous result with zero and solving for s_c produces $\frac{1}{M} \sum_i m_i s_i$.

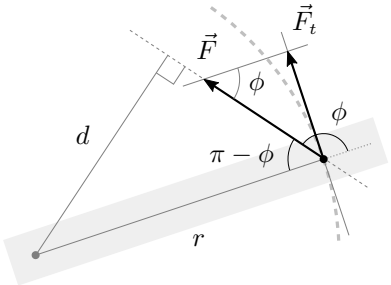
More generally:

$$s_\mathrm{c} = \frac{1}{M} \int s \, \mathrm{d}m$$

An object's mass will not provide ranges for the integration, so $\mathrm{d}m$ must be replaced with an expression of $\mathrm{d}s$ that relates the change in total mass over some interval to the change in position.

6.2 Torque

Given an object that pivots on a fixed point, the **radial line** is that which connects the pivot point with the point at which some force acts. If force $\vec{F}$ is applied such that the counterclockwise angle between the radial line and $\vec{F}$ is ϕ, then the force's tangential component $F_t = F \sin \phi$:



If the distance between the pivot and the point of application is r, the **torque** produced by this force:

$$\tau \equiv r F_t = r F \sin \phi$$

Torque is measured in **newton-meters**, Nm. Though newton-meters are equivalent to joules, torque is *not* a measure of energy, and joules are not used here.

$\vec{F}$ is directed along the **line of action**. Torque increases linearly with r, and is greatest when the line of action is perpendicular to the radial line. When the line of action is parallel, $\sin \phi$ is zero, and the force pulls or pushes the object without producing torque.

The distance between the pivot and the line of action is called the **moment arm** or **lever arm** d. The segment defining the moment arm is always perpendicular to the line of action. When ϕ is not a multiple of $\pi/2$, the moment arm combines with the radial arm and the line of action to produce a right triangle with hypotenuse of length r, and angle $\pi - \phi$ at the point of application. Because $\sin(\pi - \alpha) = \sin \alpha$:

$$d = r \sin \phi$$

The moment arm is a *distance*, not a displacement, so it is always positive. Therefore:

$$|\tau| = dF$$

When forces are applied at multiple points, the object's response is determined by the **net torque**:

$$\tau = \sum_i \tau_i$$

If an object is suspended by an axle, the net torque produced by gravity is the sum of the torque values associated with the particles in the object. For particle i, $|\tau_i| = d_i m_i g$. Because the gravitational force is perpendicular to the x-axis, and because the moment arm is perpendicular to the line of action, placing the axle at the origin allows $d_i = |x_i|$. Particles to the left of the axle produce positive values of ϕ and $\sin \phi$, while particles to the right produce negative values. Therefore:

$$\tau_i = -x_i m_i g$$

Summing these values gives:

$$\tau_\mathrm{g} = -g \sum_i m_i x_i = -gM \cdot \frac{1}{M} \sum_i m_i x_i$$
$$= -gM x_\mathrm{c}$$

where x_c is the position of the center of mass relative to the rotation axis. This allows the gravitational torque to be calculated as if the object's mass were entirely concentrated at its center of mass. The object experiences no torque if the axle coincides with the center of mass, or if it is directly above or below it.

Two equal but opposite forces applied to different points on an object are known as a **couple**. Such forces form parallel

lines of action separated by distance l. If moment arms d_0 and d_1 give the perpendicular distances from the pivot to each line, and if the pivot is somewhere between the lines, the torques will act in the same direction. Therefore:

$$|\tau| = d_0 F + d_1 F = lF$$

Normally, moving a pivot changes τ, but any movement between these lines lengthens one moment arm by the same amount that the other is shortened, so all pivots produce the same torque. This is true even if the pivot is moved *outside* the lines of action. When this happens, the torque from one force opposes the other. As a result, the *difference* between the two moment arms is l, and the resulting net torque equals lF, as before.

If the forces are constant in direction, the lines of action will change as the couple rotates. As the distance between them changes, so will the torque.

6.3 Rotational dynamics

Given a particle of mass m traveling a circular path, a tangential force:

$$F_t = ma_t = mr\alpha$$

produces *tangential* acceleration a_t. Because it is perpendicular to the radial line, the same force generates torque:

$$\tau = rF_t = mra_t = mr^2\alpha$$

that in turn produces *angular* acceleration α. r appears twice in $mr^2\alpha$, first to equate the particle's angular displacement with its movement through the circle, and again to represent the mechanical advantage produced by the moment arm.

All the particles in an object experience the same angular acceleration α, so if τ is the net torque on an object containing particles of mass m_i and radius r_i:

$$\tau = \alpha \sum_i m_i r_i^2$$

Just as an object's inertial mass represents its inherent resistance to linear acceleration, its **moment of inertia**:

$$I = \sum_i m_i r_i^2$$

gives its resistance to angular acceleration, in units $\mathrm{kg \cdot m^2}$. By extension:

$$\alpha = \frac{\tau}{I}$$

Different pivots produce different moments of inertia, just as they produce different amounts of torque for a given tangential force.

More generally, for distance r from the rotation axis:

$$I = \int r^2 \, \mathrm{d}m$$

As before, $\mathrm{d}m$ must be replaced with an expression of $\mathrm{d}r$ that relates changes in total mass to changes in position.

If a one-dimensional object is rotated about some pivot, the moment of inertia can be determined by placing the x-axis origin at the pivot and integrating. However, if the origin of the x'-axis is placed at the center of mass, and if that point is distance d from the pivot, then the x-axis coordinate for a particular point is related to the x' coordinate for that same point by $x = x' + d$. Therefore:

$$\begin{aligned} I &= \int x^2 \, \mathrm{d}m \\ &= \int (x' + d)^2 \, \mathrm{d}m \\ &= \int (x')^2 \, \mathrm{d}m + 2d \int x' \, \mathrm{d}m + d^2 \int \mathrm{d}m \end{aligned}$$

The first of these terms is the moment of inertia about the center of mass, *if* the rotation axis is parallel to the axis running through the pivot; if the axes are not parallel, the x-axis will be foreshortened relative to the x'-axis, and different moments will result. The second term is $2dM$ times the center of mass relative to the x'-axis, and because this point was placed at the origin of that axis, this evaluates to zero. The third term is d^2 times the sum of the mass M. Therefore, if I is the moment of inertia about the pivot, if I_c is the moment of inertia about a *parallel axis* through the center of mass, and if d is the distance between these axes, then I can be determined using the **parallel-axis theorem**:

$$I = I_\mathrm{c} + Md^2$$

Similar arguments extend the theorem to objects with more dimensions. Because Md^2 cannot be less than zero, I is minimized when an object is rotated about its center of mass.

An object is in **translational equilibrium** if the net force on it is zero, giving its center of mass a constant and possibly zero velocity. An object is in **rotational equilibrium** if the net torque about *every point* in the object is zero, giving it a constant and possibly zero *angular* velocity. An

object is in **total equilibrium** if net force and net torque are both zero.

Problems concerning an object in both translational *and* rotational equilibrium can be solved by identifying the forces that affect the object, expressing these forces in terms of their x, y, and z components, and then combining the relevant components into an expression showing the net torque about some pivot. Because the object is in equilibrium, the net forces and the torque can be equated with zero. The resulting system of equations can then be solved.

6.4 Rotational energy

Each particle in a rotating object has kinetic energy:

$$K_i = \frac{1}{2} m_i v_{t:i}^2 = \frac{1}{2} m_i r_i^2 \omega^2$$

Summing these gives the object's **rotational kinetic energy**:

$$K_\mathrm{r} = \frac{1}{2}\Big(\sum_i m_i r_i^2\Big)\omega^2 = \frac{1}{2} I \omega^2$$

The moment of inertia is again seen to play the role that inertial mass plays in translational motion.

If an object is not rotating about its center of mass, its gravitational potential energy could change as it rotates. In the absence of dissipative forces, however, the total mechanical energy:

$$E_\mathrm{m} = K_\mathrm{r} + U_\mathrm{g}$$

will be conserved. This allows motions that convert one type of energy to another to be understood without analyzing the forces that produce the motion.

6.5 Rolling motion

Wheels produce both rolling friction, where they contact the road, and kinetic friction, where they meet bearings. Although sleds produce only kinetic friction, wheels are more efficient than sleds. First, their bearings can be lubricated more effectively than the rails of a sled. Second, the radial distance from the outside of the wheel to the outside of the bearing grants a mechanical advantage that helps the wheel overcome kinetic friction. Assume that force F is required to pull a wheeled vehicle at a steady velocity. Because the vehicle is not accelerating, F must be opposed by an equal force that is produced by rolling and kinetic friction. This force acts between the wheel and the road, and it opposes the direction of the vehicle, causing the wheel to rotate forward. At the outside of the wheel, the rotation is opposed by rolling friction f_r, while at the bearing, it is opposed by kinetic friction f_k. Because the wheel is not accelerating, the net torque must be zero. If R is the outside wheel radius, and R_b the radius at the bearing, this requires that:

$$R_\mathrm{b} f_\mathrm{k} + R f_\mathrm{r} - RF = 0$$

Therefore:

$$F = \frac{R_\mathrm{b}}{R} f_\mathrm{k} + f_\mathrm{r}$$

The larger the wheel, the less force is required to overcome the friction at the bearing.

In one revolution, a wheel moves its center forward by one circumference, so that $\Delta s_\mathrm{c} = 2\pi R$. If v_c is the wheel's velocity, and if T is the time to complete one rotation, then it is also true that $\Delta s_\mathrm{c} = v_\mathrm{c} T$. This produces the **rolling constraint**, which relates the wheel's translational velocity to its tangential velocity:

$$v_\mathrm{c} = \frac{2\pi}{T} R = \omega R$$

Because its tangential velocity exactly opposes the wheel's translational velocity, the point P at the bottom of a wheel is instantaneously at rest if the wheel does not slip. Conversely, because its tangential velocity matches the wheel's translational velocity in magnitude *and* direction, the point at the *top* of the wheel has velocity $2\omega R$. The velocity of points between these two varies linearly with each point's distance from P, and this point can be seen as an instantaneous pivot for the wheel as a whole.

The wheel's total kinetic energy includes both rotational and translational components. The center of mass exhibits translational motion, so if the rotational energy were calculated relative to the center, it would be necessary to include translational kinetic energy when finding K. P is *motionless*, however. For the instant that P serves as a pivot, the wheel moves *around* it, and if it continued to do so, the wheel's translational motion would end.

Therefore, assuming I_P is the moment of inertia about P, the total kinetic energy:

$$K = \frac{1}{2} I_\mathrm{P} \omega^2$$

If the wheel's center of mass is at its center, then by the parallel-axis theorem, $I_\mathrm{P} = I_\mathrm{c} + MR^2$. This produces:

$$K = \frac{1}{2} I_\mathrm{c} \omega^2 + \frac{1}{2} M R^2 \omega^2$$

Because $v_c = \omega R$:

$$K = \frac{1}{2} I_c \omega^2 + \frac{1}{2} M v_c^2 = K_r + K_c$$

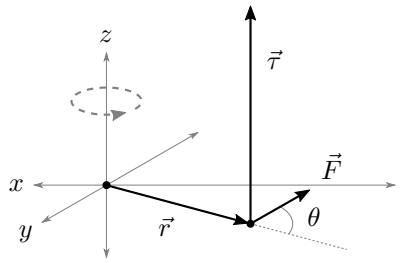
with K_c being the translational kinetic energy of the center of mass. From this it is seen that wheels with greater moments of inertia require greater amounts of energy to achieve a given speed. By extension, spheres, filled cylinders, and hollow cylinders of the same mass roll at different rates down an inclined plane, since the plane provides the same amount of potential energy for a given mass and displacement. A rotating object produces a higher angular velocity for a given amount of energy when more of its mass is concentrated near its center, so a sphere rolls faster than a filled cylinder, which in turn rolls faster than a hollow cylinder.

6.6 Angular momentum

When rotation occurs about a fixed axis, quantities like angular velocity, angular acceleration, and torque can be treated as scalars; for more general problems, these must be represented with vectors. If $\vec{r}$ is the displacement from the axis to the point of application, the torque vector:

$$\vec{\tau} \equiv \vec{r} \times \vec{F}$$

Positive values – which produce counterclockwise acceleration – are represented by vectors that point *toward* the viewer:



$\vec{\tau}$ has a direction and a magnitude, but it does *not* have a specific position. Any acceleration will occur relative to a pivot or the center of mass.

During circular motion, $\vec{v}$ and $\vec{p}$ are always perpendicular to $\vec{r}$, so that angular momentum $L = rp$. This can be generalized to include non-circular motion, where $\vec{p}$ and $\vec{r}$ are not orthogonal:

$$\vec{L} \equiv \vec{r} \times \vec{p}$$

For an object containing particles that each have angular momentum $\vec{L}_i$:

$$\vec{L} = \sum_i \vec{L}_i$$

Euler's rotation theorem guarantees that, given two axes that meet at a fixed point within the object, any combined angular displacement can be reproduced as a *single* rotation about a third axis that crosses the same point. Therefore, though an object might rotate about more than one axis over time, that motion follows a *single* (possibly moving) axis at a given instant. Each particle has one translational momentum at this instant, so its angular momentum – and the sum of all angular momenta in the object – must be a single vector as well. However, an angular momentum vector that represents a complex rotation of this type could also represent a simple rotation around a single axis, so an object's angular momentum does not uniquely define its rotation.

For a constant mass, the rate at which $\vec{L}$ changes over time:

$$\begin{aligned} \frac{\mathrm{d}\vec{L}}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}(\vec{r} \times \vec{p}) = \left(\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} \times \vec{p}\right) + \left(\vec{r} \times \frac{\mathrm{d}\vec{p}}{\mathrm{d}t}\right) \\ &= (\vec{v} \times \vec{p}) + (\vec{r} \times \vec{F}) \end{aligned}$$

Because $\vec{v}$ and $\vec{p}$ have the same direction, their cross product is zero. Therefore, just as $\mathrm{d}\vec{p}/\mathrm{d}t = \vec{F}$:

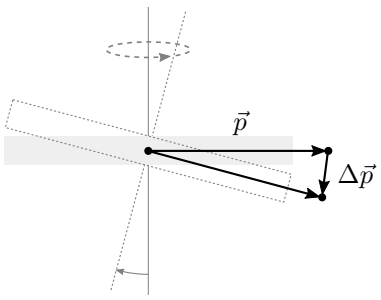
$$\frac{\mathrm{d}\vec{L}}{\mathrm{d}t} = \vec{r} \times \vec{F} = \vec{\tau}$$

The net torque affecting an object is produced both by external and *internal* forces. Because every internal force is part of an action/reaction pair, their torque contribution sums to zero. This yields the **law of conservation of angular momentum**, which states that the direction and magnitude of $\vec{L}$ are conserved within an isolated system.

While it is always true that $L = I\omega$, $\vec{L}$ is not guaranteed to point in the same direction as $\vec{\omega}$ unless the object is rotated *about an axis of symmetry*. When this is done:

$$\vec{L} = I\vec{\omega}$$

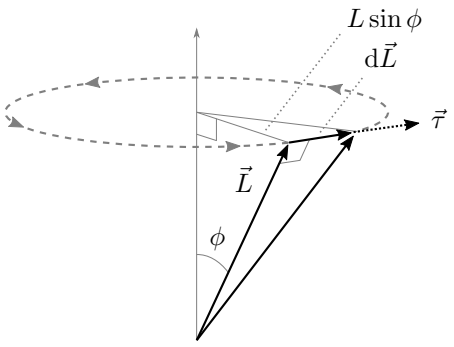
Because angular momentum has a *direction*, it is difficult to reorient the axis of a spinning object such as a gyroscope; to do so, it is necessary to change the momentum of almost every particle in the object. As the tangential velocity of each particle grows, the momentum change necessary to turn the axis to a given angle increases:



6.7 Precession

If a spinning top is not perpendicular to the floor, the top's axis will circle the perpendicular axis in the same direction that the top is spinning. This motion is called **precession**.

The top starts with angular momentum $\vec{L}$, which aligns with the spin axis at angle ϕ from the perpendicular. If $\vec{L}$ is made to start at the pivot, its end will trace a circle of radius $L \sin\phi$ that is centered on the perpendicular axis:



If the top has mass m, and if its center of gravity has displacement $\vec{r}$ relative to the pivot, gravity will produce torque $\tau = mgr \sin\phi$ about the pivot, directed at a right angle to $\vec{L}$ and tangent to the circle. By itself, this torque would rotate the top away from the perpendicular axis, but when combined with $\vec{L}$, it produces a lateral motion that follows $\vec{\tau}$. Because $\vec{\tau} = \mathrm{d}\vec{L}/\mathrm{d}t$, the top's angular momentum changes by:

$$\mathrm{d}L = mgr \sin\phi \cdot \mathrm{d}t$$

causing the end to traverse a small arc on the circle. The circle's radius is $L \sin\phi$, while the arc length is $\mathrm{d}L$, so the angle of this arc:

$$\mathrm{d}\theta = \frac{\mathrm{d}L}{L \sin\theta} = \frac{mgr \sin\phi \cdot \mathrm{d}t}{L \sin\phi}$$

Therefore, the angular velocity of the precession:

$$\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{mgr}{L} = \frac{mgr}{I\omega}$$

$\vec{\tau}$ is always perpendicular to $\vec{L}$ and tangent to the circle, so the motion continues, with ω increasing over time as friction diminishes L. Just as a satellite constantly 'falls' toward the body it orbits, the top constantly rotates toward the floor, but it does so in a way that maintains the overall structure of the system.

The direction of this motion is most easily understood by imagining a spinning wheel that faces the observer. If the wheel turns counterclockwise relative to this viewpoint, particles near the left and right edges will have translational momenta that point down and up. If the wheel is subjected to an upward-pointing torque that rotates the left edge *toward* the observer, particles near the left and right edges will be displaced in space, but their momenta will not change in magnitude or direction. Momenta near the top *will* be changed, however; these left-pointing vectors will be made to point somewhat toward the observer, while right-pointing vectors near the bottom will be made to point away. In order to conserve more of their original momentum, particles near the top will pitch *away* from the observer, while those near the bottom will pitch *toward*, rotating the wheel along a third axis that is orthogonal to both the spin axis and the torque.

The top behaves in a similar manner. As the torque rotates it away from the perpendicular axis, translational momenta on the leading edge turn upward, while those on the trailing edge turn toward the floor. To maintain more of their original momentum, particles on the leading edge *dip*, while those on the trailing edge *rise*, causing the top as a whole to lean in the direction of $\vec{\tau}$.

7 Oscillation

A periodic motion around some equilibrium position is called **oscillatory motion**. Objects that produce such motion are **oscillators**. Whereas the period T gives the time to complete one cycle, the reciprocal of this value gives the number of cycles completed in one unit of time. This is known as the **frequency**:

$$f = \frac{1}{T}$$

Just as an object's angular velocity tells its rate of rotation in radians per second, an oscillator's **angular frequency** gives the rate at which it cycles, also in radians per second:

$$\omega = \frac{2\pi}{T} = 2\pi f$$

7.1 Simple harmonic motion

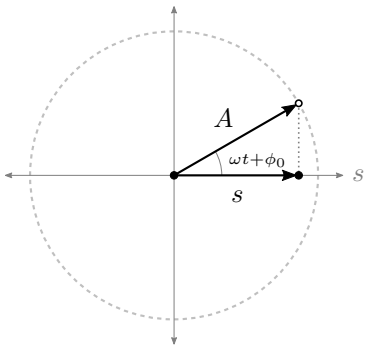
If an object in *uniform circular motion* has angular position ϕ relative to axis s, and if the motion has radius A and is centered on the origin of axis s, the s-axis position:

$$s = A \cos\phi$$

The motion along this axis is an example of **simple harmonic motion**, which produces a sinusoid when graphed against time. The motion's **phase**:

$$\phi = \omega t + \phi_0$$

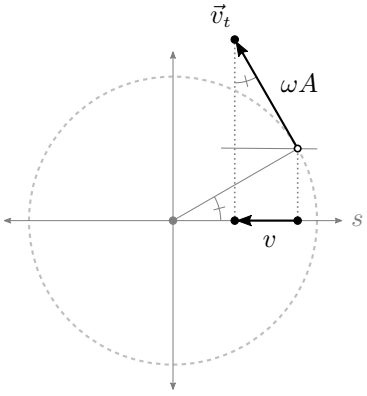
with the **phase constant** ϕ_0 giving the angular position when t is zero:



Therefore, the position at time t:

$$s = A\cos(\omega t + \phi_0)$$

On the circle, the object has tangential velocity $v_t = \omega A$. Projecting this onto the s-axis gives the velocity of the harmonic motion:



which is confirmed by calculating the derivative of the displacement:

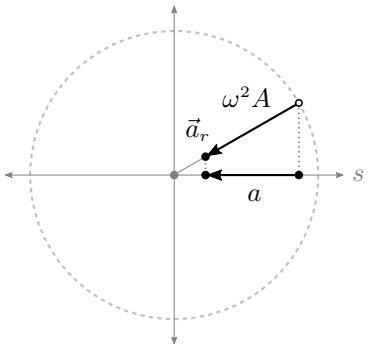
$$v = \frac{\mathrm{d}s}{\mathrm{d}t} = -\omega A\sin(\omega t + \phi_0)$$

Because sine can produce no value greater than one, the maximum speed:

$$v_{\max} = v_t = \omega A$$

The rate of change for any sinusoid is another sinusoid with the same frequency, shifted left in the graph by one quarter-cycle. In this case, when the displacement reaches its greatest magnitude, the acceleration does the same, while the speed is momentarily zero. When the displacement is zero, the acceleration is also zero, while the speed assumes its maximum value.

The acceleration necessary to maintain circular motion $a_r = \omega^2 A$:



which also matches the result obtained by differentiation:

$$a = \frac{\mathrm{d}v}{\mathrm{d}t} = -\omega^2 A\cos(\omega t + \phi_0)$$

Because $s = A\cos(\omega t + \phi_0)$, the acceleration is seen to vary linearly with the position:

$$a = -\omega^2 s$$

Since the equilibrium position s_e is zero, this is consistent with Hooke's law, which states that $F = -k\Delta s$ for spring constant k and $\Delta s = s - s_\mathrm{e}$. The negative relationship between acceleration and displacement shows that a restoring force is at work. Working back from this point, it is seen that simple harmonic motion can be produced by any restoring force that varies linearly with displacement.

Equating $F = ma$ with Hooke's law gives:

$$a = -\frac{k}{m}s$$

which produces:

$$\omega = \sqrt{\frac{k}{m}} \qquad f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \qquad T = 2\pi\sqrt{\frac{m}{k}}$$

From this it is seen that the frequency and period of the motion are *independent of its amplitude.*

Expressing the acceleration as a differential equation gives the **equation of motion** for simple harmonic motion:

$$\frac{\mathrm{d}^2 s}{\mathrm{d}t^2} = -\frac{k}{m}s$$

7.2 Energy of simple harmonic motion

As demonstrated, an object connected to an ideal spring produces simple harmonic motion after being displaced from the equilibrium position. If this position is placed at the origin, then the elastic potential energy:

$$U = \frac{1}{2}ks^2 = \frac{1}{2}kA^2\cos^2(\omega t + \phi_0)$$

Because $v = -\omega A\sin(\omega t+\phi_0)$ and $\omega = \sqrt{k/m}$, the object's kinetic energy:

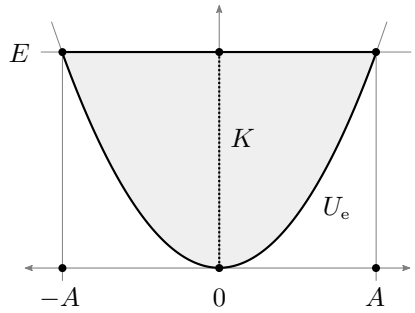
$$K = \frac{1}{2}mv^2 = \frac{1}{2}kA^2\sin^2(\omega t + \phi_0)$$

Therefore, its mechanical energy:

$$\begin{aligned} E &= K + U \\ &= \frac{1}{2}mv^2 + \frac{1}{2}ks^2 \\ &= \frac{1}{2}kA^2\sin^2(\omega t + \phi_0) + \frac{1}{2}kA^2\cos^2(\omega t + \phi_0) \end{aligned}$$

m and k play identical roles in their respective terms. The object's *mass* allows it to store *kinetic* energy, while the spring's *elasticity* allows it to store *potential* energy. All harmonic phenomena are produced by cyclical exchanges between two such energy forms, and similar terms will be found in every example of this behavior.

The energy diagram for this system contains a parabolic U function with its vertex at the origin, where the energy is entirely kinetic. If $-A$ or A is the initial displacement, the parabola intersects E at both these points, and the energy is entirely potential at each of them:



Because $\sin^2\alpha + \cos^2\alpha = 1$:

$$E = \frac{1}{2}kA^2$$

This is also seen by calculating E at displacement $-A$ or A, where there is no kinetic energy. After equating the maximum kinetic energy with the maximum potential energy:

$$\frac{1}{2}mv_{\max}^2 = \frac{1}{2}kA^2$$

solving for $v_{\max}$ gives:

$$v_{\max} = \sqrt{\frac{k}{m}}A = \omega A$$

which matches the earlier result for maximum speed.

Because:

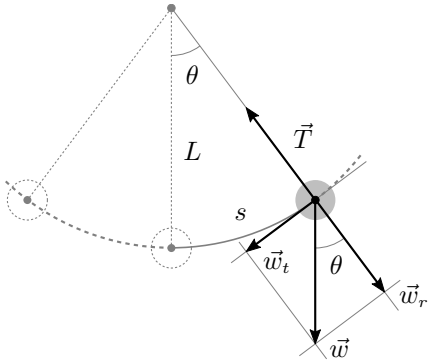
$$mv^2 + ks^2 = kA^2$$

it becomes possible to solve for different variables. For instance, the velocity at s:

$$v = \sqrt{\frac{k}{m}\left(A^2 - s^2\right)} = \omega\sqrt{\left(A^2 - s^2\right)}$$

7.3 Pendulums

The bob at the end of a perfectly rigid pendulum arm will trace a circular arc. In the absence of drag, two forces act on the bob: its weight, and the tension force exerted by the arm, which opposes the weight's radial component. If θ is the angle from the center line to the arm, then the angle between $\vec{w}$ and the radial line is also θ:



so that the tangential component:

$$w_t = -mg\sin\theta$$

Because $\vec{w}_t$ always points toward the equilibrium position, it acts as a restoring force. In this case, however, it varies with the *angular* displacement, not the translational displacement of a spring. If the arm is massless, the tangential acceleration of the bob:

$$a_t = -g\sin\theta$$

If s is the displacement along the arc from the equilibrium position, and if L is the length of the arm, then $s = L\theta$. The small angle approximation holds that $\sin\theta \approx \theta$ when $\theta \ll 1$. This allows $\sin\theta \approx s/L$, making the angular displacement approximately translational:

$$a_t \approx -\frac{g}{L}s$$

This matches the acceleration of an oscillating spring. Extending this to the other findings gives:

$$\omega \approx \sqrt{\frac{g}{L}}$$

which shows the pendulum's frequency and period to be independent (for small amplitudes) of the bob's mass. Much like k and m in an oscillating spring, g relates the storage of potential energy to a displacement, while L relates the storage of kinetic energy to a speed, since a longer arm produces greater tangential velocity for a given angular velocity.

If the arm is *not* massless, the motion must be understood not as a translational motion through a circular path, but as a *rotation* of the entire pendulum. If r is the distance from the pivot to the center of mass, the combined tangential weight produces torque:

$$\tau = -rw_t = -rmg\sin\theta$$

Using small angle approximation again:

$$\tau \approx -rmg\theta = -rw\theta$$

so that:

$$\omega \approx \sqrt{\frac{rw}{I}}$$

for small amplitudes. In this case, the moment arm and the weight relate the storage of potential energy to a displacement, while the moment of inertia I relates the storage of rotational kinetic energy to a speed.

7.4 Damped oscillation

A **damped oscillation** is one that decreases in amplitude over time. Damping is caused by dissipative forces like friction and drag. At low velocities, drag varies with speed in a roughly linear manner, so its damping force can be modeled as:

$$\vec{D} = -b\vec{v}$$

In this context, b is known as the **damping constant**. This quantity produces a force when multiplied by a velocity, so its unit is kg/s.

Simple harmonic motion is produced by a restoring force like $F_\mathrm{r} = -ks$. If that force is combined with a damping force, then the net force:

$$F = -ks - bv = ma$$

From this it is seen that:

$$a + \frac{b}{m}v + \frac{k}{m}s = 0$$

Expressing this as an equation of motion:

$$\frac{\mathrm{d}^2 s}{\mathrm{d}t^2} + \frac{b}{m}\frac{\mathrm{d}s}{\mathrm{d}t} + \frac{k}{m}s = 0$$

This is the original equation for simple harmonic motion with an added $\mathrm{d}s/\mathrm{d}t$ term to represent damping. Because the damping is produced by a dissipative force, the oscillator's energy is not conserved over time.

After solving for the displacement:

$$s = \mathrm{e}^{-bt/2m}A\cos(\omega t + \phi_0)$$

This is a sinusoid with an exponentially decaying **envelope** that gives the greatest possible displacement at each point:

$$s_\mathrm{max} = \mathrm{e}^{-bt/2m}A$$

The angular frequency:

$$\omega = \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}} = \sqrt{\omega_\mathrm{u}^2 - \frac{b^2}{4m^2}}$$

with ω_u representing the undamped angular frequency. Damping lowers the frequency relative to the undamped oscillator, but, as before, the frequency remains constant over time.

A damped oscillator's **time constant**:

$$\tau = \frac{m}{b}$$

gives the relationship between the mass, which resists the damping, and the damping constant. Therefore:

$$s_\mathrm{max} = \mathrm{e}^{-t/2\tau}A$$

Because the total energy E varies with the maximum potential energy at each point, and because the potential energy varies with the maximum displacement:

$$E = \frac{1}{2}ks_\mathrm{max}^2 = \mathrm{e}^{-t/\tau}\cdot\frac{1}{2}kA^2 = \mathrm{e}^{-t/\tau}E_\mathrm{u}$$

E_u is the energy without damping, which is also the energy when $t = 0$. The sign of the exponent is negative, so smaller τ values produce stronger damping. $e^{-1} \approx 0.37$, so when $t = \tau$, the oscillator has approximately 37% of its original energy.

Driven oscillation occurs when an oscillator is subjected to a periodic external force. The rate at which the force acts is called the **driving frequency**, and its effect is represented by the oscillator's **response curve**, which graphs the amplitude of the resulting oscillation against the driving frequency. In this curve, a peak will be found at the oscillator's **natural** or **resonant frequency** f_0, which is the rate at which it *would* oscillate in the absence of driving forces. Smaller damping constants produce taller, narrower peaks.

8 Fluids

A **fluid** is any substance that *flows*, including liquids *and* gases. In a *liquid*, molecules are connected by weak bonds that hold the liquid together while still allowing molecules to move around each other. Because these molecules are close together, liquids are largely incompressible. In a *gas*, molecules move freely without interacting, except when they happen to collide. Because these molecules are loosely distributed, gases are highly compressible.

For an object with mass m and volume V, the **mass density**:

$$\rho = \frac{m}{V}$$

The SI unit for mass density is $\mathrm{kg/m^3}$.

8.1 Pressure

Given a force $\vec{F}$ perpendicular to some area A, the **pressure** acting against the area:

$$p = \frac{F}{A}$$

Pressure is a *ratio* of force to area, not a force itself. The SI unit for pressure is the **pascal**, $\mathrm{Pa} = \mathrm{N/m^2}$.

In a liquid, pressure is produced by gravity, the force of which distributes mechanically against the bottom and sides of the container. Although gravity contributes slightly to the pressure near the bottom of a container full of gas, most of the pressure in a small container is produced by thermal effects.

Atmospheric pressure decreases with height. The **standard atmosphere** is defined as the average pressure at sea level:

$$1\,\mathrm{atm} \equiv 101,300\,\mathrm{Pa} \approx 14.7\,\mathrm{psi}$$

Given a container of *unmoving* liquid, a column can be defined that extends from any area A at depth d to the surface. This column is subject to three forces: a force p_0A due to atmospheric pressure that pushes *down* from the top, another force pA that pushes *up* from the liquid beneath, and the column's weight mg. Because the column is not moving, these forces must balance:

$$pA = p_0A + mg$$

Because $m = \rho dA$, the liquid's **hydrostatic pressure**:

$$p = p_0 + \rho g d$$

so that a change in depth produces a proportional change in pressure:

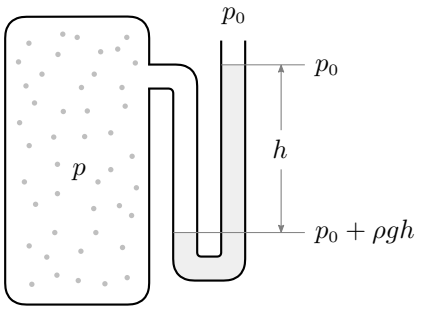
$$\Delta p = \rho g \Delta d$$

This applies to *liquids* because they are mostly incompressible, making ρ constant at all d; the relationships do *not* hold for gases. Because the ρgd term varies only with d, it follows that a change in pressure at any point produces an equal change in pressure at all other points, once equilibrium is achieved; this is called **Pascal's principle**. A contiguous volume of liquid will flow to the same level in all open areas of a container, regardless of its shape, and the pressure will be the same at all points within a given horizontal plane.

Rather than measuring the *absolute* pressure p that is used in most calculations, many gauges show the **gauge pressure**:

$$p_g = p - 1\,\mathrm{atm}$$

Zero gauge pressure represents the *ambient* pressure. The gauge pressure of the gas in some container can be measured with a *manometer*, this being a U-shaped tube, one end of which is connected to the container, and the other end of which is open. Within the tube is a dense liquid such as mercury. The liquid will settle at one level in the branch that is connected to the container, and at another in the open branch:

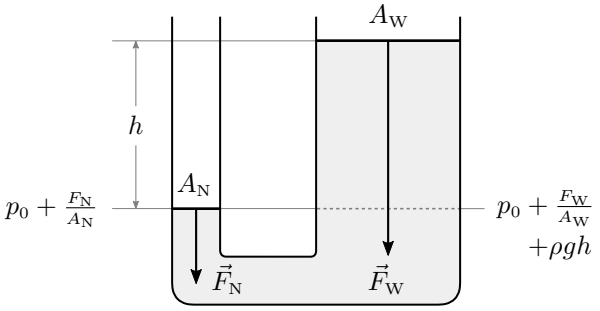


The pressure at the level of the connected branch is equal to the gas pressure; the pressure at the same point in the open branch is equal to the atmospheric pressure, which presses down on the liquid, plus ρgh, where h is the amount by which the open level exceeds the connected level. The pressure at these points is necessarily equal, so the gas pressure is found to be $p_0 + \rho gh$.

Atmospheric pressure can be measured with a *barometer*, which is constructed by sealing a tube at one end, submerging and filling it within a liquid, and then raising the sealed end above the level of the liquid. If the tube is tall enough, a vacuum will form at its top. The pressure at the surface of the liquid matches the ambient pressure, the pressure at the same point within the tube also matches the ambient pressure, and yet the pressure at that point is simultaneously equal to ρgh, where h is the amount by which the level in the tube exceeds the open level. The ambient pressure is therefore found to be ρgh. At one atmosphere, a mercury barometer measures 760 millimeters.

8.2 Hydraulics

A *hydraulic lift* is constructed by connecting a narrow vertical piston with area A_N to a wider one with area A_W:



When force $\vec{F}_\mathrm{N}$ presses down on the narrow piston, the pressure at that piston's face is $p_0 + F_\mathrm{N}/A_\mathrm{N}$. If force $\vec{F}_\mathrm{W}$ presses down on the wider piston, and if h is the vertical distance from the narrow piston to the wider one, the pressure at the same level within the wider piston must be $p_0 + F_\mathrm{W}/A_\mathrm{W} + \rho gh$. This gives:

$$\frac{F_\mathrm{N}}{A_\mathrm{N}} = \frac{F_\mathrm{W}}{A_\mathrm{W}} + \rho gh$$

so that:

$$F_\mathrm{W} = \frac{A_\mathrm{W}}{A_\mathrm{N}} F_\mathrm{N} - A_\mathrm{W} \rho gh$$

When h is small, the force is *multiplied* by a value close to the ratio of the areas. Because the pistons are oriented vertically, the effect diminishes as the weight of the liquid in the wider piston increases.

Because liquids are incompressible, displacing a volume in one piston causes a like volume to be added to the other. If the narrow piston is depressed by distance d_N, the wider piston must rise by distance:

$$d_\mathrm{W} = \frac{A_\mathrm{N}}{A_\mathrm{W}} d_\mathrm{N}$$

This shows in part how energy is conserved by the system; though the wider piston is subject to and reacts with a greater force than the narrow piston, the force acts over a shorter distance. In this vertical orientation, a portion of the input energy is also stored in the liquid as gravitational potential energy.

8.3 Buoyancy

An object submerged in any fluid, whether liquid or gas, is subject to greater pressure on its bottom surface than on its top; this difference produces the **buoyant force**, which pushes the object up. To calculate the force, consider that any shape can be modeled as a collection of vertical cylinders. If a given cylinder has height h, then the pressure difference between its bottom and top $\Delta p = \rho gh$. Because the pressure force $F = pA$, the amount by which the bottom force exceeds the top is equal to ρghA, which is itself the weight of the displaced fluid. This is called **Archimedes' principle**. Given a *completely submerged* object of volume V:

$$F_\mathrm{B} = \rho V g$$

When an object's average density matches that of the fluid surrounding it, its weight is exactly canceled by the buoyant force. Such an object is said to have **neutral buoyancy**.

An object that is *less* dense will float, with part of its volume above the fluid level so that the buoyant force exactly

matches the object's weight. Given an object with volume V_o and homogeneous density ρ_o, and given fluid density ρ_f and *displaced* fluid volume V_f:

$$\rho_\text{f} V_\text{f} g = \rho_\text{o} V_\text{o} g$$

Therefore, the ratio of the displaced fluid to the object volume as a whole:

$$\frac{V_\text{f}}{V_\text{o}} = \frac{\rho_\text{o}}{\rho_\text{f}}$$

8.4 Fluid dynamics

In a **laminar flow**, fluid moves in discrete layers or strands that do not cross, without producing swirls or eddies, and the flow rate at any point within the flow is constant over time. At higher flow rates, the layers and strands begin to mix, and flow rates change over time, producing **turbulent flow**. A flow is **irrotational** if the vector field representing the flow has zero curl at all points.

The **ideal-fluid model** offers a simplified description of motion within fluids. It assumes that the fluid is *incompressible* and *nonviscous*, and that the flow is *laminar* and *irrotational*. In this model, the trajectory followed by a small volume of fluid is called a **streamline**; a collection of adjacent streamlines is called a **flow tube**. Though a flow tube may vary in shape or cross-sectional area, it contains the same set of streamlines throughout its length. A flow tube may traverse an open body of fluid, or it may flow within a chamber or pipe, and passing between these does not affect the tube unless its cross-sectional area changes. Given fluid speed v and cross-sectional area A, the **volume flow rate** at a particular point:

$$Q = vA = \frac{\Delta s}{\Delta t} A = \frac{V}{\Delta t}$$

The SI unit for this quantity is m^3/s. The flow rate must be identical at every point along the tube's length, so V must be constant. This is expressed in the **equation of continuity** as:

$$v_1 A_1 = v_0 A_0$$

with v_0 and v_1 being the fluid speed at two points, and A_0 and A_1 the cross-sectional area at those points. This requires that the fluid move faster in narrower sections of the tube.

Because $pA = F$, a section within any flow tube is subject to two forces that press against its ends. The force on the intake has the same direction as the flow, so it performs positive work on the section equal to:

$$W_0 = F_0(\Delta s)_0 = p_0 A_0 (\Delta s)_0 = p_0 V$$

The force on the outlet *opposes* the flow, so it performs negative work. The total work performed on the section over Δt:

$$W_\text{e} = p_0 V - p_1 V$$

If the intake has vertical position y_0, and if the outlet has position y_1, the change in gravitational potential energy:

$$\begin{aligned}\Delta U &= mgy_1 - mgy_0 \\ &= \rho V g y_1 - \rho V g y_0\end{aligned}$$

The change in kinetic energy:

$$\begin{aligned}\Delta K &= \frac{1}{2} m v_1^2 - \frac{1}{2} m v_0^2 \\ &= \frac{1}{2} \rho V v_1^2 - \frac{1}{2} \rho V v_0^2\end{aligned}$$

Equating the work performed on the section with the changes in kinetic and potential energy gives the energy equation for the flow:

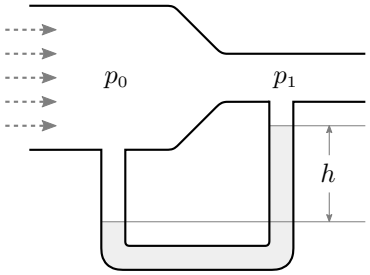
$$\begin{aligned}W_\text{e} &= \Delta K + \Delta U \\ p_0 V - p_1 V &= \frac{1}{2}\rho V v_1^2 - \frac{1}{2}\rho V v_0^2 + \rho V g y_1 - \rho V g y_0\end{aligned}$$

This produces **Bernoulli's equation**:

$$p_1 + \frac{1}{2}\rho v_1^2 + \rho g y_1 = p_0 + \frac{1}{2}\rho v_0^2 + \rho g y_0$$

which shows $p + \frac{1}{2}\rho v^2 + \rho g y$ to be constant at all points within a flow tube. When the vertical position is constant, this causes an *increase* in speed to produce a *decrease* in pressure. When the *velocity* is constant, an increase in gravitational potential energy also produces a decrease in pressure, which matches the results for hydrostatic pressure at different depths. These findings derive from the need to conserve energy within the flow tube.

The flow speed of a gas can be measured with a **Venturi tube**, which consists of a wide chamber followed by a narrow chamber, with the arms of a U-shaped tube connected to each. The tube contains a quantity of liquid:



Gas flows through the chambers; because the second chamber is narrower than the first, the gas pressure there is lower, which causes the level in that arm of the tube to rise, as in a manometer. If p_0 is the pressure in the first chamber, if p_1 is the pressure in the second, if ρ is the density of the liquid, and if h is the amount by which the level in the second arm exceeds that of the first:

$$p_1 = p_0 - \rho g h$$

By the equation of continuity:

$$v_1 = \frac{A_0}{A_1} v_0$$

After substituting into Bernoulli's equation, the potential energy terms can be discarded, since the chambers have the same vertical position. This produces:

$$p_0 + \frac{1}{2}\rho v_0^2 = (p_0 - \rho g h) + \frac{1}{2}\rho\Big(\frac{A_0}{A_1} v_0\Big)^2$$

Solving for v_0 gives:

$$v_0 = A_1\sqrt{\frac{2\rho g h}{\rho(A_0^2 - A_1^2)}}$$

Because gas is compressible, it cannot be considered an ideal fluid. However, this estimate produces adequate results at speeds much below the speed of sound.

Real fluids are at least somewhat viscous. As a solid object moves through a viscous fluid, a thin **boundary layer** of fluid adheres to it. This layer is nearly still relative to the surface of the object. At higher speeds, the boundary layer will separate from the back of the object to form a turbulent, low-pressure region called a **wake**. The difference in pressure between this area and the front contributes to the *drag* force on the object.

9 Elasticity

Hooke's law provides a basic model of the force necessary to stretch an elastic object to a given length. In its general form:

$$F = k\Delta s$$

For most objects, when F is graphed against Δs, the force is seen to vary linearly almost to the end of the *elastic region*, which ends at the **yield strength**, where permanent deformation occurs. The object will not return to its original shape after being stressed to this point, but it will continue to resist the force until the **ultimate strength** is reached, at which point it will rupture.

In Hooke's law, the spring constant is specific to the shape and material of the object. Because the object's macroscopic properties derive ultimately from molecular phenomena, its elasticity at the large scale can be understood by generalizing about molecular bonds. If a rod with cross-sectional area A is pulled with force F, the *force* on each bond must be proportional to F/A. If the rod has length s, and if it is stretched by distance Δs, the *length* by which each bond is stretched must be proportional to $\Delta s/s$. As long as the total force remains within the linear range of the elastic region, each bond can be modeled as a distinct spring. Though the exact force and displacement affecting the bonds is unknown, Hooke's law allows them to be related with:

$$\frac{F}{A} = Y\frac{\Delta s}{s}$$

The constant Y, representing the material's resistance to deformation, is called **Young's modulus**. Rigid materials have higher values for Y. The force per area F/A is the **tensile stress** affecting the object. Though tensile stress is mathematically equivalent to pressure, it acts in a *specific direction*, so it is often expressed in $\mathrm{N/m^2}$ rather than pascals. $\Delta s/s$ is the **strain** affecting the object, this being the relative amount by which it is deformed. As a whole, the equation shows that the *stress* necessary to produce a given *strain* is equal to the strain multiplied by Young's modulus. Strain has no units, so the modulus is expressed in $\mathrm{N/m^2}$, like stress. Young's modulus is also used when modeling *compressive* stress from a single direction. **Shear stress** is produced by forces that are *parallel* to the object's cross section, with the resulting strain transforming rectangular profiles into parallelograms. This type of stress is modeled with the **shear modulus**, which has the same units as Young's modulus, and is calculated in the same way.

Because Hooke's law gives $F/\Delta s = k$:

$$Y = \frac{F}{A}\cdot\frac{s}{\Delta s} = \frac{s}{A}k$$

This allows the modulus to be determined by first measuring k.

Where tensile stress *pulls* an object from a *single* direction, **volume stress** resembles pressure, in that it *pushes* an object from *all* directions. The relative amount $\Delta V/V$ by which the object's volume decreases is called the **volume strain**. Much like tensile stress, volume stress varies linearly with volume strain:

$$\Delta p = -B\frac{\Delta V}{V}$$

so that:

$$B = -\frac{\Delta p}{\Delta V/V}$$

The constant B, representing the material's resistance to compression, is called the **bulk modulus**. Less-compressible materials have higher B values. B is negated relative to Δp because an increase in pressure *decreases* the object's volume.

10 Matter and temperature

The molecules in a **solid** are closely packed, and each is held in place by molecular bonds. In a **crystal**, molecules are arranged in a periodic pattern; in an **amorphous solid**, they are arranged at random. Solids are nearly incompressible.

The molecules in a **liquid** are joined by weaker bonds that hold the liquid together without locking the molecules into place. Liquids are also largely incompressible.

The molecules in a **gas** are *not* bonded to each other, and they rarely interact at all. By volume, gasses are mostly empty space, so they are highly compressible.

An atom's **atomic mass number** is the number of protons and neutrons it contains; the number is displayed with a superscript before the atom's chemical symbol. The **atomic mass unit** u is defined to be one-twelfth the mass of an electrically-neutral ^{12}C atom. An atom's **atomic mass** is equal to its mass number multiplied by the mass unit; this provides a close approximation of the atom's true mass. The **molecular mass** of a molecule is the sum of the atomic masses for each of its atoms.

One **mole** of matter contains a number of molecules equal to the number of atoms in 12 grams of ^{12}C. **Avogadro's number** gives the number of atoms in one mole:

$$N_\mathrm{A} \approx 6.02 \times 10^{23}\,\mathrm{mol}^{-1}$$

From this it is seen that:

$$1\,\mathrm{u} \approx 1.661 \times 10^{-27}\,\mathrm{kg}$$

The **molar mass** of some substance is the mass *in grams* of one mole of that substance. This number is equal to the number of the molecular mass.

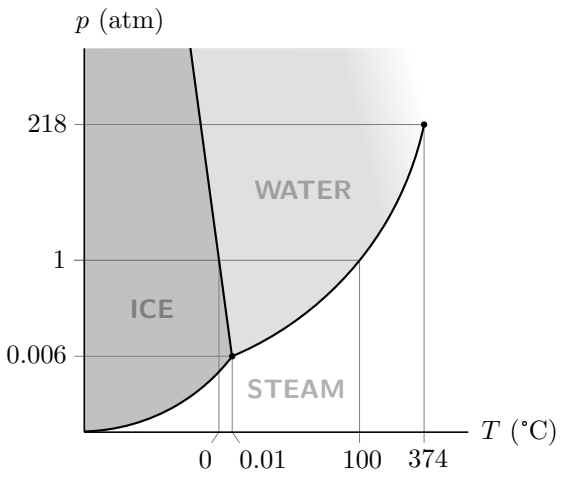
10.1 Temperature

State variables – such as volume, pressure, and temperature – describe the current state of a system, and are used to predict its future behavior. When all state variables are constant over time, a system is in **thermal equilibrium**.

Temperature is a measure of *thermal energy*, which consists of kinetic and potential energy at the molecular level. At absolute zero, there is no atomic motion, so a system's thermal energy is zero. The absolute pressure of a gas in a sealed container increases linearly with its absolute temperature. Attaching a pressure gauge to the container produces a **constant-volume gas thermometer**, which – after being calibrated at two temperatures – can infer the ambient temperature associated with any other pressure.

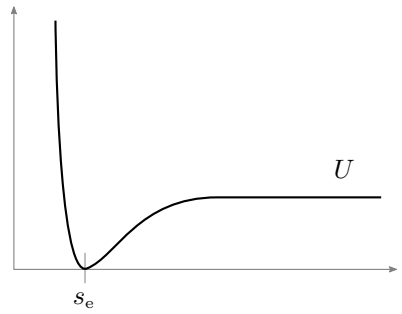
When a solid is heated, its temperature increases until it reaches the **melting** or **freezing point**. Below this point, the substance is entirely solid; above, it is entirely liquid or gas. At this particular temperature it might be any combination of solid and liquid, and though it continues to be heated, the temperature will not rise again until the substance has completely melted. The temperature then rises to the **boiling** or **condensation point**, where another phase change occurs. Temperature does *not* rise during phase changes because the incoming energy is consumed by the breaking of molecular bonds, which builds molecular potential energy without increasing kinetic energy. The system's thermal energy, by contrast, *does* increase during phase changes.

The melting and boiling points for any substance vary with pressure. A **phase diagram** displays temperature on the horizontal axis and pressure on the vertical, and within it shows the three regions where the substance is a gas, liquid, or solid. The lines separating these regions mark the phase transitions:



The gas region occupies the lower-right area, while the upper-left is split between the solid and liquid regions. The regions meet at the **triple point**, which is the only combination of temperature and pressure that allows the three phases to coexist. At pressures below the triple point, the substance **sublimes** directly from solid to gas as the temperature increases. In most substances, the solid form is denser than the liquid, so the phase boundary that rises from the triple point has a positive slope; this implies that increasing the pressure eventually causes a liquid to solidify. Because water ice is *less* dense than liquid water, the phase diagram for water shows a *negative* slope. The **critical point** is found at the far end of the line separating the gas and liquid regions. No clear distinction between gas and liquid exists at temperatures and pressures above this point; no phase transitions occur, and density varies continuously as temperature and pressure vary.

Atoms attract each other when they are close, but they *repel* each other when they are too close. In the energy diagram for this interaction, U has an increasingly steep slope at short distances, showing the strong repulsive force that prevents solids and liquids from being compressed:



The function drops to a trough at the equilibrium position, and then rises to an essentially flat line where the atoms no longer interact. The **ideal gas model** derives from a simplified version of this function, with a vertical line at the contact point, and zero values everywhere beyond. This provides a good approximation of real gas behavior, as long as the density is low and the temperature is well above boiling.

10.2 Ideal gases

Assume that n moles of a gas in thermal equilibrium are held by a container of volume V; the pressure of the gas is p, and the temperature of the gas is T. Experimentally, it is seen that pV varies linearly with nT, and the slope of this variation is the same for gases of *any substance*. This is expressed in the **ideal gas law**:

$$pV = nRT$$

where the **universal gas constant**:

$$R \approx 8.31\,\mathrm{J/mol\,K}$$

In a gas, because there are no phase changes, T varies with the average thermal energy of each molecule, so pV varies with the energy in the system as a whole.

n is constant in a sealed container, so $pV/T = nR$ is constant as well. This gives:

$$\frac{p_1 V_1}{T_1} = \frac{p_0 V_0}{T_0}$$

for all points in time. This is consistent with the notion that pV and T both vary with the total thermal energy of the system. The particular change in each variable depends on the thermodynamic process affecting the system, which constrains certain variables while leaving others open.

If N is the number of molecules in the container, then $n = N/N_\mathrm{A}$. This allows the ideal gas law to be restated as:

$$pV = Nk_\mathrm{B}T$$

with **Boltzmann's constant** giving the gas constant at the molecular scale:

$$k_\mathrm{B} = \frac{R}{N_\mathrm{A}} \approx 1.38 \times 10^{-23}\,\mathrm{J/K}$$

From this it is seen that the gas density, in molecules per cubic meter:

$$\frac{N}{V} = \frac{p}{k_\mathrm{B}T}$$

At standard temperature and pressure, the average distance between gas molecules is approximately 5.7 nanometers.

10.3 Ideal gas processes

Given a *constant quantity* of gas in a piston or other container, the **pressure-volume diagram** graphs pressure against volume, showing how these vary as the system transitions from state to state. Because pV/T is constant in a sealed container, temperature is uniquely determined for each point in the diagram. Technically, the ideal gas law applies only to systems in thermal equilibrium, and such systems do not change; for this reason, the **ideal gas process** described by the diagram is imagined to be a **quasi-static process**, which proceeds so slowly that each point can be considered an equilibrium. This makes ideal gas processes *reversible*, unlike real-world processes.

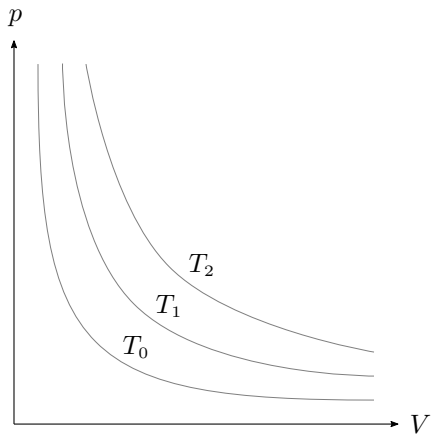
In an **isochoric process**, volume is constant over time, so that pressure varies linearly with temperature; this is represented in the pressure-volume diagram with a vertical line. Heating a constant-volume gas thermometer produces this type of process.

In an **isobaric process**, pressure is constant over time, so that volume varies linearly with temperature; this is represented with a horizontal line. This process occurs when a gas-filled piston is heated while being compressed by a constant force. As the temperature increases, the pressure increases slightly, producing a net force that expands the piston and returns it to the equilibrium pressure.

In an **isothermal process**, temperature is constant over time, so that total thermal energy is constant, and $p_1V_1 = p_0V_0$. This is represented with a hyperbolic curve called an **isotherm**:

$$p = nR \cdot \frac{T}{V}$$

for a given T:



This process can be produced by slowly compressing a gas-filled piston while it is cooled at a constant temperature, so that the work performed on the gas is exactly offset by the heat lost to the environment. This balance between work performed *on* or *by* the system and heat lost *to* or absorbed *from* the environment is part of any isothermal process. Unlike isochoric and isobaric processes, an isothermal process can theoretically be reversed without changing the temperature of the heating or cooling medium.

11 First law of thermodynamics

11.1 Ideal gas processes and work

The **working substance** in a thermodynamic system is the gas or other material that changes state to perform work.

A gas-filled piston with internal pressure p and piston area A exerts force pA on the environment. If the piston expands by $\mathrm{d}s$, the work performed on the environment is $pA\,\mathrm{d}s = p\,\mathrm{d}V$. Conversely, the work performed on the *gas*:

$$\mathrm{d}W = -p\,\mathrm{d}V$$

The same result holds for a volume of any shape that expands through area A with pressure p. Given a process that begins at V_0 and ends at V_1:

$$W = -\int_{V_0}^{V_1} p\,\mathrm{d}V$$

Note that p is likely to vary with V, and if it does, it must be expressed as a function of V.

Work performed *on* the system produces a *decrease* in volume, represented in the pressure-volume diagram as a movement from right to left. Though a low-pressure gas might seem to *pull* on the environment, it is impossible for a gas to pull anything, as its molecules are not bonded. Any compression must be produced by an *outside* force that overcomes the gas pressure; as a result, compression always represents the performance of work *on* the gas, not *by* it.

For processes that decrease the volume, $W = -\int p\,\mathrm{d}V$ is seen to be the area under the pressure-volume curve; for processes that increase the volume, W is the *negative* of that area. Because $\mathrm{d}W$ varies with p at each point, different paths between the start and end points produce different amounts of work. If a set of processes returns the

system to its starting point, then a closed figure is formed in the diagram. If the volume changes at all, some of the processes must be increasing the volume, while others must be decreasing it by a like amount, though possibly at a different pressure. The area within the figure therefore gives the *net work* performed on or by the system. If the movement around the figure is *clockwise*, then the system expands when pressure is high and contracts when it is low; this shows that the work performed *on the environment* is greater than the work performed on the system, and W is negative. Conversely, when the movement is *counterclockwise*, the work performed *on the system* is greater, and W is positive.

V is constant in an isochoric process, so no work is performed. In an *isobaric* process, p is constant, so:

$$W = -p\Delta V$$

In an *isothermal* process, volume and pressure both change. By the ideal gas law, $p = nRT/V$, so that:

$$W = -nRT \int_{V_0}^{V_1} \frac{1}{V}\,\mathrm{d}V = -nRT \ln\left(\frac{V_1}{V_0}\right)$$

Because $nRT = p_0V_0 = p_1V_1$, it is also true that:

$$W = -p_0V_0 \ln\left(\frac{V_1}{V_0}\right) = -p_1V_1 \ln\left(\frac{V_1}{V_0}\right)$$

As will be seen, if this amount of work is performed *on* the system, the same amount of energy must be lost as heat if the temperature is to remain constant.

11.2 Heat

The SI unit for heat is the joule. Historically, the unit for heat was the **calorie**, defined as the amount needed to raise the temperature of one gram of water by one degree Celsius:

$$1\,\mathrm{cal} \approx 4.186\,\mathrm{J}$$

The *large calorie*, *kilogram calorie*, or *food calorie* used to measure food energy is a different unit, with:

$$1\,\mathrm{Cal} = 1000\,\mathrm{cal} \approx 4186\,\mathrm{J}$$

Heat results from a temperature difference between a system and its environment, and, like work, it transfers energy to or from that system. If W is the work performed *on* a system, and if Q is the heat transferred *to* it:

$$\Delta E_\mathrm{s} = W + Q$$

It is also true that $\Delta E_\mathrm{s} = \Delta E_\mathrm{m} + \Delta E_\mathrm{t}$. If E_m is *constant*, so that neither the kinetic nor the potential energy of the system *as a whole* changes:

$$\Delta E_\mathrm{t} = W + Q$$

This is the **first law of thermodynamics**. For a system with constant mechanical energy, the change in thermal energy is equal to the heat absorbed by the system less the work performed *by* the system *on* its environment.

The **specific heat** or **heat capacity** of a substance is the amount of energy needed to raise the temperature of one kilogram of the substance by one kelvin. Given specific heat c and mass M:

$$\Delta E_\mathrm{t} = Mc\Delta T$$

This energy can be supplied through mechanical means, such as mixing, or as heat. By the first law, if no work is performed on the substance:

$$Q = Mc\Delta T$$

A single material has different specific heat values in its solid, liquid, and gas forms. As will be seen, the specific heat of a gas also varies with the *process* that produces the temperature change, and with the amount of work performed by that process.

The **molar specific heat** is the energy needed to raise the temperature of one mole of the substance by one kelvin. Given molar specific heat C and molar quantity n:

$$Q = nC\Delta T$$

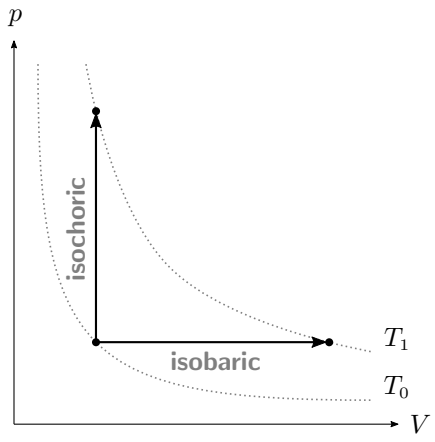
n varies with M according to the molar mass of the substance. Therefore, if M_m is the molar mass, in grams:

$$C = \frac{M_\mathrm{m}}{1000\,\mathrm{g/kg}}\,c$$

Molecular bonds must be broken or formed as a substance changes phase. The **heat of transformation** is the energy consumed or released as one kilogram of the substance undergoes such a change. The **heat of fusion** L_f gives the energy associated with a change between solid and liquid, while the **heat of vaporization** L_v gives the energy for a change between liquid and gas.

11.3 Specific heat of gasses

Assume that an isochoric process and an isobaric process start at the same pressure and volume, and that both end at different points on the same isotherm:



ΔT is the same for both processes. Because phase changes do not occur in an ideal gas, ΔE_t varies directly with ΔT, making ΔE_t the same for each path. Because $\Delta E_\mathrm{t} = W + Q$, any variation in Q for the two processes must be associated with an offsetting variation in W. This is true for *all* processes that cross between the same isotherms.

If C_V is the specific heat of the gas in the isochoric process:

$$Q_\mathrm{V} = nC_\mathrm{V}\Delta T$$

No work is performed without a change in volume, so:

$$\Delta E_\mathrm{t} = nC_\mathrm{V}\Delta T$$

This relationship between ΔE_t and ΔT holds for *all* ideal gas processes, regardless of the particular way in which they move between isotherms. In the isobaric process:

$$W_\mathrm{P} = -p\Delta V$$

According to the ideal gas law, $pV = nRT$. When p is fixed, this allows $p\Delta V = nR\Delta T$, so that:

$$W_\mathrm{P} = -nR\Delta T$$

If C_P is the specific heat of the gas in the isobaric process:

$$Q_\mathrm{P} = nC_\mathrm{P}\Delta T$$

Therefore:

$$\Delta E_\mathrm{t} = W_\mathrm{P} + Q_\mathrm{P} = -nR\Delta T + nC_\mathrm{P}\Delta T$$

This statement is *not* true for non-isobaric processes; their work is not equal to $-p\Delta V$, so it cannot be related to the ideal gas law in the same way.

After equating the two expressions of ΔE_t:

$$nC_\mathrm{V}\Delta T = -nR\Delta T + nC_\mathrm{P}\Delta T$$

and dividing by $n\Delta T$:

$$C_\mathrm{P} = C_\mathrm{V} + R$$

C_P and C_V represent the heat needed to produce a temperature change. Although $\Delta E_\mathrm{t} = nC_\mathrm{V}\Delta T$, regardless of the process, it is now seen that the *heat* necessary to produce ΔT *does* depend on the process, and in particular, the amount of work it performs, here represented by R.

Similarly, two processes that begin and end at the same points produce the same change in thermal energy. Given process H that changes volume at a higher pressure, and process L that does so at a lower pressure:

$$W_\mathrm{H} + Q_\mathrm{H} = W_\mathrm{L} + Q_\mathrm{L}$$

When the volume expands, work performed *on* the system is negative. Because $|W_\mathrm{H}| > |W_\mathrm{L}|$, this requires that $W_\mathrm{H} < W_\mathrm{L}$ and $Q_\mathrm{H} > Q_\mathrm{L}$, so the process that performs more work either absorbs more heat or releases less. When the volume contracts, $W_\mathrm{H} > W_\mathrm{L}$ and $Q_\mathrm{H} < Q_\mathrm{L}$, so the process that receives more work either releases more heat or absorbs less.

11.4 Adiabatic processes

The first law relates ΔE_t to W and Q. In an isothermal process, the temperature does not change, so $W = -Q$. In an isochoric process, no work is performed, so $\Delta E_\mathrm{t} = Q$. In an **adiabatic process**, no *heat* is exchanged, so $\Delta E_\mathrm{t} = W$. This type of process is produced by insulating the system during the volume change, or by changing the volume so quickly that there is no time for heat to transfer. Gas temperature is increased by adiabatic compression, and decreased by adiabatic expansion.

$\Delta E_\mathrm{t} = nC_\mathrm{V}\Delta T$, so in an adiabatic process:

$$W = nC_\mathrm{V}\Delta T$$

Because $\mathrm{d}W = -p\,\mathrm{d}V$:

$$-p\,\mathrm{d}V = nC_\mathrm{V}\,\mathrm{d}T$$

The ideal gas law gives $p = nRT/V$, so:

$$-\frac{nRT}{V}\,\mathrm{d}V = nC_\mathrm{V}\,\mathrm{d}T$$

$$-\frac{R}{C_\mathrm{V}} \cdot \frac{\mathrm{d}V}{V} = \frac{\mathrm{d}T}{T}$$

The **specific heat ratio**:

$$\gamma = \frac{C_\mathrm{P}}{C_\mathrm{V}}$$

Because $R = C_\mathrm{P} - C_\mathrm{V}$:

$$\frac{R}{C_\mathrm{V}} = \frac{C_\mathrm{P} - C_\mathrm{V}}{C_\mathrm{V}} = \gamma - 1$$

$$-\frac{R}{C_\mathrm{V}} = 1 - \gamma$$

Therefore:

$$(1-\gamma)\frac{\mathrm{d}V}{V} = \frac{\mathrm{d}T}{T}$$

Summing over the volume and temperature ranges:

$$(1-\gamma)\int_{V_0}^{V_1} \frac{1}{V}\,\mathrm{d}V = \int_{T_0}^{T_1} \frac{1}{T}\,\mathrm{d}T$$

$$(1-\gamma)\ln\frac{V_1}{V_0} = \ln\frac{T_1}{T_0}$$

$$\left(\frac{V_0}{V_1}\right)^{\gamma-1} = \frac{T_1}{T_0}$$

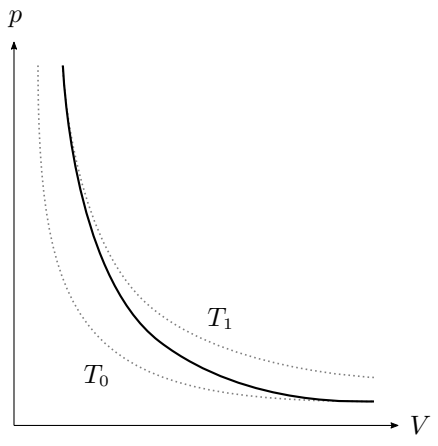
And finally:

$$T_1 V_1^{\gamma-1} = T_0 V_0^{\gamma-1}$$

Similarly, because $T = pV/nR$:

$$p_1 V_1^\gamma = p_0 V_0^\gamma$$

The process is represented in the pressure-volume diagram with an exponential curve called an **adiabat**:



Given constant k, equal to $p_n V_n^\gamma$ for any n in the process:

$$p = \frac{k}{V^\gamma}$$

Because $W = nC_\mathrm{V}\Delta T$, and because $\Delta T = \Delta(pV)/nR$:

$$W = \frac{C_\mathrm{V}}{R}\Delta(pV)$$

As already shown, $R/C_\mathrm{V} = \gamma - 1$, so:

$$W = \frac{p_1 V_1 - p_0 V_0}{\gamma - 1}$$

12 Kinetic theory

12.1 Mean free path

The molecules in a gas have random velocities. If two molecules have velocities $\vec{v}_0$ and $\vec{v}_1$, their relative velocity $\vec{v}_\mathrm{r} = \vec{v}_0 - \vec{v}_1$. Using the dot product to square this vector produces the sum of the squares of its components:

$$\vec{v}_\mathrm{r} \cdot \vec{v}_\mathrm{r} = v_{\mathrm{r}:x}^2 + v_{\mathrm{r}:y}^2 + v_{\mathrm{r}:z}^2$$

By the Pythagorean theorem, this is the square of the vector's magnitude. Following this, the average magnitude can be calculated as the root-mean-square of the relative velocity:

$$\begin{aligned}(v_\mathrm{r})_\mathrm{rms} &= \sqrt{\overline{\vec{v}_\mathrm{r} \cdot \vec{v}_\mathrm{r}}} \\ &= \sqrt{\overline{(\vec{v}_0 - \vec{v}_1) \cdot (\vec{v}_0 - \vec{v}_1)}} \\ &= \sqrt{\overline{\vec{v}_0 \cdot \vec{v}_0} - \overline{2\vec{v}_0 \cdot \vec{v}_1} + \overline{\vec{v}_1 \cdot \vec{v}_1}}\end{aligned}$$

Because the velocities are uncorrelated, the average of their correlation $\overline{\vec{v}_0 \cdot \vec{v}_1}$ is zero, producing:

$$(v_\mathrm{r})_\mathrm{rms} = \sqrt{(\overline{\vec{v}_0 \cdot \vec{v}_0}) + (\overline{\vec{v}_1 \cdot \vec{v}_1})} = \sqrt{\overline{v}_0^2 + \overline{v}_1^2}$$

However, $\overline{v}_0$ and $\overline{v}_1$ equal the average molecular velocity $\overline{v}$. Therefore, the average relative velocity:

$$(v_\mathrm{r})_\mathrm{rms} = \sqrt{2}\,\overline{v}$$

Molecular collisions can be modeled as though each molecule were a sphere with radius r. Two such spheres will collide if the distance between their centers is less than $2r$, so as each sphere moves over period Δt at relative speed $\sqrt{2}\,\overline{v}$, it traverses a cylinder with radius $2r$, and a collision occurs if the center of another sphere enters that cylinder. If the system contains N molecules, and if Δt is short

enough that no cylinders intersect, the total volume of all cylinders:

$$\begin{aligned} V_\text{c} &= \pi(2r)^2 \cdot \sqrt{2}\,\overline{v}\Delta t \cdot N \\ &= 4\sqrt{2}\pi r^2 \overline{v}\Delta t N \end{aligned}$$

If the system has volume V, the probability that one center will be within V_c during Δt:

$$P = \frac{V_\text{c}}{V} = \frac{4\sqrt{2}\pi r^2 \overline{v}\Delta t N}{V}$$

so that the probability density for an arbitrary point in time:

$$\frac{P}{\Delta t} = \frac{4\sqrt{2}\pi r^2 \overline{v} N}{V}$$

The reciprocal of this value is the expected time between collisions. Multiplying by the average speed then gives the average distance between collisions, this being known as the **mean free path**:

$$\lambda = \frac{\Delta t}{P}\overline{v} = \frac{V}{4\sqrt{2}\pi r^2 N} = \frac{1}{4\sqrt{2}\pi r^2 (N/V)}$$

Note that N/V is the molecular density. In a monatomic gas, r is approximately 0.05 nanometers.

12.2 Gas pressure

In a stationary volume of gas, the average of the molecular velocities $(v_s)_\text{avg}$ for any component s is necessarily zero. A more useful summary can be derived from the **root-mean-square speed**, equal to the square root of the average of the velocity squares:

$$v_\text{rms} = \sqrt{(v^2)_\text{avg}}$$

Because $v^2 = v_x^2 + v_y^2 + v_z^2$:

$$\begin{aligned} (v_\text{rms})^2 &= (v^2)_\text{avg} \\ &= (v_x^2)_\text{avg} + (v_y^2)_\text{avg} + (v_z^2)_\text{avg} \end{aligned}$$

However, these components are equal on average, so that:

$$(v_\text{rms})^2 = 3(v_s^2)_\text{avg}$$

for arbitrary component s. Conversely, the average of the square of any component:

$$(v_s^2)_\text{avg} = \frac{1}{3}(v_\text{rms})^2$$

As shown earlier, if one object strikes a much more massive resting object, and if the collision is perfectly elastic, the first object will rebound at nearly its original speed. Because it is so small, a single gas molecule with perpendicular velocity component v_c can be assumed to rebound at velocity $-v_\text{c}$ after striking the side of its container. If the molecule has mass m, its change in momentum:

$$\Delta p = -2mv_\text{c}$$

This momentum change is produced by an impulse, and, by Newton's third law, an equal and opposite impulse affects the container side. If F_c is the average force during the collision, and if Δt_c is the collision length, then this impulse:

$$J = F_\text{c}\Delta t_\text{c} = 2mv_\text{c}$$

Therefore, the average force *during* the collision:

$$F_\text{c} = \frac{2mv_\text{c}}{\Delta t_\text{c}}$$

If every molecule is assumed to have perpendicular speed $|v_\text{c}|$, then the half of these that are moving *toward* the side will travel $\Delta s = v_\text{c}\Delta t_\text{c}$ during Δt_c, and the ones within Δs of the side will strike it. If the side has area A, the volume containing these molecules is $A\Delta s$. If the system as a whole contains N molecules in volume V, the number that will strike the side:

$$N_\text{c} = \frac{A\Delta s}{2}\Big(\frac{N}{V}\Big) = \frac{Av_\text{c}\Delta t_\text{c}}{2}\Big(\frac{N}{V}\Big)$$

The total force from all collisions:

$$F = N_\text{c}F_\text{c} = \frac{Av_\text{c}\Delta t_\text{c}}{2}\Big(\frac{N}{V}\Big)\Big(\frac{2mv_\text{c}}{\Delta t_\text{c}}\Big) = mv_\text{c}^2 A\Big(\frac{N}{V}\Big)$$

To produce a more general result, it can be assumed that $v_\text{c} = (v_s)_\text{rms}$ for perpendicular component s. Because the square of this value is $(v_s^2)_\text{avg}$:

$$\begin{aligned} F &= m(v_s^2)_\text{avg} A\Big(\frac{N}{V}\Big) \\ &= \frac{1}{3}m(v_\text{rms})^2 A\Big(\frac{N}{V}\Big) \end{aligned}$$

Dividing by A gives the pressure against the side:

$$p = \frac{1}{3}m(v_\text{rms})^2\Big(\frac{N}{V}\Big)$$

12.3 Gas temperature

For a molecule with mass m and speed v, the *translational* kinetic energy:

$$\epsilon = \frac{1}{2}mv^2$$

Because $(v^2)_{\text{avg}} = (v_{\text{rms}})^2$, the average of this energy:

$$\epsilon_{\text{avg}} = \frac{1}{2}m(v^2)_{\text{avg}} = \frac{1}{2}m(v_{\text{rms}})^2$$

From this it follows that $(v_{\text{rms}})^2 = 2\epsilon_{\text{avg}}/m$ and:

$$p = \frac{2}{3}\epsilon_{\text{avg}}\frac{N}{V} \qquad pV = \frac{2}{3}\epsilon_{\text{avg}}N$$

By the ideal gas law, $pV = Nk_{\text{B}}T$, so that:

$$\epsilon_{\text{avg}} = \frac{3}{2}k_{\text{B}}T \qquad T = \frac{2}{3}\cdot\frac{\epsilon_{\text{avg}}}{k_{\text{B}}}$$

for molecules with *only translational energy*. Note that energy varies linearly with temperature, as expected. This supports the assumption that molecular collisions are perfectly elastic. If they were not, kinetic energy would decrease with each collision, causing the temperature to drop over time.

By equating $\frac{3}{2}k_{\text{B}}T$ with $\frac{1}{2}m(v_{\text{rms}})^2$, it is seen that the molecular velocity in this gas:

$$v_{\text{rms}} = \sqrt{\frac{3k_{\text{B}}T}{m}}$$

12.4 Thermal energy and specific heat

A system's thermal energy includes the translational and vibrational kinetic energy of the molecules within it, along with the potential energy associated with stretched or compressed molecular bonds:

$$E_{\text{t}} = K_{\text{m}} + U_{\text{m}}$$

The molecules in a monatomic gas have no bonds, and their kinetic energy is entirely translational, so that:

$$E_{\text{t}} = K_{\text{m}} = N\epsilon_{\text{avg}}$$

for a system of N molecules. Therefore:

$$E_{\text{t}} = \frac{3}{2}Nk_{\text{B}}T = \frac{3}{2}nRT$$

By extension:

$$\Delta E_{\text{t}} = \frac{3}{2}nR\Delta T$$

But ΔE_{t} is also related to ΔT by the specific heat of the gas, so:

$$nC_{\text{V}}\Delta T = \frac{3}{2}nR\Delta T$$

Therefore, in any *monatomic* gas, the specific heat during an isochoric process:

$$C_{\text{V}} = \frac{3}{2}R$$

which is confirmed by experiment.

An independent parameter that partially defines the state of some system is called a **degree of freedom**. Together, the degrees of freedom define the **phase space**, which encompasses all possible states for the system. A molecule's translational kinetic energy can be expressed as:

$$\epsilon = \frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2 = \epsilon_x + \epsilon_y + \epsilon_z$$

If the molecule is not bonded to others, as in a gas, then its potential energy is zero; if it is monatomic, its rotational kinetic energy is zero as well. This gives just three degrees of freedom for the storage of energy in a monatomic gas. The **Equipartition theorem** states that the thermal energy of a system in thermal equilibrium is divided equally among its degrees of freedom. Moreover, if energy varies quadratically with a given degree of freedom, then the energy associated with that degree:

$$\epsilon_{\text{Q}} = \frac{1}{2}Nk_{\text{B}}T = \frac{1}{2}nRT$$

Because ϵ_x, ϵ_y, and ϵ_z vary quadratically with v_x, v_y, and v_z, it is seen that the thermal energy of a monatomic gas is equal to $\frac{3}{2}Nk_{\text{B}}T$ or $\frac{3}{2}nRT$, as expected.

The molecules in a solid have three degrees of freedom to store kinetic energy, along with three to store the potential energy of compressed or stretched bonds. Because elastic potential energy varies quadratically with displacement, the thermal energy of a *solid*:

$$E_{\text{t}} = 3Nk_{\text{B}}T = 3nRT$$

By equating this with the specific heat expression, it is predicted that the specific heat of a *solid*:

$$C = 3R$$

which is close to observed values.

Diatomic molecules have three degrees of freedom for translational kinetic energy, and two for rotational energy about the axes perpendicular to the bond. Kinetic and potential energy along the bond's axis would seem to require two more degrees, but at standard temperatures, quantum effects prevent energy from being stored this way. Thus, in a *diatomic* gas:

$$E_{\mathrm{t}} = \frac{5}{2} N k_{\mathrm{B}} T = \frac{5}{2} n R T$$

and:

$$C_{\mathrm{V}} = \frac{5}{2} R$$

which is also close to observed values. At lower temperatures, the rotational degrees are lost, and at higher temperatures, the two vibrational degrees along the axis become active.

12.5 Second law of thermodynamics

Pressure and temperature are *macroscopic* phenomena that summarize and abstract *microscopic* events. A given macroscopic state – such as a particular concentration of thermal energy within a mass – can be produced by a number of different microscopic configurations, and **entropy** is a measure of that number. Some macroscopic states are associated with many more configurations than others, and as the microscopic structure evolves in an essentially random manner, these states – and the macroscopic phenomena they produce – become probabilistically inevitable. This effect is expressed by the **second law of thermodynamics**, which states that the entropy of an isolated system never decreases; instead, it increases until thermal equilibrium is reached, and then it remains constant.

As a result, when two systems touch, molecular collisions cause thermal energy to pass from the hotter system to the cooler one; eventually all degrees of freedom in both systems have the same average energy, giving both systems the same temperature. Assume that systems A and B start at different temperatures. Though the temperatures will change, the total thermal energy must remain constant:

$$E_{\mathrm{AB}} = E_{\mathrm{A:0}} + E_{\mathrm{B:0}} = E_{\mathrm{A:1}} + E_{\mathrm{B:1}}$$

If the systems have the *same specific heat*, and if they contain N_{A} and N_{B} molecules respectively, their average thermal energy per molecule at equilibrium:

$$\frac{E_{\mathrm{A:1}}}{N_{\mathrm{A}}} = \frac{E_{\mathrm{B:1}}}{N_{\mathrm{B}}} = \frac{E_{\mathrm{AB}}}{N_{\mathrm{A}} + N_{\mathrm{B}}}$$

so that:

$$E_{\mathrm{A:1}} = \frac{N_{\mathrm{A}}}{N_{\mathrm{A}} + N_{\mathrm{B}}} E_{\mathrm{AB}}$$

$$E_{\mathrm{B:1}} = \frac{N_{\mathrm{B}}}{N_{\mathrm{A}} + N_{\mathrm{B}}} E_{\mathrm{AB}}$$

13 Heat engines and refrigerators

The gas in a piston presses outward with force $\vec{F}_{\mathrm{g}}$. If the piston is not to be pushed from the cylinder, the environment must counter with an opposing force $\vec{F}_{\mathrm{e}}$. During a quasi-static process, $\vec{F}_{\mathrm{g}} = -\vec{F}_{\mathrm{e}}$, so if W_{E} is the work performed *by* the gas, and if W is the work performed *on* it:

$$W_{\mathrm{E}} = -W$$

This allows the first law of thermodynamics to be restated as:

$$Q = W_{\mathrm{E}} + \Delta E_{\mathrm{t}}$$

This shows that when heat is transferred *to* the system, either work is performed on the environment, or the system's thermal energy is increased.

If the gas in a piston is heated, and if the force that compresses the piston is decreased as it expands so that pV remains constant, then, by the ideal gas law, T will also remain constant. Since ΔE_{t} is zero in this case, the first law requires that $W_{\mathrm{E}} = Q$, and the process is seen to convert heat to work with perfect efficiency. The process does not end where it started, however, so eventually it will be unable to produce more work.

A **thermodynamic cycle** combines processes in a way that returns the system to its original state. A **thermal reservoir** is a system with a heat capacity so great that its temperature and thermal energy can be considered constant even after another system transfers heat to or from it. A **heat engine** uses a thermodynamic cycle to perform useful work. During each cycle, it extracts heat Q_{H} from a high-temperature reservoir and exhausts heat Q_{C} to a low-temperature reservoir; note that, contrary to normal practice, Q_{C} represents energy *lost* by the engine, which absorbs $Q = Q_{\mathrm{H}} - Q_{\mathrm{C}}$ in total. Because each cycle returns the engine to its original state, $\Delta E_{\mathrm{t}} = 0$. Thus, by the first law:

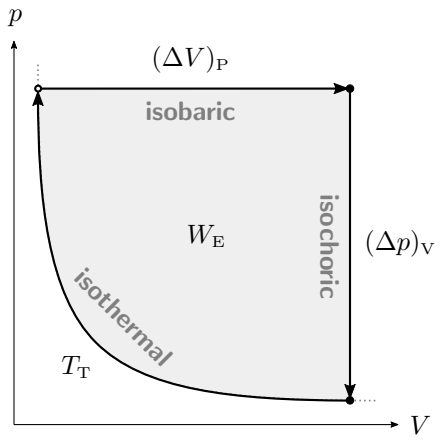
$$W_{\mathrm{E}} = Q_{\mathrm{H}} - Q_{\mathrm{C}}$$

The engine's **thermal efficiency** is the ratio of useful work to heat input:

$$\eta = \frac{W_\mathrm{E}}{Q_\mathrm{H}} = 1 - \frac{Q_\mathrm{C}}{Q_\mathrm{H}}$$

Typical heat engines have thermal efficiencies of 10% to 40%.

Assume that a particular heat engine proceeds through three processes: an isobaric process that performs work by heating a volume of gas, an isochoric process that cools the gas, and an isothermal process that returns the gas to its original state:



Recall that **isochoric** processes have *constant volume* so that:

$$W = 0 \qquad Q = nC_\mathrm{V}\Delta T \qquad \Delta E_\mathrm{th} = Q$$

isobaric processes have *constant pressure* so that:

$$W = -p\Delta V \qquad Q = nC_\mathrm{P}\Delta T \qquad \Delta E_\mathrm{th} = Q + W$$

and **isothermal** processes have *constant temperature* so that:

$$W = -pV\ln\left(\frac{V_1}{V_0}\right) \qquad Q = -W \qquad \Delta E_\mathrm{th} = 0$$

They are not used in this cycle, but **adiabatic** processes exchange *no heat* so that:

$$W = \frac{p_1V_1 - p_0V_0}{\gamma - 1} \qquad Q = 0 \qquad \Delta E_\mathrm{th} = W$$

In the isobaric process, $Q_\mathrm{P} = nC_\mathrm{P}(\Delta T)_\mathrm{P}$ of heat is transferred to the gas, causing $-W_\mathrm{P} = p_\mathrm{P}(\Delta V)_\mathrm{P}$ of work to be performed on the environment. By the ideal gas law, $p\Delta V = nR\Delta T$, so $-W_\mathrm{P} = nR(\Delta T)_\mathrm{P}$.

In the isochoric process, no work is performed, and $Q_\mathrm{V} = nC_\mathrm{V}(\Delta T)_\mathrm{V}$ of heat is exchanged; because $(\Delta T)_\mathrm{V}$ and Q_V are negative, this represents a *release* of thermal energy. The gas is now at its original temperature, but it has less pressure and more volume. $(\Delta T)_\mathrm{V} = -(\Delta T)_\mathrm{P}$, so:

$$Q_\mathrm{P} + Q_\mathrm{V} = n(C_\mathrm{P} - C_\mathrm{V})(\Delta T)_\mathrm{P} = nR(\Delta T)_\mathrm{P}$$

of heat has been absorbed.

In the isothermal process, $W_\mathrm{T} = -nRT_\mathrm{T}\ln(V_\mathrm{V}/V_\mathrm{P})$ of work is performed on the gas to compress it, and an equivalent quantity $-Q_\mathrm{T}$ of heat is released to maintain the temperature while this happens. At the end of one cycle:

$$Q_\mathrm{P} + Q_\mathrm{V} + Q_\mathrm{T} = nR(\Delta T)_\mathrm{P} + nRT_\mathrm{T}\ln(V_\mathrm{V}/V_\mathrm{P})$$

of heat has been transformed into:

$$-W_\mathrm{P} - W_\mathrm{V} - W_\mathrm{T} = nR(\Delta T)_\mathrm{P} + nRT_\mathrm{T}\ln(V_\mathrm{V}/V_\mathrm{P})$$

of work on the environment. As a result:

$$Q_\mathrm{P} + Q_\mathrm{V} + Q_\mathrm{T} = -W_\mathrm{P} - W_\mathrm{V} - W_\mathrm{T}$$

as required by the first law.

A **refrigerator** uses a thermodynamic cycle to move heat in the direction *opposite* that predicted by the second law. In each cycle, it extracts Q_C of heat from a low-temperature reservoir and exhausts Q_H to a high-temperature reservoir, so that it absorbs $Q = Q_\mathrm{C} - Q_\mathrm{H}$ in total. Like the heat engine, $\Delta E_\mathrm{t} = 0$ at the end of each cycle, so that:

$$W = Q_\mathrm{H} - Q_\mathrm{C}$$

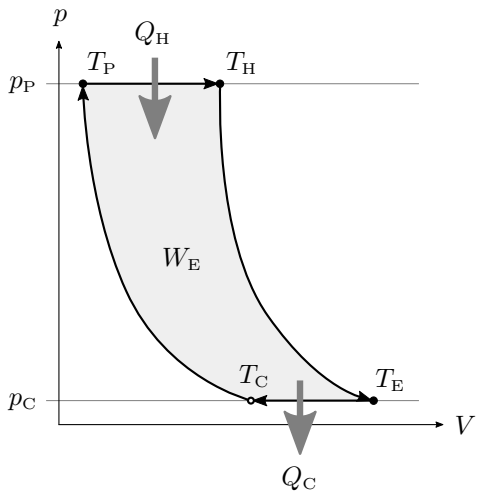
A refrigerator's **coefficient of performance** relates its cooling effect to the work required to produce the cooling:

$$K = \frac{Q_\mathrm{C}}{W}$$

Because the second law prohibits the spontaneous transfer of heat from a low-temperature reservoir to a high-temperature one, W is always greater than zero. By extension, it is impossible to produce a perfectly efficient heat engine; if such an engine did exist, it could be used to produce the work required by a refrigerator. The engine would absorb heat from the high-temperature reservoir, convert it entirely to work, and the refrigerator would return that energy back to the reservoir as waste heat, along with a quantity of heat from the low-temperature reservoir. Taken as a whole, this system would decrease total entropy, which is impossible.

13.1 Brayton cycle

The **Brayton cycle** is used by gas turbine engines:



At the beginning of the cycle, gas is passed through a compressor, producing adiabatic compression that increases the gas temperature. The compressed gas flows through a chamber where it is heated, typically by being mixed with fuel and ignited, or sometimes with a heat exchanger; the chamber is open at the exhaust end, so this heating occurs isobarically. The gas then expands adiabatically through a turbine to produce work, until the starting pressure is reached. At this point, the gas is exhausted, or it is cooled with a heat exchanger before possibly being returned to the engine. In either case, the cooling occurs isobarically, and the gas returns to its initial state. Air-breathing jet engines also use this process, but their turbines extract only enough energy from the gas flow to drive the compressor and possibly a fan; the remaining energy is left to produce thrust.

Though the gas temperature is increased by the compressor, actual heating occurs only during the combustion phase. If T_P is the temperature after compression, and T_H that after combustion, the input heat:

$$Q_H = nC_P(T_H - T_P)$$

Similarly, if T_E is the temperature after the adiabatic expansion, and if T_C is the starting temperature, the exhaust heat:

$$Q_C = nC_P(T_E - T_C)$$

Therefore, the thermal efficiency:

$$\eta_B = 1 - \frac{nC_P(T_E - T_C)}{nC_P(T_H - T_P)}$$
$$= 1 - \frac{T_E - T_C}{T_H - T_P}$$

pV^γ is constant during an adiabatic process. By the ideal gas law, $V^\gamma = (nRT/p)^\gamma$, so:

$$pV^\gamma = p^{1-\gamma}(nRT)^\gamma$$

Because nR is constant, $p^{(1-\gamma)}T^\gamma$ and $p^{(1-\gamma)/\gamma}T$ are constant as well, giving:

$$p_C^{(1-\gamma)/\gamma}T_C = p_P^{(1-\gamma)/\gamma}T_P$$

Therefore:

$$T_C = \left(\frac{p_P}{p_C}\right)^{(1-\gamma)/\gamma}T_P$$

The **pressure ratio**:

$$r_p = \frac{p_P}{p_C}$$

relates the maximum pressure to the minimum, so that:

$$T_C = r_p^{(1-\gamma)/\gamma}T_P$$

By the same reasoning:

$$T_E = r_p^{(1-\gamma)/\gamma}T_H$$

Returning these to the thermal efficiency equation:

$$\eta_B = 1 - \frac{r_p^{(1-\gamma)/\gamma}(T_H - T_P)}{T_H - T_P}$$
$$= 1 - r_p^{(1-\gamma)/\gamma}$$
$$= 1 - \frac{1}{r_p^{(\gamma-1)/\gamma}}$$

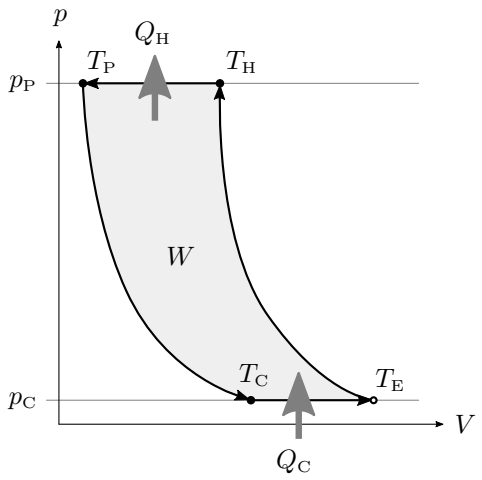
Therefore, higher pressure ratios convert more of the input heat to work.

It can be shown that the work performed by one Brayton cycle:

$$W_{E:B} = nR\left(1 + \frac{1}{\gamma - 1}\right)(T_H - T_P + T_C - T_E)$$

$T_H - T_P$ and $T_C - T_E$ give the temperature changes induced by the two reservoirs.

Some heat engine cycles can be *reversed* to create refrigerators:

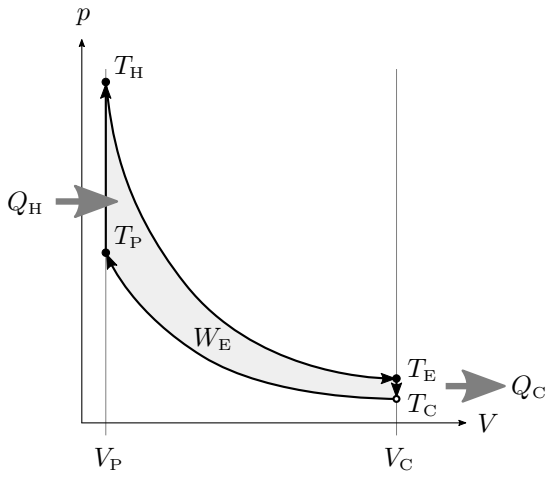


When the Brayton cycle is reversed, it starts, as before, with an adiabatic compression, but the direction is reversed so that the compression occurs in the high-volume portion of the cycle; this increases the temperature without heating the gas, making it hotter than the high-temperature reservoir outside the refrigerator. The gas cools isobarically in this reservoir before being expanded adiabatically to lower its temperature below that inside the refrigerator. Finally, the gas is heated isobarically within the refrigerator; this cools the interior and returns the gas to its initial state.

In total, W work has been performed on the gas, Q_C heat has been absorbed, and $Q_H = Q_C + W$ has been released. Note that, in addition to changing the direction of the process, it is necessary to change the *temperatures* of the two reservoirs. In the engine, heat is transferred from the high-temperature reservoir to the *gas*, so the reservoir must have temperature T_H or *greater*. In the refrigerator, heat is transferred from the gas to the *reservoir*, so the reservoir must have temperature T_P or *lower*. The diagram shows that $T_P < T_H$, so no single reservoir can meet these criteria. A similar problem affects the low-temperature reservoir.

13.2 Otto cycle

The **Otto cycle** is used by spark-ignition engines, commonly known as *gasoline engines*:



At the beginning of each two-stroke cycle, a mix of fuel and air is injected into the piston where it is compressed adiabatically. The fuel mixture is then ignited by a spark plug. Because the fuel burns so quickly, there is no time for the piston to expand, producing an isochoric pressure and temperature increase. The hot gas expands the piston adiabatically to its original volume, where an exhaust value opens, dropping the pressure and temperature isochorically to their initial values. Many engines implement four-stroke cycles that use the next compression and expansion strokes to clear exhaust and take up air and fuel, but this is thermodynamically equivalent to the two-stroke cycle.

It can be shown that the work performed by one Otto cycle:

$$W_{\text{E:O}} = \frac{nR}{1-\gamma}(T_P - T_C + T_E - T_H)$$

Given **compression ratio**:

$$r_v = \frac{V_C}{V_P}$$

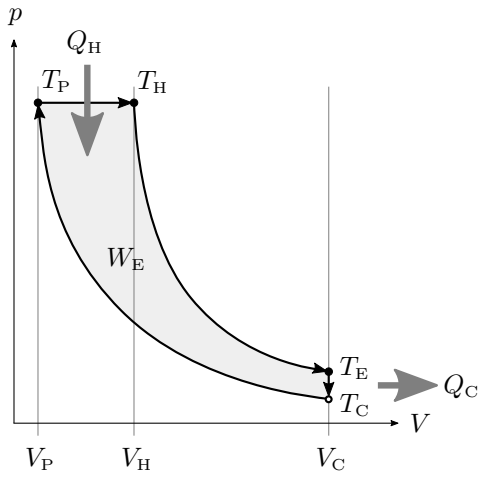
the cycle's thermal efficiency:

$$\eta_O = 1 - \frac{1}{r_v^{\gamma-1}}$$

Higher compression ratios increase efficiency, but they also produce higher temperatures near the end of the compression stroke; if the temperature increases too much, the fuel will ignite on its own, while the piston is still compressing. Fuels with higher **octane ratings** can be compressed more without causing early ignition.

13.3 Diesel cycle

The **Diesel cycle** also has two-stroke and four-stroke variants that are thermodynamically equivalent:



The two-stroke cycle starts by adiabatically compressing a volume of air, increasing its temperature. Fuel is slowly added to the hot air, where it ignites spontaneously, producing an isobaric expansion. The gas is allowed to expand adiabatically, and then it is exhausted, lowering the pressure and temperature to their starting values. Because the gas contains no fuel when it is compressed, higher compression ratios can be used, producing greater thermal efficiency than other combustion engines.

It can be shown that the work performed by one Diesel cycle:

$$W_{\mathrm{E:D}} = nR\Big(\frac{1}{\gamma - 1}(T_{\mathrm{P}} - T_{\mathrm{C}} + T_{\mathrm{E}} - T_{\mathrm{H}}) + (T_{\mathrm{P}} - T_{\mathrm{H}})\Big)$$

Given **cutoff ratio**:

$$r_{\mathrm{c}} = \frac{V_{\mathrm{H}}}{V_{\mathrm{P}}}$$

the cycle's thermal efficiency:

$$\eta_{\mathrm{D}} = 1 - \frac{1}{r_{\mathrm{v}}^{\gamma-1}}\Big(\frac{r_{\mathrm{c}}^{\gamma} - 1}{\gamma(r_{\mathrm{c}} - 1)}\Big)$$

with $r_{\mathrm{v}} = V_{\mathrm{C}}/V_{\mathrm{P}}$ as in the Otto cycle.

13.4 Carnot cycle

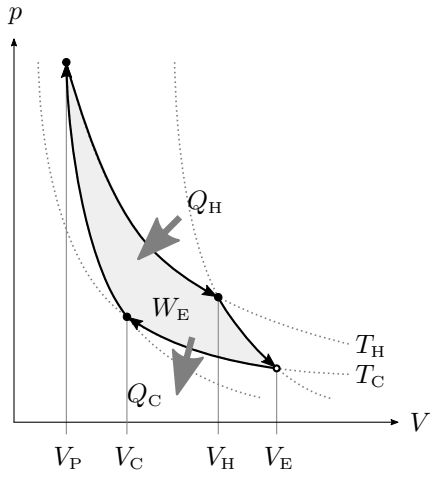
Though some thermodynamic cycles can be reversed to produce refrigerators, it is also necessary that their high-temperature reservoirs be decreased in temperature, and their low-temperature reservoirs increased. However, if a *perfectly reversible engine* were imagined to exist, this device could operate as a heat engine or as a refrigerator just by changing its direction of operation.

A heat engine extracts Q_{H} heat from a high-temperature reservoir and converts it to $-W$ work and Q_{C} waste heat. If the engine is perfectly reversible, it can also act as a refrigerator that uses W work to extract Q_{C} heat from the same low-temperature reservoir, and releases Q_{H} heat to the same high-temperature reservoir. If the engine's output is used to drive the refrigerator, then Q_{H} is absorbed from the high-temperature reservoir, and the same amount is released there by the refrigerator; similarly, Q_{C} is added to the low-temperature reservoir, from which the same amount is then extracted. Taken as a whole, the system changes neither reservoir.

A reversible heat engine is therefore the most efficient engine possible for a given pair of reservoirs, since a more efficient engine could produce the same amount of work by absorbing less heat from the high-temperature reservoir, and exhausting less to the low-temperature one; when combined with the refrigerator, this would decrease total entropy, which is impossible. The same reasoning shows that no refrigerator can be more efficient than a perfectly reversible refrigerator. Moreover, all reversible engines must be equally efficient.

To produce a perfectly reversible engine, it is necessary that the engine be *frictionless* so that all work performed by the gas is output as work. Similarly, all input work must be transferred to the gas without loss. All heat transfers must be reversible as well, but because of the second law, transfers produced by temperature differences *cannot* be reversed. An isobaric expansion proceeds to a high-temperature, high-volume state by absorbing heat from a high-temperature reservoir; after this is done, the only way to return the system to its original low-temperature, low-volume state is to cool it, and that would require a change in the reservoir temperature. An isochoric process proceeds to a high-temperature, high-pressure state in the same way, and again, this cannot be reversed except by cooling. Therefore, any heating or cooling must be performed with *isothermal* processes. These can be reversed by changing the direction of the work, this determining whether the gas absorbs heat from the reservoir (as it performs work on the environment) or whether it releases heat (as work is performed upon it).

An engine that meets these criteria is called a **Carnot engine**:



At the start of the **Carnot cycle**, the gas has temperature T_C equal to that of the low-temperature reservoir. The gas is compressed isothermally to increase its pressure without changing its temperature; as this happens, Q_C heat is passed to the reservoir, and $W_C = Q_C$ work is performed on the gas. Next, it is compressed adiabatically to raise its pressure again, and to increase its temperature to that of the high-temperature reservoir, T_H. The gas is allowed to expand isothermally at this temperature, absorbing Q_H from the high-temperature reservoir, and performing $-W_H = Q_H$ work on the environment. Finally, it is allowed to expand adiabatically, performing additional work as it drops to the starting pressure and temperature.

$W = -nRT\ln(V_1/V_0)$ in an isothermal process, so if V_E is the volume after the adiabatic expansion, and V_C that after the isothermal compression:

$$Q_C = W_C = -nRT_C \ln\left(\frac{V_C}{V_E}\right) = nRT_C \ln\left(\frac{V_E}{V_C}\right)$$

Similarly, if V_P is the volume after the adiabatic compression, and V_H is that after the isothermal expansion:

$$Q_H = -W_H = nRT_H \ln\left(\frac{V_H}{V_P}\right)$$

$TV^{\gamma-1}$ is constant in an adiabatic process, so:

$$T_C V_E^{\gamma-1} = T_H V_H^{\gamma-1} \qquad T_C V_C^{\gamma-1} = T_H V_P^{\gamma-1}$$

and:

$$V_E = V_H\left(\frac{T_H}{T_C}\right)^{1/(\gamma-1)} \qquad V_C = V_P\left(\frac{T_H}{T_C}\right)^{1/(\gamma-1)}$$

Dividing V_E by V_C:

$$\frac{V_E}{V_C} = \frac{V_H}{V_P}$$

In general $\eta = 1 - Q_C/Q_H$, so:

$$\eta_C = 1 - \frac{Q_C}{Q_H} = 1 - \frac{T_C \ln(V_E/V_C)}{T_H \ln(V_H/V_P)}$$

but $V_E/V_C = V_H/V_P$ so:

$$\eta_C = 1 - \frac{T_C}{T_H}$$

This gives the Carnot cycle efficiency, and the maximum efficiency for any heat engine using reservoirs T_C and T_H.

It can be shown that the coefficient of performance for a Carnot refrigerator, and thus the maximum coefficient for any refrigerator using reservoirs T_C and T_H:

$$K_C = \frac{T_C}{T_H - T_C}$$

14 Waves

Mechanical waves traverse a material medium. A disturbance moves some volume of material away from its equilibrium position, and a restoring force pushes it back. For water waves, the restoring force is gravity; for waves on a string, it is the tension force along the string's length. Though the wave moves, and though it transfers energy through the medium, it does not permanently displace any part of the medium.

There are two types of mechanical wave motion. In a **transverse** or **shear wave**, particles oscillate in a direction *perpendicular* to the wave's travel, as in the waves on a string. In a **longitudinal** or **compression wave**, the oscillation is *parallel* to the travel, as when sound passes through a gas or liquid. Some waves, like those in water, produce both types of motion.

Oscillations produce **displacement** within the medium at a given position and time. In a string, this is the perpendicular displacement of a particle at some position along the length. Fluid particles do not have fixed positions, but if some volume is chosen that is large relative to the mean free path of the particles, and small relative to the wave motion, then displacement can be understood as the movement of this volume

Given a string of length L and mass m, the string's **linear density**:

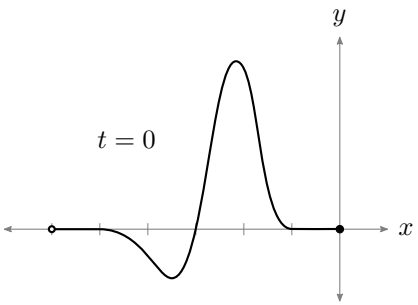
$$\mu = \frac{m}{L}$$

As will be seen, the wave speed depends entirely on the restoring force and the linear density, and *not* on the amplitude, frequency, or shape of the wave. If T is the tension force within the string, the speed:

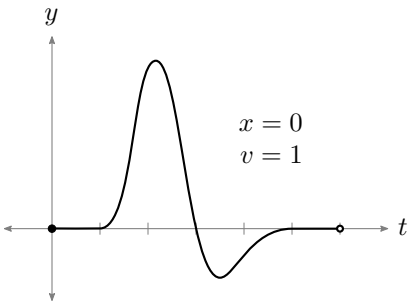
$$v = \sqrt{\frac{T}{\mu}}$$

so that v *increases* with T, and *decreases* with μ.

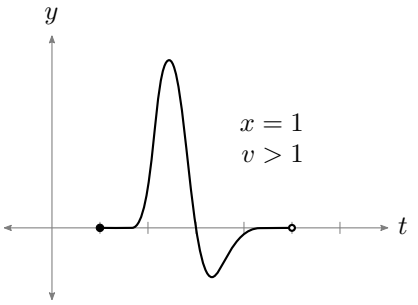
A one-dimensional wave can be partially represented with a **snapshot graph**, which shows the displacement throughout the medium at a single point in time:



The wave can also be represented with a **history graph**, which shows the displacement over time at a single point within the medium:



If the wave shape is constant over time, each graph will contain a reversed image of the other. The image will be scaled horizontally according to the wave's speed, and translated according to the time or position at which each graph is fixed:



If Δy is the displacement change over a short interval in the snapshot graph, the average slope over that interval will be $\Delta y/\Delta x$. Because the graphs are *reversed* relative to each other, a displacement *increase* in one will be matched by a *decrease* in the other, thus reversing the sign of the slope in the history graph. Since $\Delta x = v\Delta t$, this allows:

$$\frac{\Delta y}{\Delta x} = \frac{-\Delta y}{v\Delta t}$$

so the slopes at corresponding points in the two graphs are related by:

$$\frac{\partial y}{\partial x} = -\frac{1}{v}\frac{\partial y}{\partial t} \qquad \frac{\partial y}{\partial t} = -v\frac{\partial y}{\partial x}$$

Neither graph describes the wave throughout space *and* time; though the snapshot captures the wave at every point in space, it does so only at a particular *time*, and though the history quantifies the wave at every point in time, it does so only at a particular *position*. For a complete definition, it is necessary to define the wave's **displacement** D as a function of both x and t.

If the wave shape is constant, and if it moves at constant speed v, then shifting the waveform at time t to the left by vt yields the waveform as it stood at time zero. Conversely, for any t, the displacement at position x matches the displacement that was vt to the *left* of x when t was zero. Therefore, *any* constant traveling wave must be a function of a single expression, $x - vt$. When t is fixed, the resulting function of x produces a snapshot graph. When x is fixed, the resulting function of $-vt$ yields a history graph, with the change in direction being produced by the negative sign, and the horizontal scaling by v.

14.1 Sinusoidal waves

The **wavelength** of a periodic waveform:

$$\lambda = vT$$

is the distance traveled during one period. Similarly:

$$v = \lambda f$$

As λ increases, greater distances are traveled during each period, and as f increases, those distances are traversed more frequently.

Simple harmonic motion produces sinusoidal waves. In a history graph, the peaks of a periodic waveform are one

period apart, while in a snapshot graph, the peaks are one wavelength apart. To produce this type of periodicity, the snapshot graph must be a function of x/λ. Given amplitude A and starting phase ϕ_0:

$$D(x, t=0) = A\sin\left(2\pi\frac{x}{\lambda} + \phi_0\right)$$

$x/\lambda = xf/v$, so as f increases, the function oscillates more frequently; as v increases, each oscillation stretches over a greater distance. Since:

$$\begin{aligned} D(x+\lambda, t=0) &= A\sin\left(2\pi\frac{x+\lambda}{\lambda} + \phi_0\right) \\ &= D(x, t=0) \end{aligned}$$

the function is periodic over λ, as expected.

Replacing x with $x - vt$ releases the time constraint, producing a traveling wave:

$$\begin{aligned} D(x,t) &= A\sin\left(2\pi\frac{x-vt}{\lambda} + \phi_0\right) \\ &= A\sin\left(2\pi\left[\frac{x}{\lambda} - \frac{t}{T}\right] + \phi_0\right) \end{aligned}$$

The position x is divided by the wavelength just as the time t is divided by the period, so this function is periodic in space over λ, and in time over T.

Just as the angular frequency gives the number of radians traversed per unit of *time*:

$$\omega = 2\pi f = \frac{2\pi}{T}$$

the **wave number** gives the radians traversed per unit of *distance*:

$$k = \frac{2\pi}{\lambda}$$

Although k is also used to represent the spring constant, the two values are unrelated.

Because $\lambda = 2\pi/k$ and $f = \omega/2\pi$:

$$v = \lambda f = \frac{2\pi}{k}\cdot\frac{\omega}{2\pi} = \frac{\omega}{k}$$

and:

$$\omega = vk$$

so that radians per time is equal to distance per time multiplied by radians per distance.

Because $2\pi/\lambda = k$ and $2\pi/T = \omega$:

$$D(x,t) = A\sin(kx - \omega t + \phi_0)$$

The **phase** of the wave:

$$\phi = kx - \omega t + \phi_0$$

so the displacement function can also be written:

$$D(x,t) = A\sin\phi$$

The wave speed is the rate at which a given peak or trough travels through space. Such a point has a constant *displacement* as it follows the wave, so it must also have a constant *phase* as it follows the wave. If ϕ is constant, its time derivative:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = k\frac{\mathrm{d}x}{\mathrm{d}t} - \omega = 0$$

From this it is again seen that the speed:

$$v = \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\omega}{k}$$

At time t, the **phase difference** between points x_A and x_B:

$$\begin{aligned} \Delta\phi &= (kx_\mathrm{B} - \omega t + \phi_0) - (kx_\mathrm{A} - \omega t + \phi_0) = k\Delta x \\ &= 2\pi\frac{\Delta x}{\lambda} \end{aligned}$$

This follows from the fact that k is the number of radians per unit of distance, while $\Delta x/\lambda$ is the number of cycles within Δx.

14.2 Wave speed in strings

The waves on a string produce transverse sinusoidal motion. If the displacement occurs in the y dimension, the transverse velocity of a particle at position x:

$$v_y = \frac{\partial y}{\partial t} = -\omega A\cos(kx - \omega t + \phi_0)$$

By extension, the acceleration:

$$a_y = \frac{\partial v_y}{\partial t} = -\omega^2 A\sin(kx - \omega t + \phi_0)$$

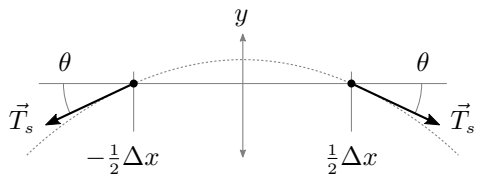
The acceleration is strongest at the peak or trough of each cycle. At the peak:

$$a_{y:\mathrm{P}} = -\omega^2 A = -v_s^2 k^2 A$$

If Δx is a very short length of string centered around the peak of one cycle, the net force on that length:

$$F_{y:\mathrm{P}} \approx ma_{y:\mathrm{P}} = \mu\Delta x\cdot a_{y:\mathrm{P}} = -\mu\Delta x\cdot v_s^2 k^2 A$$

Excepting gravity or drag, the only force affecting any segment is the tension force within the string. If the angle between the horizontal axis and the tension vectors at either end is θ:



then this force is equal to the sum of the y components of the tension vectors:

$$F_{y:\mathrm{P}} = 2T_s \sin\theta$$

Because Δx is small, θ is also small. One of the small angle approximations allows $\sin u \approx \tan u$ when $u \ll 1$, so:

$$F_{y:\mathrm{P}} \approx 2T_s \tan\theta$$

If the peak is centered around the y-axis, then its displacement is greatest there, and:

$$y_{\mathrm{P}} = A\cos(kx)$$

$\tan\theta$ gives the slope at $\frac{1}{2}\Delta x$, yet the slope is also equal to:

$$\frac{\mathrm{d}y_{\mathrm{P}}}{\mathrm{d}x} = -kA\sin(kx)$$

so that:

$$\tan\theta = -kA\sin\left(\frac{k\Delta x}{2}\right)$$

Because Δx is small, $k\Delta x/2$ is also small. Another small angle approximation allows $\sin u \approx u$ when $u \ll 1$, so:

$$\tan\theta \approx -\frac{k^2 A\Delta x}{2}$$

Therefore:

$$F_{y:\mathrm{P}} = -k^2 A\Delta x \cdot T_s$$

Equating this with the earlier result for $F_{y:\mathrm{P}}$ gives:

$$-\mu\Delta x \cdot v_s^2 k^2 A = -k^2 A\Delta x \cdot T_s$$

so that:

$$\mu v_s^2 = T_s \quad \text{and} \quad v_s = \sqrt{\frac{T_s}{\mu}}$$

This tells the speed for a sinusoidal wave. Because any waveform can be decomposed into sinusoids, and because the speed depends only on the string tension and the linear density, all components have the same speed. This makes the result valid for a string waveform of any shape.

14.3 Speed of sound

Sound produces longitudinal waves within a fluid. Assume that a wave pulse is traveling from right to left along the x-axis through a flow tube with cross-sectional area A, and that it approaches a tube section of length Δx. If the reference frame is centered on the pulse, then the section is seen to be moving at speed v while the pulse remains stationary. If the pressure within the tube is p, and if the pressure inside the pulse is $p + \Delta p$, then, as the section meets the pulse, force:

$$F = -\Delta p A$$

is exerted on the section. The section's volume:

$$V = A \cdot \Delta x = A \cdot v\Delta t$$

with Δt being the time for the section to cross any point before the pulse. If the fluid has density ρ at equilibrium, the section's mass:

$$m = \rho A v \Delta t$$

The force will produce acceleration Δv that slows the section. Because $F = ma$:

$$-\Delta p A = \rho A v \Delta t \cdot \frac{\Delta v}{\Delta t}$$
$$-\Delta p = \rho v \Delta v$$

so that:

$$\rho v = -\frac{\Delta p}{\Delta v} \quad \text{and} \quad \rho v^2 = -\frac{\Delta p}{\Delta v/v}$$

If the tube is divided into a number of like sections, each will contain the same mass of fluid. Though the wave disturbs the fluid, it does not permanently displace it, so, in the aggregate, and relative to the pulse, it must travel at a constant mass flow rate:

$$v_0 \rho_0 A = v_1 \rho_1 A$$

As the force accelerates the section, it also compresses it, producing volume change ΔV. The original volume $V = Av\Delta t$. Assuming:

$$\Delta V = A\Delta v \Delta t$$

relates the acceleration to the volume change in a way that maintains the mass flow rate. Therefore:

$$\frac{\Delta V}{V} = \frac{A\Delta t \cdot \Delta v}{A\Delta t \cdot v} = \frac{\Delta v}{v}$$

and:

$$\rho v^2 = -\frac{\Delta p}{\Delta V/V} = B$$

B is the fluid's bulk modulus, which in this case relates the pressure difference within the pulse to the proportional decrease in the section volume. Any compression or rarefaction produced by the pulse is assumed to be adiabatic.

As a result, the speed of sound in a fluid:

$$v = \sqrt{\frac{B}{\rho}}$$

This is analogous to the finding for wave speed in a string, and for simple harmonic motion in general, with B or T_s representing the restoring force that pushes the system toward its equilibrium, and ρ or μ representing the system's ability to store mechanical energy, which carries it past that equilibrium. This result also applies to longitudinal sound waves in a solid. Sound can produce transverse waves in solids as well, but those move at a different speed.

The bulk modulus of a gas varies with temperature. pV^γ is constant in an adiabatic ideal gas process, so:

$$p = \frac{k}{V^\gamma}$$

for some constant k. The derivative of pressure with respect to volume:

$$\frac{\mathrm{d}p}{\mathrm{d}V} = -\gamma k V^{-\gamma-1}$$

The bulk modulus:

$$B = -V\frac{\Delta p}{\Delta V} = -V\frac{\mathrm{d}p}{\mathrm{d}V}$$

so that:

$$B = -V \cdot -\gamma k V^{-\gamma-1} = \gamma \frac{k}{V^\gamma} = \gamma p$$

If n is the number of moles in the volume, and M the molar mass, then:

$$\rho = \frac{nM}{V}$$

Therefore:

$$v = \sqrt{\frac{B}{\rho}} = \sqrt{\gamma p \cdot \frac{V}{nM}}$$

Because $pV = nRT$:

$$v = \sqrt{\frac{\gamma RT}{M}}$$

so that the speed of sound in an ideal gas increases with temperature, and decreases with the molar mass of the gas.

14.4 Wave power and intensity

A wave that spreads outward from a point within a plane is called a **circular wave**. A wave that spreads outward from a point in space is called a **spherical wave**. **Wave fronts** are the regions at which the wave crests. These appear as a set of concentric circles in a circular wave, or as concentric spheres in a spherical wave, each one wavelength apart. At a sufficiently great distance from the wave source, wave fronts resemble parallel lines or planes. In a three-dimensional wave, these are known as **plane waves**. True plane waves do not occur in nature, but they are useful as a simple model.

Because plane waves are identical throughout two of their dimensions, they are described adequately by the basic displacement function $D(x,t)$. To characterize a circular or spherical wave, it is necessary to replace the x variable with r, the straight-line distance to the source. Unlike the wave fronts in a one-dimensional wave (which are points) the fronts in a multidimensional wave increase in size as r increases. Because energy is conserved, the amplitude must decrease over distance, so that A becomes a function of r:

$$D(r,t) = A(r)\sin(kr - \omega t + \phi_0)$$

A wave's power P is the rate at which it transfers energy; its **intensity** I is its power *per unit of area*. If a three-dimensional wave has surface area a at some distance from the source, the intensity over that surface:

$$I = \frac{P}{a}$$

Intensity is measured in $\mathrm{W/m^2}$. If the distance is r, the intensity of a *spherical* wave:

$$I = \frac{P}{4\pi r^2}$$

Because $I_\mathrm{B}/I_\mathrm{A} = r_\mathrm{A}^2/r_\mathrm{B}^2$, the ratio of the intensities at two distances can be found even if the power is unknown.

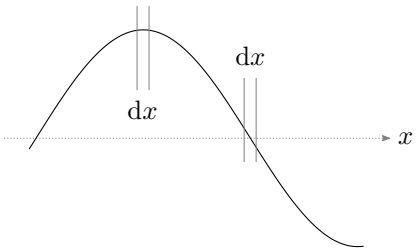
A mechanical wave pulse tranfers energy from one part of the medium to another. The particles in the wave oscillate transversely, longitudinally, or both, but within either axis, the motion follows a waveform that can be decomposed into one or more sinusoids. Each sinusoidal path is an example

of simple harmonic motion. As already demonstrated, the mechanical energy of a particle moving this way $E = \frac{1}{2}kA^2$, with k being the spring constant. The wave's power is a measure of the rate at which this energy is transferred, so, in *any* wave, power and intensity vary with the square of the amplitude.

More specifically, in a string carrying a transverse sinusoidal wave, the transverse velocity of any short segment:

$$v_y = \frac{\partial y}{\partial t} = -\omega A \cos(kx - \omega t + \phi_0)$$

This velocity is momentarily zero at each peak or trough, and it is greatest where the string crosses the x-axis. If the segment spans length $\mathrm{d}x$ when projected onto the x-axis, it can be seen to approximate the unstretched, equilibrium length at every peak or trough, while being stretched to its greatest extent where it crosses the axis:



This shows that the elastic potential energy is also zero at peaks and troughs, and that it reaches its maximum value near the axis.

The total mass of the segment varies with its length, so that $\mathrm{d}m = \mu\,\mathrm{d}x$ for linear density μ. Therefore, the kinetic energy in the segment:

$$\begin{aligned}\mathrm{d}K &= \frac{1}{2}\,\mathrm{d}m \cdot v_y^2 \\ &= \frac{1}{2}\mu\,\mathrm{d}x \cdot \omega^2 A^2 \cos^2(kx - \omega t + \phi_0)\end{aligned}$$

Dividing by $\mathrm{d}t$ gives the rate at which kinetic energy enters and leaves the segment. Because $\mathrm{d}x/\mathrm{d}t$ is the wave speed v:

$$\frac{\mathrm{d}K}{\mathrm{d}t} = \frac{1}{2}\mu v \omega^2 A^2 \cos^2(kx - \omega t + \phi_0)$$

Over a whole number of cycles, the average of the square of cosine is one-half. Therefore:

$$\overline{\left(\frac{\mathrm{d}K}{\mathrm{d}t}\right)} = \frac{1}{4}\mu v \omega^2 A^2$$

The potential energy that passes through the segment over a whole number of cycles must equal the kinetic energy, so that the *average power*:

$$\overline{P} = 2\overline{\left(\frac{\mathrm{d}K}{\mathrm{d}t}\right)} = \frac{1}{2}\mu v \omega^2 A^2$$

Note that the *instantaneous* power varies with the phase of the wave.

As will be seen, the string's impedance $Z = \sqrt{\mu T}$. Because $v = \sqrt{T/\mu}$, $\mu v = Z$, allowing the power to be expressed in terms of impedance:

$$\overline{P} = \frac{1}{2} Z \omega^2 A^2$$

14.5 Impedance

When a traveling wave encounters a *boundary* in the medium, it may be reflected, or part of the wave's energy may be passed through the boundary, leaving the rest to be reflected. Depending on the type of the wave, the reflection may also be inverted.

If a string wave meets a perfectly rigid boundary, it will be reflected. The fixed end of the string does not move, so any force exerted by the string must be opposed by an equal and opposite force. If the wave causes the string to pull *up*, this force will pull *down* on the string, and the reflection will be inverted. As will be seen, reflections are produced by changes in *impedance*, which describes a medium's resistance to harmonic motion; in this case, the wave traveling the low-impedance string encounters an infinitely high impedance at the rigid boundary, and is completely reflected. The wave will also be reflected if the region past the boundary has a *lower* impedance. This can be understood by imagining that the string connects to a massless, frictionless ring that follows a pole in the transverse direction. Beyond the ring is a massless string. The ring travels freely, so nothing resists the transverse motion; however, as the string pulls in the *longitudinal* direction, an equal and opposite force pulls back, reflecting the wave from the zero-impedance region *without* inverting it.

Sound is also reflected by impedance changes. A flute can be open at one or both ends. When the sound inside a flute strikes a *closed* end, the high-pressure region at each peak presses against the rigid boundary, producing an equal and opposite force that reflects the wave. Therefore, although the material at the closed end has a *higher* impedance, the wave is *not* inverted. Sound is also reflected at the *open* end, where the impedance decreases. Impedance is partly determined by the medium's bulk modulus, and although the air inside a flute has the same modulus as

that outside, it is surrounded by a rigid tube that prevents high-pressure regions from expanding (except axially) and that similarly impedes the contraction of low-pressure regions. This causes air inside the tube to resist pressure changes relative to the same air outside, thus increasing the impedance. When a high-pressure region is created just outside the flute, it disperses more quickly than it can inside; this creates a low-pressure region that passes back into the flute, reflecting and inverting the wave. So, although *string* waves are inverted when reflected by impedance *increases*, *sound* waves are inverted when reflected by impedance *decreases*.

Impedance can be defined more precisely by considering the transverse force necessary to drive a wave from one end of a string. If T is the string's tension, and if θ is the angle between the string end and the x-axis, the force must equal and oppose the transverse tension component:

$$T_y = T\sin\theta$$

If the wave amplitude is very small, θ will be small as well. The small angle approximations allow $\sin u \approx \tan u$ when $u \ll 1$, so that:

$$T_y \approx T\tan\theta$$

Because $\tan\theta$ is the slope at this point:

$$T_y = T\frac{\partial y}{\partial x}$$

However, as already seen:

$$\frac{\partial y}{\partial x} = -\frac{1}{v}\frac{\partial y}{\partial t}$$

with $\partial y/\partial t$ being the transverse velocity v_y. Therefore:

$$T_y = -\frac{T}{v}v_y$$

Because it varies linearly with v_y, and because it acts to *oppose* that motion, T_y is seen to be a *damping force*. The **impedance**:

$$Z \equiv \frac{T}{v}$$

is the *damping constant* for that force:

$$T_y = -Zv_y$$

As will be seen, in addition to defining the force necessary to *drive* the wave, this defines the force necessary to *absorb* the wave without producing a reflection. Because $v = \sqrt{T/\mu}$, it is also true that:

$$Z = \sqrt{\mu T}$$

To understand the effect of an impedance change, consider the behavior of the string wave at a boundary. The displacement of a traveling wave can be expressed as a function of its phase, $x - vt$, so that:

$$D(x,t) = f(x - vt)$$

If $g(s) = f(-v \cdot s)$, then it can also be expressed as:

$$D(x,t) = g\Big(t - \frac{x}{v}\Big)$$

If a source wave g_S approaches a boundary from the *left*, then the source and the reflection g_R will superpose, while the transmitted wave g_T continues on to the right. The wave speed is determined entirely by the medium, so the reflection will have the same speed as the source, but an opposite direction:

$$D_\mathrm{S}(x,t) = g_\mathrm{S}\Big(t - \frac{x}{v_\mathrm{S}}\Big) \quad D_\mathrm{R}(x,t) = g_\mathrm{R}\Big(t + \frac{x}{v_\mathrm{S}}\Big)$$

Therefore, the combined displacement on the *left*:

$$D_\mathrm{L}(x,t) = g_\mathrm{S}\Big(t - \frac{x}{v_\mathrm{S}}\Big) + g_\mathrm{R}\Big(t + \frac{x}{v_\mathrm{S}}\Big)$$

while that on the *right*:

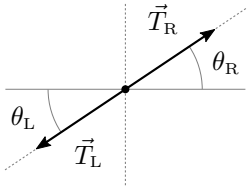
$$D_\mathrm{R}(x,t) = g_\mathrm{T}\Big(t - \frac{x}{v_\mathrm{T}}\Big)$$

Assume the boundary is at $x = 0$. The string is continuous at all points, so $D_\mathrm{L}(0,t) = D_\mathrm{R}(0,t)$ and:

$$g_\mathrm{S}(t) + g_\mathrm{R}(t) = g_\mathrm{T}(t)$$

Note that – by replacing x with zero – a *contingent* statement $D_\mathrm{L}(x,t) = D_\mathrm{R}(x,t)$ that is true only where $x = 0$ is transformed into a *general* statement that is true for *any* argument t. This will be used later to make more general claims about the functions, at points where x is *not* zero.

The point on the boundary has negligible mass; the net transverse force must therefore be zero, or the point would quickly move to a different position that does balance the forces. Assuming θ_L is the angle between the x-axis and the tension $\vec{T}_\mathrm{L}$ on the *left*, and θ_R the corresponding angle on the *right*:



the transverse tension components must be equal:

$$T_{\mathrm{L}} \sin\theta_{\mathrm{L}} = T_{\mathrm{R}} \sin\theta_{\mathrm{R}}$$

If the wave amplitude is again assumed to be very small, the transverse tension can be related to each slope, so that:

$$T_{\mathrm{L}} \frac{\partial D_{\mathrm{L}}(x,t)}{\partial x}\bigg|_{x=0} = T_{\mathrm{R}} \frac{\partial D_{\mathrm{R}}(x,t)}{\partial x}\bigg|_{x=0}$$

Differentiating and replacing x with zero again:

$$-\frac{T_{\mathrm{L}}}{v_{\mathrm{S}}} g'_{\mathrm{S}}(t) + \frac{T_{\mathrm{L}}}{v_{\mathrm{S}}} g'_{\mathrm{R}}(t) = -\frac{T_{\mathrm{R}}}{v_{\mathrm{T}}} g'_{\mathrm{T}}(t)$$
$$-Z_{\mathrm{L}} \cdot g'_{\mathrm{S}}(t) + Z_{\mathrm{L}} \cdot g'_{\mathrm{R}}(t) = -Z_{\mathrm{R}} \cdot g'_{\mathrm{T}}(t)$$

After negating both sides, and after assuming that the displacements (and therefore the integration constants) are zero when $t = 0$, integration produces:

$$Z_{\mathrm{L}} \cdot g_{\mathrm{S}}(t) - Z_{\mathrm{L}} \cdot g_{\mathrm{R}}(t) = Z_{\mathrm{R}} \cdot g_{\mathrm{T}}(t)$$

Combining this with the earlier result leads to:

$$g_{\mathrm{R}}(t) = k_{\mathrm{R}} \cdot g_{\mathrm{S}}(t) \qquad g_{\mathrm{T}}(t) = k_{\mathrm{T}} \cdot g_{\mathrm{S}}(t)$$

with the *reflection* and *transmission coefficients*:

$$k_{\mathrm{R}} = \frac{Z_{\mathrm{L}} - Z_{\mathrm{R}}}{Z_{\mathrm{L}} + Z_{\mathrm{R}}} \qquad k_{\mathrm{T}} = \frac{2Z_{\mathrm{L}}}{Z_{\mathrm{L}} + Z_{\mathrm{R}}}$$

determining the amplitudes of the reflected and transmitted waves in this string. Note that $1 + k_{\mathrm{R}} = k_{\mathrm{T}}$.

Recall that after a perfectly elastic collision, the speed of objects A and B:

$$v_{\mathrm{A:1}} = \frac{m_{\mathrm{A}} - m_{\mathrm{B}}}{m_{\mathrm{A}} + m_{\mathrm{B}}} v_{\mathrm{A:0}} \qquad v_{\mathrm{B:1}} = \frac{2m_{\mathrm{A}}}{m_{\mathrm{A}} + m_{\mathrm{B}}} v_{\mathrm{A:0}}$$

The reflection and transmission coefficients take the same form, and in both cases, the relations conserve energy. The kinetic energy of an object, $K = \frac{1}{2}mv^2$, while the average power of a wave:

$$\overline{P} = \frac{1}{2} Z\omega^2 A^2$$

In the wave, impedance takes the place of *mass*, while amplitude takes that of *velocity*. The range of outcomes is also equivalent to those of an elastic collision:

- If $Z_{\mathrm{L}} \ll Z_{\mathrm{R}}$, the reflection retains most of the source wave's amplitude, while the transmitted wave is very small;
- If $Z_{\mathrm{L}} < Z_{\mathrm{R}}$, the reflected and transmitted waves are both *smaller* than the source;
- If $Z_{\mathrm{L}} = Z_{\mathrm{R}}$, there is no reflection, and the transmitted wave is identical to the source;
- If $Z_{\mathrm{L}} > Z_{\mathrm{R}}$, the reflection is smaller than the source, while the transmitted wave is *larger*;
- If $Z_{\mathrm{L}} \gg Z_{\mathrm{R}}$, the reflection is almost as large as the source, while the transmitted wave has nearly twice the amplitude.

Note that when $Z_{\mathrm{L}} < Z_{\mathrm{R}}$, k_{R} is *negative*, producing an inverted reflection.

Because $g_{\mathrm{R}}(t) = k_{\mathrm{R}} \cdot g_{\mathrm{S}}(t)$ and $g_{\mathrm{T}}(t) = k_{\mathrm{T}} \cdot g_{\mathrm{S}}(t)$ are *general* claims about the g_{S}, g_{R}, and g_{T} functions, t can be replaced by *any* argument, including $t + x/v_{\mathrm{S}}$. This works because the functions do nothing more than map the wave's *phase* (here represented by t) to its displacement. $t + x/v_{\mathrm{S}}$ relates the change in phase to the change in position, and as long as it does that correctly, the phase and displacement relationships between g_{S}, g_{R}, and g_{T} continue to hold. As a result:

$$g_{\mathrm{R}}\Big(t + \frac{x}{v_{\mathrm{S}}}\Big) = k_{\mathrm{R}} \cdot g_{\mathrm{S}}\Big(t + \frac{x}{v_{\mathrm{S}}}\Big)$$

Recalling that $D_{\mathrm{R}}(x,t) = g_{\mathrm{R}}(t + x/v_{\mathrm{S}})$ and $D_{\mathrm{S}}(x,t) = g_{\mathrm{S}}(t - x/v_{\mathrm{S}})$:

$$D_{\mathrm{R}}(x,t) = k_{\mathrm{R}} \cdot D_{\mathrm{S}}(-x,t)$$

When the boundary is placed on the y-axis, the reflection's displacement at point $-x$ is some fraction of the displacement the source *would* have if it reached point x.

A similar operation applies to the g_{T} equation. Replacing t with $t - x/v_{\mathrm{T}}$ produces:

$$g_{\mathrm{T}}\Big(t - \frac{x}{v_t}\Big) = k_{\mathrm{T}} \cdot g_{\mathrm{S}}\Big(t - \frac{x}{v_t}\Big)$$

However, this expression of g_{S} does not match $D_{\mathrm{S}}(x,t) = g_{\mathrm{S}}(t - x/v_{\mathrm{S}})$, which uses v_{S}. Therefore:

$$g_{\mathrm{T}}\Big(t - \frac{x}{v_t}\Big) = k_{\mathrm{T}} \cdot g_{\mathrm{S}}\Big(t - \frac{(v_{\mathrm{S}}/v_{\mathrm{T}})x}{v_{\mathrm{S}}}\Big)$$

so that:

$$D_{\mathrm{T}}(x,t) = k_{\mathrm{T}} \cdot D_{\mathrm{S}}\Big(\frac{v_{\mathrm{S}}}{v_{\mathrm{T}}}x, t\Big)$$

The $v_{\mathrm{S}}/v_{\mathrm{T}}$ term shows the speed of the transmitted wave to be $v_{\mathrm{T}}/v_{\mathrm{S}}$ times that of the source, since displacements that *would* occur at $v_{\mathrm{S}}/v_{\mathrm{T}} \cdot x$ *instead* occur at x. This is to be expected, since $v_{\mathrm{T}}/v_{\mathrm{S}} \cdot v_{\mathrm{S}} = v_{\mathrm{T}}$.

14.6 Light

Light travels at different speeds in different materials. If c is the speed of light in a vacuum, and if v is its speed within some material, then the **index of refraction** or **refractive index** for that material:

$$n = \frac{c}{v}$$

This is the *reciprocal* of the relative speed of light within the material. As will be seen, the refractive index also affects the amount by which light bends as it passes from one medium to another. The refractive index of a vacuum is one, and that of air is very close to one. Liquids and solids have greater refraction indices than gases, so light travels more slowly in those materials. Diamond has the very high refractive index of 2.41. A material's refractive index varies slightly for different colors of light. The index is higher for shorter wavelengths like violet, so those colors travel more slowly through the material.

The wavelength of visible light varies from 400nm to 700nm. Because its speed changes as light enters different materials, and because $v = \lambda f$, either the wavelength or the frequency must change. In a vacuum, a given wave may oscillate at frequency f. As the wave enters a material, this oscillation induces a response that propagates the wave. Because the stimulus is periodic at frequency f, the response must be periodic at that same rate, so the frequency must be the same. Only the wavelength changes.

Inside the material, $\lambda_\mathrm{M} = v/f$. It is also true that $v = c/n$, so:

$$\lambda_\mathrm{M} = \frac{c}{fn}$$

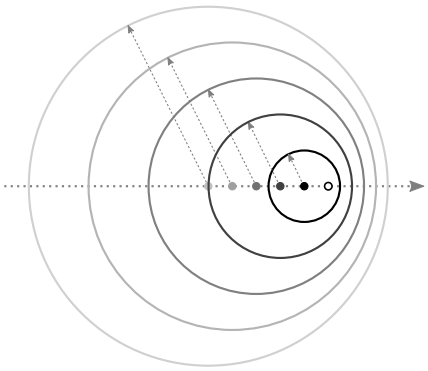
However, the wavelength in a vacuum $\lambda_\mathrm{V} = c/f$, so:

$$\lambda_\mathrm{M} = \frac{\lambda_\mathrm{V}}{n}$$

As the refractive index increases, the speed and the wavelength decrease.

14.7 Doppler Effect

The center of a wave front in a circular or spherical wave is the point at which the front was generated. When the source and the medium are both stationary, concentric fronts are produced. When the source moves, the fronts are spaced unevenly, creating the **Doppler effect**:



Note that the speed of the source does *not* add to that of the wave. A mechanical wave's speed is determined by the bulk modulus and density of the medium, as always.

A stationary source produces wave fronts that are one wavelength apart. If the medium is stationary, and if the source *approaches* an observer at speed v_S, then, in the period T necessary for a *mechanical* wave front to travel λ from its origin, the source has moved $v_\mathrm{S}T$ closer to the observer. The distance between the last front and the next, as measured between the source and the observer:

$$\lambda' = \lambda - v_\mathrm{S}T$$

Because $\lambda = v/f$ and $T = 1/f$:

$$\frac{v}{f'} = \frac{v}{f} - \frac{v_\mathrm{S}}{f}$$
$$vf = (v - v_\mathrm{S})f'$$

so that the observed frequency:

$$f' = \frac{v}{v - v_\mathrm{S}}f = \frac{1}{1 - v_\mathrm{S}/v}f$$

If the observer moves at speed v_O *toward* a *stationary* source, the wave fronts will remain concentric, but their speed relative to the observer will increase to $v' = v + v_\mathrm{O}$. Therefore:

$$f' = \frac{v + v_\mathrm{O}}{\lambda} = \frac{v + v_\mathrm{O}}{v}f = \left(1 + \frac{v_\mathrm{O}}{v}\right)f$$

If the medium remains stationary while the source and the observer move *toward each other*, then, because $v = \lambda f$:

$$v + v_\mathrm{O} = (\lambda - v_\mathrm{S}T)f'$$

Following from this:

$$f' = \frac{v + v_\mathrm{O}}{\lambda - v_\mathrm{S}T} = \frac{v + v_\mathrm{O}}{v/f - v_\mathrm{S}/f} = \frac{v + v_\mathrm{O}}{v - v_\mathrm{S}}f$$

If the source moves *away* from the observer, or if the observer moves *away* from the source, the sign of v_S or v_O is reversed.

The Doppler effect is also observed in electromagnetic waves; however, these waves have no medium, and the speed of light is constant relative to any reference frame. If a light source *approaches* some observer, Einstein's theory of relatively can be used to show that the observed wavelength:

$$\lambda'_{\mathrm{c}} = \sqrt{\frac{c - v_{\mathrm{S}}}{c + v_{\mathrm{S}}}} = \sqrt{\frac{1 - v_{\mathrm{S}}/c}{1 + v_{\mathrm{S}}/c}}$$

14.8 Standing waves

A system is **linear** if it exhibits both *homogeneity* and *additivity*. If the system is modeled by a function $F(x)$, **homogeneity** requires that:

$$F(a \cdot x) = a \cdot F(x)$$

while **additivity** requires:

$$F(x_{\mathrm{A}} + x_{\mathrm{B}}) = F(x_{\mathrm{A}}) + F(x_{\mathrm{B}})$$

In physics, these results are known as the **superposition principle**. When applied to overlapping waves, the principle implies that the displacement at some point is equal to the sum of the displacements that would be produced by each wave individually.

Assume that two linear waves with the same amplitude, frequency, and wavelength are traveling in opposite directions. If their phase constants are both zero, the displacement of their superposition:

$$D(x,t) = a\sin(kx - \omega t) + a\sin(kx + \omega t)$$

Because $\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta$:

$$\begin{aligned} D(x,t) &= a\big[(\sin kx)(\cos\omega t) - (\cos kx)(\sin\omega t)\big] \\ &\quad + a\big[(\sin kx)(\cos\omega t) + (\cos kx)(\sin\omega t)\big] \\ &= 2a(\sin kx)(\cos\omega t) \end{aligned}$$

Though the displacement is periodic over x and t, it is not a function of $x - vt$, so it is not a traveling wave. Instead, the superposition produces a **standing wave**. Where x and t are joined to produce the single phase of a traveling wave, two *separate* phases define the standing wave, one in space, and one in time. The wave's general shape in *space* is defined by the **amplitude function**:

$$A(x) = 2a\sin kx$$

so that:

$$D(x,t) = A(x)\cos\omega t$$

A simple traveling wave appears as a sinusoid that moves along the x-axis. By contrast, a simple standing wave appears as a *stationary* sinusoid that varies over time in amplitude. As $\cos\omega t$ changes in magnitude and sign, the wave drops from its maximum amplitude of $2a$, momentarily disappears, then reappears with its peaks and troughs reversed, these growing until the amplitude reaches its minimum value of $-2a$. The points where $A(x) = 0$ are called **nodes**, and these points never move, even transversely. Nodes occur where $kx = m\pi$ for any integer m. Because $k = 2\pi/\lambda$:

$$x = m\frac{\pi}{k} = m\frac{\lambda}{2}$$

This places a node at every half-multiple of the wavelength. The points halfway between the nodes are called **antinodes**, and they produce the greatest displacement.

If a string is fixed at both ends, any disturbance will create reflections with the same frequency and wavelength as the source. The fixed points impose **boundary conditions** that limit the displacement at those points. If the first point is placed at the origin, and if the string has length L:

$$\begin{aligned} D(0,t) &= 0 \\ D(L,t) &= 0 \end{aligned}$$

To meet the second condition at all times, the amplitude function must produce a node at the end of the string:

$$2a\sin kL = 0$$

Therefore:

$$kL = \frac{2\pi}{\lambda}L = n\pi$$

for some integer $n > 0$. This allows wavelengths:

$$\lambda_n = \frac{2L}{n}$$

and frequencies:

$$f_n = \frac{v}{\lambda_n} = n\frac{v}{2L}$$

The lowest frequency is called the **fundamental frequency**:

$$f_1 = \frac{v}{2L}$$

while the frequencies in general are called **harmonics**:

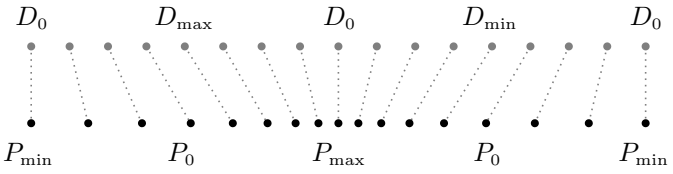
$$f_n = nf_1$$

Note that the fundamental is counted as one of the harmonics, so the 'second' harmonic is $2f_1$, *not* $3f_1$.

The standing wave corresponding to a given n is called a **normal mode** of the string. For each such mode, in a string that is fixed at both ends, n gives the number of antinodes. Note that the fundamental mode produces only *one* antinode, representing *half* a wavelength.

If the string were *open* at one end, there would be nothing to oppose transverse motion at that point. In the idealized example, where a frictionless ring travels a transverse pole, any non-zero slope would produce a transverse force that immediately moved the end to an equilibrium position. Because the peaks and troughs of the standing wave are the only points where the slope is zero, they are the only points where the wave displacement would be stable. Therefore, an open end will produce an *antinode* rather than a node.

Sound produces variations in *displacement* and *pressure*, with the phase of the pressure amplitude $\pi/2$ *behind* that of the displacement. Pressure variations are therefore shifted *forward* in the snapshot graph:



When discussing *sound*, the word 'amplitude' can apply to the *displacement* or the *pressure* of the wave. When it is associated with *displacement*, results equivalent to a string wave are obtained. The closed end of a flute produces a *displacement node*, because the air cannot vibrate longitudinally through the rigid boundary, and a *pressure antinode*, as the boundary alternately compresses and rarefies air that would otherwise have been displaced. Just as the open string end cannot support a non-zero transverse tension, the open flute end cannot maintain a region that varies from the air pressure outside the flute. The only points in the wave that are not compressed or rarefied are the displacement *antinodes*, so an antinode will be found at the open end.

If a string or flute is *open* at both ends, the node at each end will be replaced with an *antinode*. In other respects, the system will match one that is closed at both ends, producing wavelengths $\lambda_n = 2L/n$ and frequencies $f_n = nv/2L$ for every integer $n > 0$. Each mode will contain n half-wavelengths.

If only *one* end is open, the fundamental mode will contain a node and an antinode, these spanning just one *quarter* of a wavelength. Doubling this frequency would produce a mode with *two nodes*, however, which is impossible for this system, so only *odd-numbered* modes will be supported. These will produce wavelengths $\lambda_n = 4L/n$ and frequencies $f_n = nv/4L$ for odd n. Each mode will contain a combined number of $n + 1$ nodes and antinodes.

14.9 Interference

Interference occurs when waves overlap. A standing wave is a special form of interference produced by two same-frequency waves traveling in opposite directions. If two waves with the same amplitude and frequency travel in the *same* direction, and if their sources are x_A and x_B distant from some point, their displacements at that point:

$$D_\mathrm{A}(x_\mathrm{A}, t) = a\sin(kx_\mathrm{A} - \omega t + \phi_{0:\mathrm{A}})$$
$$D_\mathrm{B}(x_\mathrm{B}, t) = a\sin(kx_\mathrm{B} - \omega t + \phi_{0:\mathrm{B}})$$

The distance between the sources:

$$\Delta x = x_\mathrm{B} - x_\mathrm{A}$$

is called the **path-length difference**. The **phase difference** includes the phase change produced by that length, plus the difference in starting phases:

$$\begin{aligned}\Delta\phi &= (kx_\mathrm{B} - \omega t + \phi_{0:\mathrm{B}}) - (kx_\mathrm{A} - \omega t + \phi_{0:\mathrm{A}}) \\ &= k\Delta x + \Delta\phi_0\end{aligned}$$

The waves' superposition:

$$D = D_\mathrm{A} + D_\mathrm{B} = a\sin\phi_\mathrm{A} + a\sin\phi_\mathrm{B}$$

One of the sum-to-product identities allows:

$$\sin\alpha + \sin\beta = 2\cos\Big(\frac{\alpha-\beta}{2}\Big)\sin\Big(\frac{\alpha+\beta}{2}\Big)$$

so that:

$$\begin{aligned}D &= a\cdot 2\cos\left[\frac{-(k\Delta x + \Delta\phi_0)}{2}\right] \\ &\qquad \cdot \sin\left[\frac{(kx_\mathrm{A} + kx_\mathrm{B}) - 2\omega t + (\phi_{0:\mathrm{A}} + \phi_{0:\mathrm{B}})}{2}\right] \\ &= \Big(2a\cos\frac{\Delta\phi}{2}\Big)\sin\big(kx_\mathrm{avg} - \omega t + (\phi_0)_\mathrm{avg}\big)\end{aligned}$$

after exploiting the fact that $\cos -\alpha = \cos\alpha$.

The result is a sinusoid of the same frequency, with path-length and starting phases that are halfway between those

of the sources. Unlike the standing wave, this is a function of $x - vt$, so it is a traveling wave. Its amplitude:

$$A = \left|2a\cos\frac{\Delta\phi}{2}\right|$$

is greatest when the phase difference is an *even* multiple of π. It is *zero* where the difference is an *odd* multiple of π.

If two waves with *different* frequencies have the same amplitude and a starting phase of π, their displacements at $x = 0$:

$$D_{\mathrm{C}}(0,t) = a\sin(-\omega_{\mathrm{C}}t + \pi)$$
$$D_{\mathrm{D}}(0,t) = a\sin(-\omega_{\mathrm{D}}t + \pi)$$

Because $\sin(-\alpha + \pi) = \sin(\alpha)$, their superposition:

$$D = D_{\mathrm{C}} + D_{\mathrm{D}} = a(\sin\omega_{\mathrm{C}}t + \sin\omega_{\mathrm{D}}t)$$

After applying the same sum-to-product identity:

$$D = 2a\cos\left[\frac{1}{2}(\omega_{\mathrm{C}} - \omega_{\mathrm{D}})t\right]\sin\left[\frac{1}{2}(\omega_{\mathrm{C}} + \omega_{\mathrm{D}})t\right]$$

If ω_{C} and ω_{D} are close, $\omega_{\mathrm{mod}} = \frac{1}{2}(\omega_{\mathrm{C}} - \omega_{\mathrm{D}})$ will be small, and D will be perceived as a single amplitude-modulated frequency halfway between the source frequencies:

$$D = \big(2a\cos(\omega_{\mathrm{mod}}t)\big)\sin(\omega_{\mathrm{avg}}t)$$

This pattern is called a **beat**. Because the amplitude crosses zero *twice* as it varies from $-2a$ to $2a$, the beat frequency is equal to *twice* the modulation frequency.

Despite what an observer may see or hear, it is equally valid to understand D as a low frequency ω_{mod} that is modulated at a very fast rate ω_{avg}. Any product of two signals is technically a form of *ring modulation*, and the sum and difference of the modulated frequencies:

$$\omega_{\mathrm{mod}} + \omega_{\mathrm{avg}} = \omega_{\mathrm{C}} \qquad |\omega_{\mathrm{mod}} - \omega_{\mathrm{avg}}| = \omega_{\mathrm{D}}$$

produces the source frequencies, as expected.

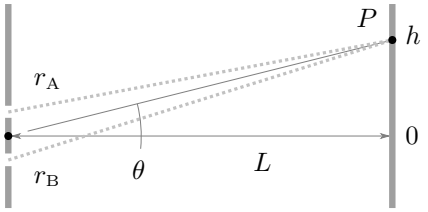
15 Wave optics

An obstacle that blocks *part* of a wave does not cast a clear shadow; instead, the wave fronts *curve* after they pass its edges to fill the shadowed region. This effect is called **diffraction**.

15.1 Double-slit experiment

The *double-slit experiment* uses diffraction to demonstrate the wavelike nature of light. Two slits approximately 100µm wide are cut into a plate some 500µm apart. Even if the wave fronts are flat when they reach the plate, they emerge as two sets of *curved* fronts, as though each slit were a new source. These overlap to produce an **interference fringe**, containing alternating bands of constructive and destructive interference.

Given a screen that is parallel to the plate at horizontal distance L, distances r_{A} and r_{B} will separate any screen point P from the slits. Assume that the y-axis origin is exactly between the slits. If a line is drawn from this point to P, and if the angle from the horizontal axis to the line is θ:



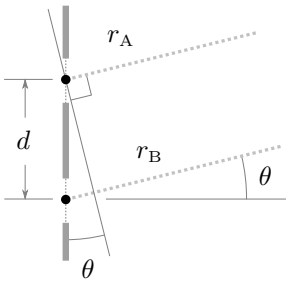
then P will have vertical position:

$$h = L\tan\theta$$

If the light source is placed where $y = 0$, the same wave fronts will enter each slit at the same time, and the waves that exit will have the same phase. Each of these waves will cover every point on the screen, but perfect constructive interference will occur at P only if the path-length difference is an integer multiple of the wavelength, so that:

$$\Delta r = r_{\mathrm{B}} - r_{\mathrm{A}} = m\lambda$$

for integer m. If an arc with radius r_{A} is centered on P, it will cross through the first slit and near the second, with the distance to the second slit being Δr. If vertical distance d between the slits is very small relative to L, this arc will form an essentially straight line. Because d is small, distances r_{A} and r_{B} can be assumed to roughly parallel the line between the midpoint and P, so that both vary by angle θ from the horizontal axis:



This makes the angle from the plate to the arc line θ as well. Naturally, r_A and r_B are *not* parallel, but this assumption allows their length difference to be estimated:

$$\Delta r \approx d \sin\theta$$

so that constructive interference occurs where:

$$d \sin\theta_m = m\lambda$$

If P is near the midpoint, θ will be small, allowing $\sin\theta_m \approx \theta_m$ and:

$$\theta_m \approx m\frac{\lambda}{d}$$

Another small angle approximation gives $\tan\theta_m \approx \theta_m$, so the vertical position of this point on the screen:

$$h_m = L\tan\theta_m \approx L\theta_m = m\frac{\lambda L}{d}$$

The midpoint on the screen is equidistant from each slit. A bright band appears at this point, and at evenly-spaced points above and below it.

The intensity of a wave varies with the square of its amplitude. If I_1 is the intensity at a point on the screen when only *one* slit is open, and if a is the amplitude of this light when it reaches the screen, then $I_1 = ka^2$ for some constant k. In general, when two waves of equal frequency and amplitude are superposed, their combined amplitude:

$$A = \left|2a\cos\frac{\Delta\phi}{2}\right|$$

The phase is the same for both waves as they emerge from the slits, so:

$$\Delta\phi = \frac{2\pi}{\lambda}\Delta r = \frac{2\pi}{\lambda}d\sin\theta \approx \frac{2\pi}{\lambda}d\tan\theta = \frac{2\pi}{\lambda}d\cdot\frac{h}{L}$$

producing:

$$A = \left|2a\cos\Big(\frac{\pi d}{\lambda L}h\Big)\right|$$

for position h on the screen. The intensity of this superposition:

$$I_2 = k\cdot 4a^2\cos^2\Big(\frac{\pi d}{\lambda L}h\Big)$$

Because $I_1 = ka^2$, this can be related to the single-slit intensity:

$$I_2 = 4I_1\cos^2\Big(\frac{\pi d}{\lambda L}h\Big)$$

The intensity of the interference fringe therefore varies from zero to four times that produced by a single slit.

15.2 Diffraction gratings

Diffraction gratings bend different colors of light in different directions, somewhat like a prism. They can be implemented as *reflection* or *transmission* gratings.

A **reflection grating** is created by engraving thousands of thin, parallel grooves in the surface of a mirror. As light is reflected, each ridge acts as a separate light source, producing an interference pattern that reinforces different wavelengths at different angles.

A **transmission grating** produces the same type of interference, but it is made by scoring thousands of thin slits (typically 1000 per millimeter) within a transparent material. If a light source is aligned with the center of the grating, and if it is distant enough to produce something like plane waves, the light emerging from each slit will have the same phase, and an effect much like the double slit experiment will result.

As before, if two adjacent slits are separated by vertical distance d, and if they vary in distance to P by Δr, the paths to the screen can be assumed to be parallel, allowing Δr to be estimated as $d\sin\theta$. The same reasoning applies to all the slits above and below these two. Only at one point is the angle exactly θ, and with each iteration, the slits are d farther from that point. However, if the screen is sufficiently far from the grating, the accumulated error will remain small for some distance along the grating.

Constructive interference will occur where $\Delta r = m\lambda$, so that:

$$\sin\theta_m = m\frac{\lambda}{d}$$

Before, each angle was assumed to be very small, so that $\sin\theta_m \approx \theta_m$. Because d in the grating is much smaller relative to λ, the angles are relatively large, and the small angle

approximations cannot be used. The vertical distance to each point of constructive interference:

$$h_m = L \tan \theta_m$$

The integer m is known as the **order** of the diffraction.

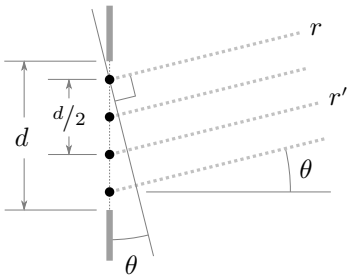
If one slit allows a straight transit from the source to the screen, the slits immediately adjacent can be assumed to be parallel and equidistant, and so on, so that many slits produce phase-consistent paths to the center. Therefore, the *middle* of the screen, where $m = 0$, is a point of constructive interference for *any* wavelength. When a combination of frequencies is directed at the grating, this point displays a bright band that contains each of these components. The same frequencies are scattered at *different angles* above and below, producing a symmetrical pattern that can be used to identify components within the source.

If there are n slits, and if the amplitude at the screen produced by just one of these is a, then the combined amplitude will vary from zero to na. Intensity varies with the square of the amplitude, so if I_1 is the intensity produced by a single slit, the combined intensity must vary from zero to $n^2 I_1$. Because I gives the power per unit *area*, and because energy is conserved, the total area of the bands must decrease as the intensity grows. If it were possible for the waves to combine without interfering, they would cover the screen evenly, and their intensity would be nI_1 rather than $n^2 I_1$. The *width* of the bands therefore varies with $1/n$. As the number of slits increases, the bands grow *brighter* and *thinner*.

15.3 Single-slit diffraction

Huygens' principle is a simple geometric model that predicts diffraction effects in a wave. In this model, each point on a wave front produces a hemispherical *wavelet* that expands only in the 'forward' direction. The next wave front takes the shape of a surface that is *tangent* to all the wavelets.

Light passing through a single slit also produces an interference pattern. If the slit has width d, any point along its length can be associated with another point, also within the slit, that is $d/2$ away:



Each point within some pair generates a hemispherical wavelet that covers the entire screen. For a given point P, one wavelet travels distance r, while the other travels r'. If the angle from the center of the slit to P is θ, and if the paths are again assumed to be parallel, the length difference:

$$\Delta r \approx \frac{d}{2} \sin \theta$$

If this value happens to be $\lambda/2$, the waves in each pair will destructively interfere. Because every path from the slit to P is part of a similar pair, that point will be completely dark.

This approach can be extended to any pair distance that divides the slit into an *even* number of lengths. Given pair distance $d/2n$, for non-zero integer n, destructive interference occurs where:

$$\frac{d}{2n} \sin \theta_n = \frac{\lambda}{2}$$

$$\sin \theta_n = n \frac{\lambda}{d}$$

When the slit is small, θ_n must be larger to produce the necessary path-length difference, which is why smaller apertures produce more pronounced diffraction effects. θ_n also increases as larger n values divide the slit into smaller fractions. This places higher-order minima at the outside of the pattern. Note also that, when the wavelength is *greater* than the slit width, no non-zero integer n can satisfy this equation, and the difference is *never* great enough to produce perfect destructive interference.

If λ is small relative to d, θ will be small as well, allowing:

$$\theta_n \approx n \frac{\lambda}{d}$$

Surprisingly, this resembles the findings for *constructive* interference in the double-slit experiment. The difference results from the geometry of the slits and the way that d is defined. In both cases, different point distances along the slit produce different path lengths, which in turn create varying phase relationships between the paths. In fact, *any* phase relationship can be produced if the points along the

slit are chosen freely. In the double slit, d gives the distance between the two *centers*; this produces a whole-wavelength difference, and constructive interference results. In the single slit, d gives the distance between the two *edges*, and even if wavelet sources were negligible in size, these would be the only two points that were d apart, so perfect constructive interference does not occur. In the single slit, the $d/2n$ distances produce destructive interference. The same distances can be found in the double slit, but at least one point in each distance pair is *blocked* by the material between the slits, so perfect destructive interference does not occur.

If $\tan\theta \approx \theta$, the position on the screen:

$$y_n = n\frac{\lambda L}{d}$$

The central maximum spans the distance between the first-order minima at $\pm\lambda L/d$, so its width on the screen is $2\lambda L/d$. This distance grows as the wavelength increases relative to the size of the slit, so narrower slits produce *wider* patterns. Because light wavelengths are very small, the interference fringes produced by objects of mundane scale are too narrow to be seen, and light passing through even a small hole produces a sharp-edged beam. The hole must be as small as one micron to produce strong light diffraction that fills the space beyond the barrier, the way sound fills a room.

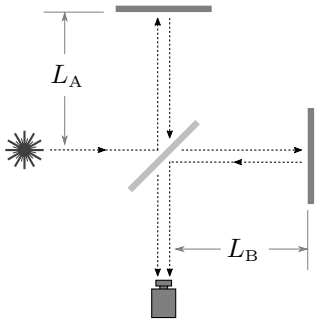
Light passing through a circular aperture creates *Airy's disk*, a round central maximum surrounded by concentric circles. Analyzing this interference pattern is more difficult, but it can be shown that the angle of the first-order minimum:

$$\sin\theta_1 \approx 1.22\frac{\lambda}{d}$$

This gives the central maximum a diameter of approximately $2.44\lambda L/d$.

15.4 Interferometry

Interferometers use interference to make measurements. In the **Michelson interferometer**, light is passed through a beam splitter that divides the light into two roughly orthogonal beams. Each beam encounters a mirror that returns it to the splitter, where the beams divide again. Some of the light is lost, while the rest joined into a single beam that is intercepted by a detector:



Because the beams are generated by the same light source, they start with the same wavelength and phase. They travel the same distance from the source to the splitter, and later from the splitter to the detector, but they travel different distances to and from the mirrors. This creates a phase difference that produces interference at the detector. Depending on the precise geometry of the mirrors, the interference could appear as a pattern of concentric rings or as parallel bands.

If the distance from the splitter to the first mirror is L_A, and that to the second L_B, then the path-length difference:

$$\Delta r = 2L_B - 2L_A$$

Constructive interference will occur at the center of the pattern when Δr is an integer multiple of the wavelength, so that:

$$L_B - L_A = m\frac{\lambda}{2}$$

for integer m.

Reduction gears can be used to make precise changes to one of the distances. As this is done, alternately constructive and destructive interference will occur at the detector, so that:

$$C = \frac{2\Delta L}{\lambda}$$

cycles are observed as the distance changes by ΔL. This can be used to measure the wavelength very precisely, or a known wavelength can be used to measure the change in distance.

This type of interferometer can also be used to measure the refractive index of a gas. If a container of length d is interposed between the splitter and one of the mirrors, one beam will travel distance $2d$ through the container as it passes from the splitter to the mirror and back. If the container is evacuated, and if the wavelength of the light in vacuum is λ_V, then:

$$m_V = \frac{2d}{\lambda_V}$$

wavelengths will be spanned. If gas is slowly added to the container, the refractive index will increase, and the light within will *decrease* in wavelength. When the gas reaches the target pressure, its refractive index will be n, and the light's wavelength λ_M will equal λ_V/n, so that:

$$m_\mathrm{M} = n\frac{2d}{\lambda_\mathrm{V}}$$

wavelengths span the container. Therefore:

$$\Delta m = m_\mathrm{M} - m_\mathrm{V} = (n-1)\frac{2d}{\lambda_\mathrm{V}}$$

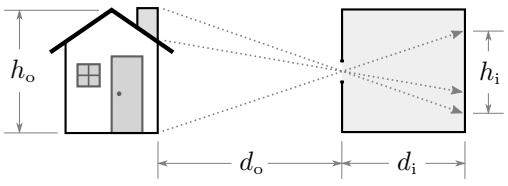
Each increment to Δm represents one wavelength difference relative to the count in the vacuum, this producing one complete cycle of constructive and destructive interference. Counting these cycles at the detector establishes Δm, so that the final refractive index:

$$n = 1 + \frac{\lambda_\mathrm{V}}{2d}\Delta m$$

16 Ray optics

The **ray model** presents light as a collection of *rays*, each moving in a straight line away from the source. When it interacts with matter, a ray may be *scattered*, causing it to change direction, or *absorbed*, so that it stops; otherwise it will continue in the same direction indefinitely. When a ray encounters a *boundary* between materials, it may be *reflected*, or it may be *refracted*, so that it bends in a new direction. This model is used at larger scales, where any apertures traversed by the light are much larger than the light's wavelength. At smaller scales, with apertures of one millimeter or less, diffraction effects must be considered.

The ray model can be used to explain the *camera obscura*, a darkened room with a small aperture that is open to the outside. Rays that emanate from objects outside the room pass through the aperture and illuminate the far wall, producing an image. An object near the ground produces rays that travel *up* to cross the aperture, and these continue at this angle until they reach the wall. In this way, the image is both inverted and flipped left-to-right. The image is also magnified or reduced relative to the size of the object:



Rays emanating from the top and bottom of the object produce one triangle as they travel toward the aperture, and another after they cross the aperture and travel toward the wall. These triangles are similar, so their heights h_i and h_o vary with their widths d_i and d_o in the same way:

$$\frac{h_\mathrm{i}}{d_\mathrm{i}} = \frac{h_\mathrm{o}}{d_\mathrm{o}}$$

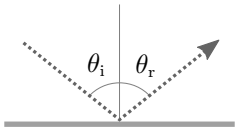
Therefore, the camera's *magnification*:

$$m = \frac{h_\mathrm{i}}{h_\mathrm{o}} = \frac{d_\mathrm{i}}{d_\mathrm{o}}$$

If the aperture were small enough, only a single ray could pass through from any given point on the object. In practice, a *cone* of light passes from each such point, causing the image on the wall to blur. Smaller apertures sharpen the image by passing narrower cones, but they admit less light, making it dimmer as well. If the aperture is small enough, blurring will instead be caused by diffraction.

16.1 Reflection

Very smooth surfaces produce **specular reflection**, like a mirror. For any incident ray, there is a plane containing it that is perpendicular to the reflective surface; when specular reflection occurs, the reflected ray will *also* be found in that plane. Moreover, if the **angle of incidence** θ_i is measured from the incident ray to a line that is normal to the surface, and if the **angle of reflection** θ_r is measured similarly:

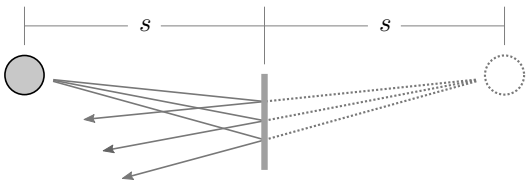


then the ray will reflect so that $\theta_\mathrm{r} = \theta_\mathrm{i}$. Together, these guarantees are known as the **law of reflection**. Unlike refraction, reflection bends light of all wavelengths at the same angle.

At the smallest scale, this law also applies to *rough* surfaces, but irregularities on these cause the normal line and the reflected angle to vary widely from point to point. Because visible light contains wavelengths between 400nm and 700nm, irregularities below one micron will produce specular reflection. Most surfaces, with larger irregularities, produce **diffuse reflection**.

Many rays emanate from each point on an object, and these follow many different paths, allowing the point to be viewed

from different angles. If some of the rays encounter a mirror, they will reflect in a way that produces equal angles of incidence and reflection. If the incident rays were reversed and transposed to the other side of the mirror, they would precisely align with the reflected rays, just as if they and the reflected rays originated from an object on that side:



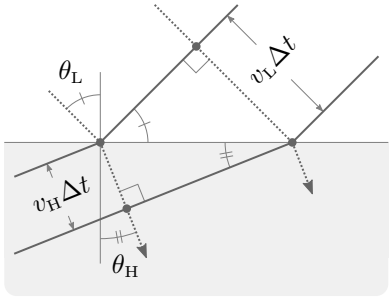
The reflected rays thus produce a **virtual image** of the original that can be viewed from different angles, just like a real object. If the object has distance s from the mirror, the image will appear to be the same distance from the other side.

16.2 Refraction

When rays encounter a smooth boundary between one transparent medium and another, some of the light may be reflected, and some may be *transmitted* into the new material. If the angle of incidence is non-zero, the transmitted rays will also change direction at the boundary. This is called **refraction**.

As already shown, a material's refractive index $n = c/v$, with v being the speed of light in the material. Each ray is part of a wave front that is perpendicular to the ray, so if it enters the new material at an angle, one side of the front will meet the boundary before the other, and the line of their intersection will sweep across the boundary as the front advances. Wherever the front meets the boundary, a disturbance occurs that transmits the wave to the new medium. If the second material has a higher refractive index, the new front will be slower, yet the intersection line travels at a rate consistent with the higher speed of the first wave. The new front must therefore be *flatter* relative to the boundary so that its intersection line sweeps across at the same rate. This flattening also produces the decrease in wavelength that is expected when the refractive index increases.

The amount of refraction can be determined geometrically. If two wave fronts are crossing the boundary, then two intersection lines will be produced. Viewing the fronts edge-on shows these lines as points. The rays that cross these points produce an irregular quadrilateral that can be divided along the boundary into two right triangles, with these sharing a hypotenuse:



If v_L is the speed of light in the first material, and v_H that in the second, then the lengths of the ray sides will be $v_L\Delta t$ and $v_H\Delta t$. The angle of incidence θ_L and the **angle of refraction** θ_H are both measured relative to a line that is normal to the boundary. It can be shown that θ_L and θ_H also occur within the triangles, so the side lengths:

$$v_L\Delta t = h\sin\theta_L \qquad v_H\Delta t = h\sin\theta_H$$

for hypotenuse length h. After solving both for h:

$$\frac{v_L\Delta t}{\sin\theta_L} = \frac{v_H\Delta t}{\sin\theta_H}$$

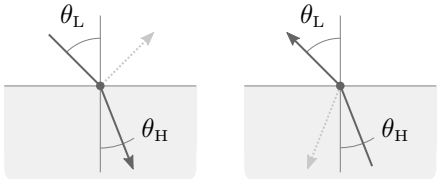
$$\frac{1}{v_L}\sin\theta_L = \frac{1}{v_H}\sin\theta_H$$

and multiplying by c:

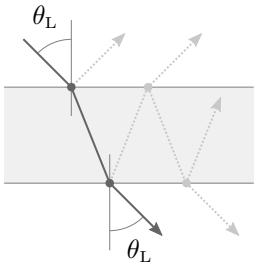
$$n_L\sin\theta_L = n_H\sin\theta_H$$

This is called **Snell's law**, or the **law of refraction**. Note that *both* refraction indices affect the refraction angle, not just that of the second material. When the ray enters a material with a *higher* refractive index, it bends *toward* the normal, but when the refractive index *decreases*, the ray bends *away* from it. The refractive index varies somewhat with the wavelength of the light, with shorter wavelengths having higher indices, and therefore moving more slowly in a given material. The relationship between wavelength and refractive index is called **dispersion**, and this is what causes a prism to split white light into a range of colors.

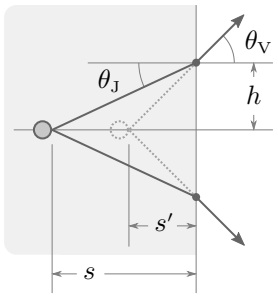
Snell's law treats the incident and refraction angles the same way, so reversing the ray will cause it to refract in a way the returns it to the original angle:



Therefore, a ray passing through a sheet with parallel faces will emerge at its original angle, though it will be *displaced* in a direction that follows the sheet. If the sheet has a *lower* refractive index, the ray will be shifted *forward*; if the sheet has a *higher* refractive index, the ray will be shifted *back*:



When an object in one material is viewed from outside that material, the surface between them acts as a lens, bending rays that emanate from the object and changing the angles at which they are perceived. This produces a virtual image that is closer to or farther from the viewer than the object really is:



The **optical axis** is a line passing through an optical system, about which the system has rotational symmetry; in this case, a line perpendicular to the boundary. Assume the axis passes through the object and the viewer. If a cone of rays leave the object at angle θ_J, the incident angle at the boundary will also be θ_J. If the refraction angle is θ_V, then the virtual image appears where that same angle would have been produced without refraction.

The distance from an object to the center of a lens (in this case, the boundary) is called the **object distance** s. The distance from the lens center to the image is called the **image distance** s'. By convention, virtual images are associated with *negative* image distances. The incident and refracted rays meet at distance h from the optical axis. In both cases, this distance is equal to the slope of the ray multiplied by the real or virtual object's distance from the boundary:

$$h = s\tan\theta_J \qquad h = -s'\tan\theta_V$$

Equating these allows:

$$s' = -\frac{\tan\theta_J}{\tan\theta_V}s$$

Rays that are almost parallel to the optical axis are called **paraxial rays**. Because the pupil is very small relative to an ordinary viewing distance, only paraxial rays reach the retina from the object. One of the small angle approximations allows $\tan u \approx \sin u$ when $u \ll 1$, so:

$$s' \approx -\frac{\sin\theta_J}{\sin\theta_V}s$$

But Snell's law requires that:

$$\frac{\sin\theta_J}{\sin\theta_V} = \frac{n_V}{n_J}$$

so the image distance, relative to the object distance:

$$s' = -\frac{n_V}{n_J}s$$

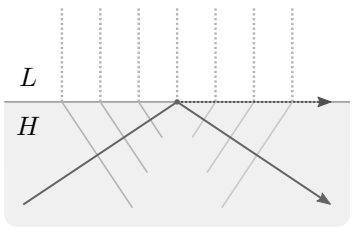
This is true whether the material containing the object has a higher *or* lower refractive index. If the index is *lower*, refraction will bend the ray *toward* the normal, and the object will appear to be *farther* from the surface than it is.

16.3 Total internal reflection

When light passes from material H to material L, the angle of refraction is found by solving Snell's law for θ_L:

$$\theta_L = \sin^{-1}\left(\frac{n_H}{n_L}\sin\theta_H\right)$$

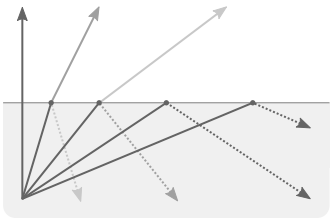
However, that equation has no solution when $n_H/n_L \cdot \sin\theta_H$ is greater than one, or less than negative one. Geometrically, these limits appear because the wavelengths in L are longer than those in H, and yet the wave front edges in both materials must align at the boundary. As the incident angle becomes more oblique, or as the speed of light in H decreases, the incident fronts intersect the boundary at smaller intervals. The refraction fronts must produce the same intervals, and they can be shortened by bending the ray *away* from the normal line so that the fronts become more perpendicular to the boundary. Eventually, however, the refraction angle θ_L reaches 90°, and the intervals can be made no shorter:



The incident angle θ_{H} that produces this is called the **critical angle**:

$$\theta_{\mathrm{c}} = \sin^{-1}\left(\frac{n_{\mathrm{L}}}{n_{\mathrm{H}}}\right)$$

When the incident angle equals or exceeds the critical angle, refraction is no longer possible, and the ray is completely reflected. This is called **total internal reflection**, and it occurs *only* when light passes from a higher refractive index to a lower one. When the incident angle is just below the critical angle, some refraction occurs, but the light is mostly reflected. As the incident angle decreases, more light is refracted, and less is reflected:



Total internal reflection allows fiber optic cable to transmit light impulses over hundreds of kilometers with little intensity loss.

16.4 Scattering

Light may be scattered by dust or droplets suspended in a transparent medium. Because most such particles are large relative to the wavelength of light, the light is typically reflected; if the particles are not colored, this produces a white haze in the medium, as in fog or clouds.

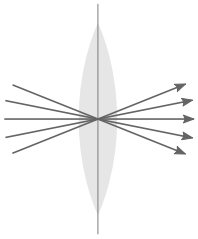
When light is scattered by particles much smaller than its wavelength, **Rayleigh scattering** results. The sun produces a broad range of wavelengths, and when viewed from space, it is white. When the sun is high, molecules in the atmosphere cause Rayleigh scattering, but this is most likely to affect smaller wavelengths. The sky appears blue because short wavelengths are scattered horizontally before reaching the ground. The sun itself looks somewhat yellow because the blue and purple components have been partly lost. When the sun is very low, its light travels much farther through the atmosphere, and most of its blue light is lost before it reaches the viewer. As a result, the sun appears red, and any scattering that does occur at the end must affect the remaining wavelengths, producing a red or orange sky.

16.5 Thin lenses

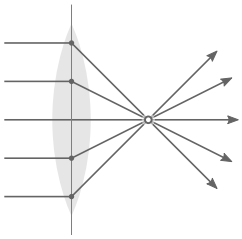
A **focal point** is one toward which rays that are *parallel* to the optical axis are made to *converge*, or one from which they are made to *diverge*. The **focal length** is the distance from the lens surface to a focal point. A **focal plane** is one that is parallel to the lens plane, and that contains a focal point.

Every lens has a **lens plane** that is centered between its faces and perpendicular to the optical axis. A lens uses refraction to bend light rays, and this occurs at the lens *surface*, which is some non-zero distance from the plane. A **thin lens** is one that is very thin relative to its focal length, and the object and image distances. In this type of lens, refraction can be assumed to occur *in the lens plane*, which simplifies calculations while introducing relatively little error.

The faces of a thin lens are almost parallel near the center, so rays passing through the center will not be bent. If the lens is assumed to have a zero width, they will not even be displaced, as normally happens when rays traverse a sheet:

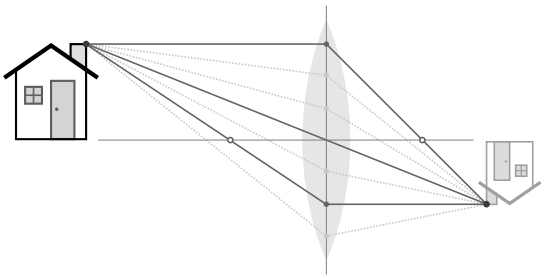


A **converging lens** is thicker in the middle, so it bends incoming parallel rays *toward* its far focal point, on the side that is *away* from the light source:



Such a lens also has a focal point on the *near* side, at the same distance from the plane. If rays radiate from this point, they will be bent to produce *parallel* rays on the far side.

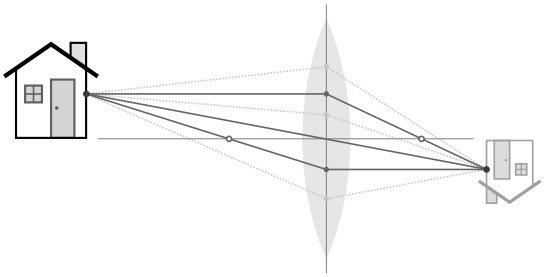
Every point on an object produces rays that travel in many directions. When the object distance is greater than the focal length, a converging lens causes these rays to meet at a point on its far side. This point is called the **real image** of the object point:



The real image can be located by following a number of **special rays** with known geometries:

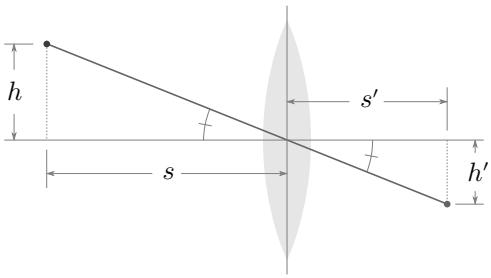
- Because rays that converge on one focal point correspond to parallel rays on the other side, an incoming parallel ray must cross the far focal point;
- Conversely, a ray that crosses the near focal point produces a parallel ray on the far side;
- A ray that crosses the center is unaffected by the lens.

All these rays meet at the real image, and they meet on the side of the optical axis that is below or above the object point. Every point in the **object plane** produces a point in the **image plane**, and object points that are closer to the axis produce image points that are also closer, so the image as a whole is inverted:



For a given object distance, the points in the image plane are the only ones where the rays intersect, and thus the only ones that focus the image perfectly. If a screen is placed in the image plane, a clear representation of the object will be visible. Increasing the focal length causes the bottom special ray to meet the lens at a steeper angle, and farther from the center, and it causes the top special ray to exit at a flatter angle. Both effects *increase* the image size, and move the image *away* from the lens, and these changes are *linear* with respect to the focal length.

The center-crossing ray produces similar triangles on either side of the lens:



If the object and image distances are s and s', and if the object and image heights are h and h':

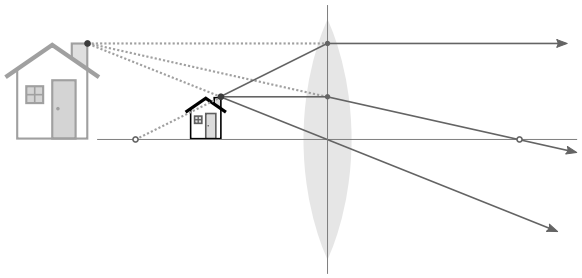
$$\frac{h}{s} = \frac{h'}{s'}$$

The lens magnification relates the image height to the object height. By convention, both sides of the preceding equation are negated, so that the magnification:

$$M = -\frac{s'}{s}$$

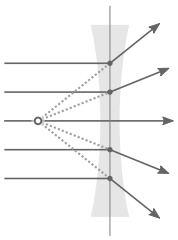
This allows *inverted* images to be represented with negative values. Note that the magnification is associated with a particular object or image distance; it is determined *by* the lens, but it is not a property *of* the lens itself.

When an object is moved *inside* the focal distance of a converging lens, the rays on the other side cease to converge. The special rays can still be used to establish the effect of the lens; in particular, a ray that *would* have passed through the near focal point still produces a parallel ray on the far side:

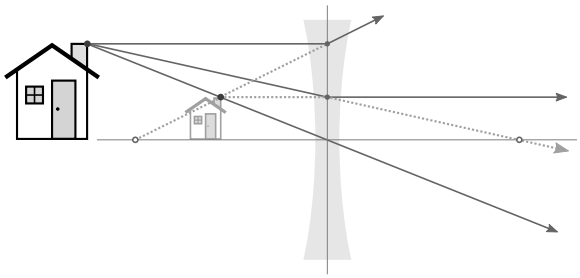


Instead of converging, the rays appear to *diverge* from a point on the near side. This *magnifies* the object by producing an upright virtual image that is *larger* and *more distant* than the original. This image cannot be projected onto a screen, however, without a second lens to focus it. A virtual image is considered to have a *negative* image distance, so the magnification equation produces a positive value, as expected. In this case, increasing the focal length makes the image *smaller* and brings it *closer* to the lens.

A **diverging lens** is *thinner* in the middle, and this causes incoming parallel rays to spread out, as though they diverged from a single point on the near side:



The lens has a similar focal point on the far side, and it produces a virtual image when the object is *outside* the focal length:



The image position is determined much as before:

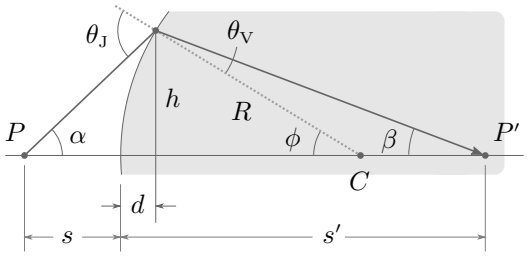
- Incoming parallel rays appear to diverge from the near focal point;
- Rays that *would* cross the far focal point produce parallel rays on the far side;
- Rays that cross the center are unaffected.

Increasing the focal length of a diverging lens makes the image *larger* and moves it *farther* from the lens.

Lenses can be placed in series so the real image projected by one lens becomes the *object* of another. When this is done, the second lens can be analyzed as though the projected image were a physical object.

16.6 Spherical lenses

A lens can be modeled more accurately by understanding that light refracts at the lens *surface*, rather than in the lens plane. Consider a refractive material with a spherical face on one side. If an object is placed at point P in some less-refractive medium, then a ray that emanates from P will bend *toward* the normal when it reaches the boundary; if P is outside the focal point, this will return the ray to the optical axis at point P'. Every line that is normal to the spherical surface crosses the center of the sphere at point C. The angle of incidence θ_J is measured relative to this line, as is the refraction angle θ_V. Because the refraction angle is greater than zero, C must lie between the surface and P':



If the angle between the incident ray and the optical axis is α, and if the angle between the normal and the axis is ϕ, then the third angle in this triangle must equal $\pi - \phi - \alpha$. That angle and θ_J must sum to π, so:

$$\theta_\text{J} = \phi + \alpha$$

Similarly, if the angle between the refracted ray and the axis is β, then the obtuse angle in that triangle must be $\pi - \beta - \theta_\text{V}$. That angle, when added to ϕ, must produce π, so:

$$\theta_\text{V} = \phi - \beta$$

Snell's law requires that $n_\text{J} \sin\theta_\text{J} = n_\text{V} \sin\theta_\text{V}$. If the incident ray is paraxial, a small angle approximation allows $n_\text{J}\theta_\text{J} \approx n_\text{V}\theta_\text{V}$, and:

$$n_\text{J}(\phi + \alpha) = n_\text{V}(\phi - \beta)$$

Both triangles have altitude h, which produces a number of right triangles. If the altitude meets the optical axis at distance d from the surface, if the object and image distances are s and s', and if the sphere has radius R:

$$\tan\phi = \frac{h}{R-d} \qquad \tan\alpha = \frac{h}{s+d} \qquad \tan\beta = \frac{h}{s'-d}$$

The small angle approximations allow $\tan u \approx u$, and the smallness of the angles also implies that d is very near zero. Therefore:

$$\phi \approx \frac{h}{R} \qquad \alpha \approx \frac{h}{s} \qquad \beta \approx \frac{h}{s'}$$

$$n_\text{J}\Big(\frac{h}{R} + \frac{h}{s}\Big) = n_\text{V}\Big(\frac{h}{R} - \frac{h}{s'}\Big)$$

and:

$$\frac{n_\text{J}}{s} + \frac{n_\text{V}}{s'} = \frac{n_\text{V} - n_\text{J}}{R}$$

This relates the image distance to the object distance and the radius. The incident angle is not referenced, so *all* paraxial rays that emanate from object point P converge at real image point P'.

In this example, the surface is convex with respect to the object, and its material is more refractive, but the same equation can be shown to apply to *concave* surfaces, and to those that are *less* refractive. In general, *concave* surfaces must be represented by *negative* radii. In cases where the surface produces a virtual image, s' is also expected to be negative.

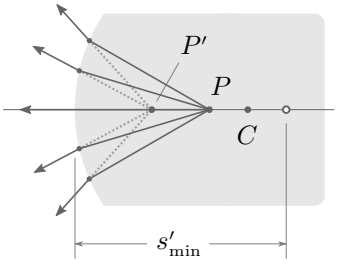
Earlier it was determined that, when viewed from medium V through a flat surface, the image distance of an object within medium J is $-n_V/n_J \cdot s$. As R approaches infinity, the spherical surface becomes flatter, and $(n_V - n_J)/R$ approaches zero. This causes s' to approach $-n_V/n_J \cdot s$, as expected.

$(n_V - n_J)/R$ is constant, so as s' decreases, s must increase. Eventually n_V/s' comes to equal $(n_V - n_J)/R$, and the equation can no longer be satisfied by moving the object away. At that point, the object is infinitely distant, and the incident rays are parallel to the optical axis. The resulting image distance therefore marks a focal point:

$$s'_{\min} = \frac{n_V}{n_V - n_J} R$$

This is the closest the real image can come to the surface.

Reversing a refracted ray causes it to emerge at its original angle, so moving the object to the image position will produce a real image at the original object position. The original refraction angle must be greater than zero, however, so the new object position must be farther from the surface than C. In fact, it must be farther than $s'_{\min}$, because the rays on the other side are already parallel when it is at that point. If the object is placed *between* the surface and $s'_{\min}$, the refracted rays will *diverge*, and a virtual image will be created, as when an object is placed within the focal length of a converging lens:

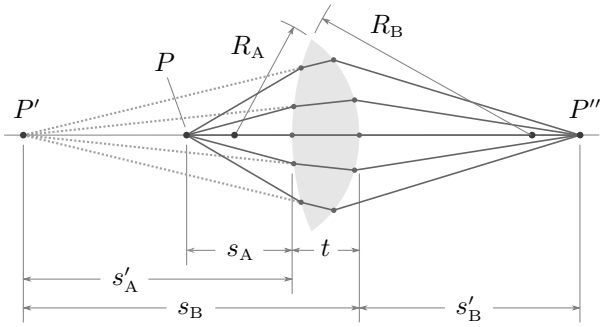


A similar constraint affects s. When n_J/s equals $(n_V - n_J)/R$, the refracted rays become parallel, and s marks another focal point. Decreasing the object distance beyond this point:

$$s_{\min} = \frac{n_J}{n_V - n_J} R$$

causes the rays to diverge from a virtual image.

These findings can be used to understand a lens with two spherical faces, of radius R_A and R_B. Assume that the lens material has refractive index n, while the material around the lens has an index very close to one, as does air. If the object P is inside the focal length of the first surface, its rays will be made to diverge, producing a virtual image P' behind the object:



If the object and image distances are s_A and s'_A:

$$\frac{1}{s_A} + \frac{n}{s'_A} = \frac{n-1}{R_A}$$

The refracted rays approach the second surface just as they would if the virtual image were a real object within material n. The process can therefore be repeated to find the final image at P''.

s'_A is measured relative to the first surface, and it represents a *negative* number. The second image distance s_B spans a forward interval, so if t is the thickness of the lens, then $s_B = t - s'_A$. If this is also considered to be a thin lens:

$$s_B = -s'_A$$

Therefore the second object and image distances are related by:

$$-\frac{n}{s'_A} + \frac{1}{s'_B} = \frac{1-n}{R_B}$$
$$= -\frac{n-1}{R_B}$$

Adding this to the previous result:

$$\frac{1}{s_A} + \frac{1}{s'_B} = \frac{n-1}{R_A} - \frac{n-1}{R_B}$$

If s and s' are the distances for the system as a whole, this produces the **thin lens equation**:

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

with the focal length f being determined by the **lens maker's equation**:

$$\frac{1}{f} = (n-1)\Big[\frac{1}{R_{\mathrm{A}}} - \frac{1}{R_{\mathrm{B}}}\Big]$$

As expected, s approaches the focal length as s' approaches infinity, and vice versa.

In this example, the first surface is convex relative to the object, and the second is concave, but the same equation can be shown to apply to other configurations, if concave radii are again represented with negative numbers. This includes diverging lenses, *meniscus lenses*, which have radii that are both convex or both concave, and even lenses with one flat surface, this being represented by an infinite radius.

16.7 Resolution

Because a material's refractive index increases slightly as wavelengths get shorter, the focal length of a lens is somewhat shorter for violet light than it is for red. The color-specific blurring this produces is called **chromatic aberration**.

When spherical lenses were analyzed in the thin lens equation, the incident rays were assumed to be paraxial so that small angle approximations could be used. Unless the object is very distant relative to the lens diameter, however, larger angles will be found near the outside of the lens, and some rays will be focused on slightly different points. This blurring effect is called **spherical aberration**, and it becomes stronger as the lens diameter increases. Converging lenses and diverging lenses produce offsetting effects, so spherical aberration can be largely canceled by placing these in series.

Light can be modeled as a collection of rays, but it is still a wave. A lens focuses the waves that pass through it, but it also acts as an aperture, diffracting the waves. As already seen, when light passes through a circular aperture, it creates a circular interference pattern called Airy's disk. If a screen is placed in the focal plane, the central maximum in this pattern has width:

$$w \approx 2.44\frac{\lambda f}{d}$$

with d being the diameter of the lens. An object that is very small or very distant would be expected to create a very small image, but w is not affected by the object size, it is determined only by the size of the aperture, the focal length, and the wavelength of the light. w is therefore the **minimum spot size** that can be produced by a particular optical system. It is difficult to produce a lens with a diameter greater than the focal length, so in practice, resolution is limited by the wavelength. Though spherical aberration can be managed by placing lenses in series, or by using non-spherical lenses, this diffraction effect can never be eliminated.

Rayleigh's criterion can be used to determine whether two object points, such as distant stars, can be resolved by a given system. The central maximum in Airy's disk is surrounded by the first-order minimum, a dark ring where destructive interference occurs. The angular displacement of this minimum:

$$\theta_1 \approx 1.22\frac{\lambda}{d}$$

According to Rayleigh's criterion, two objects of equal brightness are resolvable if their angular separation is greater than or equal to θ_1. For this reason, θ_1 is considered to be the **angular resolution** of an optical system.

17 Wave-particle duality and quantization

17.1 Spectroscopy

An **optical spectrometer** splits a single ray of light into distinct spectral components. This can be accomplished by focusing light onto a diffraction grating, which produces constructive interference for specific components at specific angles. The spectrum is then focused onto a detector or a photographic plate.

Different light sources produce different spectrum types. When matter is heated until it glows, it produces a **continuous spectrum** containing smooth intensity variations over a range of wavelengths. This type is produced by the sun or by incandescent lights. When gas is ionized to generate light within a discharge tube, it produces a **discrete spectrum** containing sharp, bright **spectral lines** at specific wavelengths.

Different chemical elements produce different spectra when excited within a discharge tube. Hydrogen produces lines with wavelength:

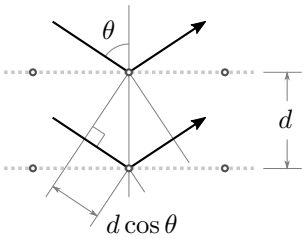
$$\lambda = \frac{91.19\mathrm{nm}}{1/m^2 - 1/n^2}$$

for integers m and n such that $m \geq 1$ and $n > m$. The lines where $m = 2$ are within the visible light range; these are called the **Balmer series**. Within each series, the line spacing progressively decreases as n increases, and as the wavelength approaches the series limit at $\lambda = 91.19\text{nm} \cdot m^2$.

17.2 X-ray diffraction

X-rays have wavelengths between 0.1nm and 10nm. Usually these pass through solids without being absorbed or reflected, but they sometimes interact with an atom, causing a portion of their energy to be radiated as a new spherical wave. The atoms in a crystal are spaced regularly, for the most part. When x-rays are aimed at a crystal, it is possible for these spherical waves to interfere constructively, producing reflections at particular angles. This phenomenon is called **x-ray diffraction**.

If one ray interacts with an atom in any plane of the crystal, and if another ray interacts with an atom directly below it, then constructive interference will occur if the path-length difference between the rays is an integer multiple of the wavelength. The difference is determined geometrically by connecting the rays with two perpendicular lines that pass through the atom in the first plane:

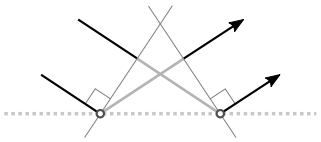


This produces two right triangles with hypotenuse d, equal to the distance between the planes. If the angle of incidence relative to the planes is θ, the adjacent sides will have length $d\cos\theta$, and constructive interference will occur where:

$$2d\cos\theta = m\lambda$$

for integer m. This is called the **Bragg condition**. The interplanar distance must be somewhat larger than half the wavelength to produce the necessary path-length difference. If it is much larger, however, many reflection angles will result, and they will be difficult to resolve.

If another ray interacts with an atom *beside* the first, the same distances will be traveled, and the reflections will again be phase-consistent:



Atoms in different planes may not be directly above or below each other. An imaginary atom in the plane above *would* produce a phase-consistent reflection, however, so the atom below *also* produces one, regardless of its position in the plane. The reflections ultimately depend only on λ, θ, and d.

The atoms in one crystal can be divided into many different families of planes, each oriented to produce a different incident angle θ, and each with a different distance d between the planes. Every such family is capable of producing reflections.

At angles that do *not* meet the Bragg condition, a range of different phases are encountered in different planes, producing strong destructive interference. Any reflections therefore appear as sharp, narrow intensity peaks. This allows a crystal to be used as an *x-ray monochromator*, a device that selects a particular component from a beam containing multiple wavelengths.

17.3 Photon model of light

Though light is a wave, it sometimes behaves like a stream of particles. An image projected by a bright light source shows fine and apparently continuous detail. If light were entirely wavelike, a dimmer but equally detailed image would be produced as the light is attenuated, but instead a speckle pattern emerges. Each dot in this pattern shows the impact of a single **photon**, the particle manifestion of light. According to the **photon theory of light**:

- Light is composed of massless particles called **photons**. In a vacuum, photons travel at the speed of light:

  $$c \approx 3.00 \times 10^8\,\text{m/s}$$

- The energy of a single photon is given by the **Planck-Einstein relation**:

  $$E_\text{P} = hf$$

  with f being the light's frequency, and h the **Planck constant**:

  $$h \approx 6.63 \times 10^{-34}\,\text{Js}$$

- The aggregate behavior of a large number of photons approaches that of a classical wave.

Paradoxically, light's wavelike qualities persist even in its particle manifestation. The double-slit experiment shows the wavelike nature of light by diffracting the output of a single light source at two narrow openings. When the light is relatively intense, smooth gradations are seen in the interference pattern that results. If the light is greatly attenuated, a speckle pattern emerges, yet the dots still conform *on average* to the original interference pattern. This happens even when photons traverse the system *one at a time*, implying that each single photon passes though *both* slits and interferes *with itself* before reaching the screen.

17.4 Matter waves

Just as light waves sometimes resemble particles, particles of *matter* sometimes resemble *waves*. If a beam of electrons is aimed at a crystal, and if all the electrons travel at the same speed, they will reflect at angles consistent with the Bragg condition, for some wavelength. This implies that the electrons have reflected and superposed *as a wave* to produce constructive and destructive interference, just as light does during x-ray diffraction. If the speed of the electrons is changed, the reflection angles will change also, but every angle will again be consistent with the Bragg condition for some wavelength.

It can be shown that the wavelength of these **matter waves**:

$$\lambda = \frac{h}{p}$$

This is called the **de Broglie wavelength**. p is the particle's momentum, so massive or fast-moving particles have smaller wavelengths. Particles or objects of any size can be analyzed as matter waves, but larger objects have wavelengths that are too small to produce wave phenomena in real conditions.

At the quantum scale, the wavelike nature of matter introduces constraints that are not seen in classical systems. Assume that a particle of mass m is bouncing between the walls of a container with length L. This is called the **'particle in a box' model**. If the collisions are perfectly elastic, the particle will bounce indefinitely, and its waveform will overlap itself to produce a standing wave.

As already seen, a standing wave with fixed points at zero and L has wavelength $\lambda_n = 2L/n$, for some integer $n > 0$. From this it follows that:

$$\frac{h}{p_n} = \frac{2L}{n}$$

so that:

$$p_n = \Big(\frac{h}{2L}\Big)n$$

The particle's momentum is therefore *quantized* by increments that grow larger as the container becomes smaller. Because kinetic energy $E = p^2/2m$:

$$E_n = \frac{(h/2L)^2 n^2}{2m} = \frac{h^2}{8mL^2}n^2$$

so the particle's energy is *also* quantized. n is called the **quantum number**, and E_n is the **energy level** of the particle. The energy quanta become smaller as m and L increase.

The standing wave can be no less than half a wavelength long, so n can be no less than one, and the particle's kinetic energy cannot be less than E_1. As a result, the particle can *never be at rest*. The energy level can be expressed relative to this least energetic state:

$$E_n = n^2 E_1$$

18 Electric charge

An atomic nucleus is around $10^{-14}\,\mathrm{m}$ in diameter. Surrounding the nucleus is an **electron cloud** with a diameter of approximately $10^{-10}\,\mathrm{m}$. Electrons are often said to 'orbit' the nucleus, but wave-particle duality prevents them from following specific trajectories over such a small distance.

Electric charge is a property of *protons* and *electrons*, and is found nowhere else. The proton has a mass of $1.67 \times 10^{-27}\,\mathrm{kg}$, and it carries a positive charge, known as the **fundamental unit of charge**:

$$\mathrm{e} = 1.60 \times 10^{-19}\,\mathrm{C}$$

The electron has a much smaller mass of $9.11 \times 10^{-31}\,\mathrm{kg}$, but it carries an equal negative charge, $-\mathrm{e}$. As will be seen, every charge produces *electrostatic forces* that repel like charges and attract opposite ones. A charge's effect on other charges is described by its *electric field*.

An object's charge is determined solely by the number of protons N_p and electrons N_e it contains:

$$q = (N_\mathrm{p} - N_\mathrm{e})\mathrm{e}$$

Charge is therefore *quantized*. An object with no net charge is **electrically neutral**.

Protons are not readily added to or removed from nuclei, so objects become charged by gaining or losing electrons. The gain or loss of electrons by a single atom is called **ionization**. The **law of conservation of charge** holds that charge cannot be created or destroyed, it can only be *transferred* between objects.

The electrons in the outer shell of an atom are called **valence electrons**. In metals, valence electrons are loosely bound to their nuclei. This creates a 'sea of electrons' around the positively-charged **ion cores** that contain the nuclei and non-valence electrons. The aggregate movement of charges through a material is called **current**. A charge that physically moves is called a **charge carrier**. In metals, electrons serve as charge carriers, and their free movement makes metals electrically conductive. Ionic solutions are also conductive, but their charge carriers are ions. It is not necessary for any charge carrier to travel the full length of the conductor; a most carriers travel only a short distance, advancing the charge incrementally and causing other carriers to be displaced in turn.

The valence electrons in an *insulator* are tightly bound to their nuclei. An insulator's surface can be rubbed to produce a charge, particularly if either material contains complex organic molecules, which are easily broken to produce *molecular ions*. The charge cannot move, so it remains on the surface, in the area that was rubbed.

Charges propagate very quickly in a conductor, and an isolated conductor soon reaches **electrostatic equilibrium**, where the net force on every charge is zero, and all charges are at rest. If the object contains more electrons than protons, or vice versa, these *excess* charges will move away from each other, and, as the conductor reaches equilibrium, they will spread across its surface.

If a charged object touches an uncharged conductor, the charge will be shared between them, at least partly discharging the first object. A conductor that is **grounded** shares its charge with the entire earth, allowing it to absorb a charge of any practical size. Humid air is a poor conductor, but it will also discharge an object over time.

If a charged object is placed near an uncharged conductor, some opposite charges in the conductor will be drawn *toward* the object, and some like charges will be pushed *away*. This is called **charge polarization**. The object must not *touch* the conductor or it will be discharged. The force that attracts opposite and repels like charges decreases with distance. Because the opposite charges are *closer* to the charged object, this creates a **polarization force** that attracts the conductor, even though it is electrically neutral as a whole.

If a charged object is used to polarize a neutral conductor, and if the opposite side of the conductor is momentarily grounded, the excess charge *on that side* will be discharged. When the object is moved away, the conductor will be left with a net charge *opposite* that of the object. This is called **charge by induction**.

Insulators cannot be polarized in the aggregate because they do not contain mobile charge carriers. Individual atoms *can* be polarized by displacing their electron clouds in one direction and their nuclei in the other. Two opposite and slightly separated charges are called an **electric dipole**, and these also produce a polarization force. Placing a charged object near an insulator creates *induced* dipoles. Water, by contrast, has a molecular structure that acts as a *permanent* dipole. Because electrostatic force decreases with distance, the *near* charges in the dipoles attract slightly more than the *far* charges repel, and the object as a whole is pulled toward the charge.

Like other gases, air is an insulator, but it normally contains small numbers of electrons that have been freed by background radiation. In the presence of a strong electric field, these can accelerate quickly before striking other molecules. In air, if an electron's kinetic energy is $2.0 \times 10^{-18}\,\mathrm{J}$ or more, another electron will be freed by the collision, then both will be accelerated again, potentially yielding an exponential proliferation of charge carriers. This is called **electrical breakdown**, and it allows gases to conduct electricity. When the free electrons rejoin the ionized nuclei, they emit light. Sparks, arcing, and lightning are all examples of electrical breakdown, and discharge tubes use it to produce light. Other electrical breakdown processes can occur in solids and liquids.

18.1 Coulomb's law

A **point charge** represents an idealized charged object; it has charge and mass, but no size. Two objects can be represented with point charges if they are both much smaller than the distance that separates them.

Given two *static* point charges q_A and q_B, separated by distance r, the magnitude of the **electrostatic force** acting on either is given by **Coulomb's law**:

$$F = \frac{K|q_\mathrm{A}||q_\mathrm{B}|}{r^2}$$

K is the **electrostatic constant**:

$$K \approx 8.99 \times 10^9\,\mathrm{Nm^2/C^2}$$

The force is directed along the line that joins the points, with opposite charges *attracting*, and like charges *repelling*. The SI unit of charge is the **coulomb**, C. Though Coulomb's law applies specifically to static charges, it approximates the force between moving charges if their relative speed is much less than the speed of light.

As will be seen, *electric flux* describes the strength of the electric field passing through a surface. A medium's **permittivity** determines the amount of charge needed to produce flux, with lower permittivity values producing greater amounts. The lowest possible permittivity is the **vacuum permittivity**, also called the *permittivity of free space* or the *permittivity constant*:

$$\epsilon_0 = \frac{1}{4\pi K} \approx 8.85 \times 10^{-12}\,\mathrm{C^2/Nm^2}$$

This constant allows Coulomb's law to be expressed as:

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{|q_\mathrm{A}||q_\mathrm{B}|}{r^2}$$

18.2 Electric fields

Though the elements in a charge pair *interact* to produce the electrostatic force, it is common for one to be designated as the **source charge**. Its contribution is represented as an **electric field** that can be combined with another charge to determine the force. If the source is a *point charge* Q at the origin, the field:

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}\,\hat{r}$$

Electric field strength is measured in N/C. Given a second, positive charge q within $\vec{E}$, the force *on* q:

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

If Q is positive, the field will point *away* from the origin, like $\hat{r}$. Because q is also positive, the direction of $\vec{F}$ will match that of $\vec{E}$, and the charges will *repel*, as expected. If exactly one of the charges were negative, the field strength would be negative, and the field would be understood to point *toward* the origin.

Though dipoles are neutral as a whole, they do produce electric fields. Assume that dipole charges q and $-q$ are separated by distance s, and that they lie along the y-axis, with the dipole's center at the origin. If q is above $-q$, a positive charge that is itself *above* q will be repelled by the positive pole more than it is attracted by the negative, and one *below* $-q$ will be attracted more than repelled. Therefore, the field will point *up* wherever the y-axis is greater than $s/2$ or less than $-s/2$. It will point *down* between those points.

When charges combine to produce a field, the field value at any point is equal to the vector sum of the values produced by the various charges. Along the y-axis, each vector has only one spatial component, so the vector sum can be represented with a single number. For points *outside* the dipole, the total field strength:

$$\begin{aligned} E_\mathrm{a:dip} &= \frac{q}{4\pi\epsilon_0}\Big[\frac{1}{(y-\frac{1}{2}s)^2} - \frac{1}{(y+\frac{1}{2}s)^2}\Big] \\ &= \frac{q}{4\pi\epsilon_0}\Big[\frac{2sy}{(y-\frac{1}{2}s)^2(y+\frac{1}{2}s)^2}\Big] \end{aligned}$$

Note that this gives the *length* of the field vector, which by default points *away* from the source charges. The positive and negative signs in the total field strength do *not* correlate directly with the acceleration expected of a charge in this field; in particular, the negative values *below* the dipole do *not* indicate that a positive charge would accelerate in the negative direction. Instead, every positive field strength *above* indicates a vector that points *away*, and every negative strength *below* indicates one that points *toward*. The field and the acceleration therefore point *up* everywhere along the axis.

If the dipole length s is very small relative to the charge distance y, this reduces to:

$$E_\mathrm{a:dip} \approx \frac{1}{4\pi\epsilon_0} \cdot \frac{2qs}{y^3}$$

The **dipole moment** $\vec{p}$ represents the **polarity** or separation of charge in a dipole. It points from the negative charge *to the positive*, and it has magnitude $p = qs$, measured in Cm.

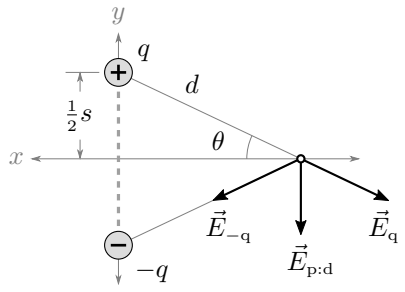
Because the field always points *up*, the direction can be indicated by $\vec{p}$, and the signed value y can be replaced with r, the unsigned distance from the center of the dipole. Therefore, the charge along the dipole axis:

$$\vec{E}_\mathrm{a:dip} = \frac{1}{4\pi\epsilon_0} \cdot \frac{2\vec{p}}{r^3}$$

This would not produce a correct direction or magnitude *between* the charges, but it has already been assumed that r is far outside the dipole. Because each charge is partially canceled by the other, the field strength decreases with the *cube* of the distance, rather than the *square*, as it would for a point charge.

Consider the perpendicular plane that bisects the dipole. If a given point on this plane is assumed to lie along the

x-axis, the field at that point can be divided into x and y components:



The point is equidistant from each charge, so the x components cancel, the y components combine, and the field points down everywhere in the plane.

The point, the origin, and charge q combine to form a right triangle with hypotenuse d. The strength of the q field:

$$E_{\mathrm{q}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{d^2}$$

If θ is the angle between the hypotenuse and the x-axis, the y component of the q field:

$$E_{\mathrm{y:q}} = E_{\mathrm{q}} \sin\theta$$

Because:

$$\sin\theta = \frac{1}{2}s \Big/ d \quad \text{and} \quad d = \sqrt{x^2 + \left(\frac{1}{2}s\right)^2}$$

$E_{\mathrm{y:q}}$ can be expressed in terms of s and x:

$$E_{\mathrm{y:q}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{x^2 + (\frac{1}{2}s)^2} \cdot \frac{\frac{1}{2}s}{\sqrt{x^2 + (\frac{1}{2}s)^2}}$$

Charge $-q$ produces an equal y component, so the dipole field as a whole, throughout the bisecting plane:

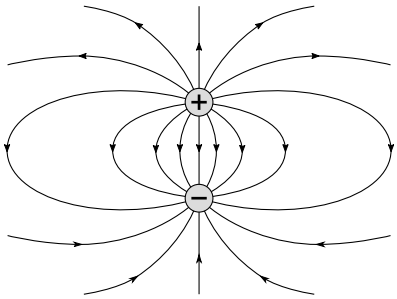
$$E_{\mathrm{p:dip}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{qs}{\left[x^2 + (\frac{1}{2}s)^2\right]^{3/2}}$$

If s is again assumed to be very small relative to x, and if x is replaced with r, the distance to the dipole center, then the charge in the plane that bisects the dipole:

$$\vec{E}_{\mathrm{p:dip}} \approx -\frac{1}{4\pi\epsilon_0} \cdot \frac{\vec{p}}{r^3}$$

The value is negated because $\vec{p}$ points toward the positive charge, yet the field in this plane points down.

A field's structure can be visualized with **field lines** that show its direction and general magnitude within a plane containing the charges:



Each line starts at a charge. The line is drawn by following the field's direction, which leads eventually to another, opposite charge, or away from the charges and to infinity. A tangent to any line shows the field's direction; since this is unique at a given point, the lines can never cross. The lines are marked with arrowheads that point away from the positive charge and toward the negative. The number of lines in an area correlates roughly with the strength of the field.

18.3 Uniform charge distributions

Consider a thin rod bearing charge Q that is *uniformly distributed* over its length. Within the perpendicular plane that bisects the rod, the field can be calculated much as it was for the dipole. If the rod is aligned with the y-axis and centered on the origin, any point in the plane can be assumed to lie along the x-axis. Because the charge is uniform, the rod can be divided into a number of small sections, each carrying charge ΔQ. Every section *above* is matched by another *below* that is the same distance from the plane; because these bear the same charge, their y components cancel, and their x components add. Summing the x component for each section gives the field as a whole.

The section at y_i combines with the origin and the plane point to form a right triangle with hypotenuse d_i. If the section acts as a point charge, and if θ_i is the angle at the point, the section's x field component:

$$E_{x:i} = \frac{1}{4\pi\epsilon_0} \cdot \frac{\Delta Q}{d_i^2} \cos\theta_i$$

To integrate, it is necessary to express the function in terms of y. Because $\cos\theta_i = x/d_i$ and $d_i = \sqrt{x^2 + y_i^2}$:

$$\begin{aligned} E_{x:i} &= \frac{1}{4\pi\epsilon_0} \cdot \frac{\Delta Q}{x^2 + y_i^2} \cdot \frac{x}{\sqrt{x^2 + y_i^2}} \\ &= \frac{1}{4\pi\epsilon_0} \cdot \frac{x\Delta Q}{(x^2 + y_i^2)^{3/2}} \end{aligned}$$

Summing to produce the magnitude of the entire field:

$$E_{\text{p:rod}} = \frac{1}{4\pi\epsilon_0} \sum_i \frac{x\Delta Q}{(x^2 + y_i^2)^{3/2}}$$

ΔQ must also be related to y. Given rod length L, the **linear charge density**:

$$\lambda = \frac{Q}{L}$$

so that $\Delta Q = \lambda \Delta y$. Therefore:

$$\begin{aligned} E_{\text{p:rod}} &= \frac{\lambda}{4\pi\epsilon_0} \sum_i \frac{x\Delta y}{(x^2 + y_i^2)^{3/2}} \\ &= \frac{\lambda}{4\pi\epsilon_0} \int_{-\frac{1}{2}L}^{\frac{1}{2}L} \frac{x}{(x^2 + y^2)^{3/2}} \, \mathrm{d}y \end{aligned}$$

This integral can be solved with trigonometric substitutions. If $u = \arctan(y/x)$, then $y = x \tan u$ and $\mathrm{d}y = x \sec^2 u \, \mathrm{d}u$:

$$\begin{aligned} \int_a^b \frac{x}{(x^2 + y^2)^{3/2}} \, \mathrm{d}y &= \int_{y=a}^{y=b} \frac{x^2 \sec^2 u}{(x^2 + x^2 \tan^2 u)^{3/2}} \, \mathrm{d}u \\ &= \int_{y=a}^{y=b} \frac{x^2 \sec^2 u}{\big((x^2)(1 + \tan^2 u)\big)^{3/2}} \, \mathrm{d}u \\ &= \int_{y=a}^{y=b} \frac{x^2 \sec^2 u}{x^3 \sec^3 u} \, \mathrm{d}u \\ &= \int_{y=a}^{y=b} \frac{1}{x} \cos u \, \mathrm{d}u \\ &= \frac{1}{x} \sin u \Big|_{y=a}^{y=b} \end{aligned}$$

$\arctan(y/x)$ returns the angle at the point, so it turns out that $u = \theta$. Because $\sin\theta = y/\sqrt{x^2 + y^2}$:

$$\begin{aligned} E_{\text{p:rod}} &= \frac{\lambda}{4\pi\epsilon_0} \cdot \frac{y}{x\sqrt{x^2 + y^2}} \Big|_{-\frac{1}{2}L}^{\frac{1}{2}L} \\ &= \frac{\lambda}{4\pi\epsilon_0} \cdot \frac{L}{x\sqrt{x^2 + (\frac{1}{2}L)^2}} \end{aligned}$$

Finally, $\lambda L = Q$, so the field strength in the plane that bisects the rod, at distance r from the rod's center:

$$E_{\text{p:rod}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r\sqrt{r^2 + (\frac{1}{2}L)^2}}$$

This can be used to determine the field around a straight **line of charge** with infinite length:

$$\begin{aligned} E_{\text{line}} &= \frac{1}{4\pi\epsilon_0} \cdot \lim_{L\to\infty} \frac{Q}{r\sqrt{r^2 + (\frac{1}{2}L)^2}} \\ &= \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r \cdot \frac{1}{2}L} \\ &= \frac{\lambda}{2\pi\epsilon_0 r} \end{aligned}$$

This is a good approximation for the field near a straight wire, except near the ends.

Consider the field along the axis passing through a **ring of charge** with radius R. If the ring is centered on the origin within the yz plane, each point on the x-axis has the same distance d from every point on the ring. Because the field produced by a given ring section is balanced by the section opposite, the y and z components cancel. The x component:

$$\begin{aligned} E_{x:i} &= \frac{1}{4\pi\epsilon_0} \cdot \frac{\Delta Q}{d^2} \cos\theta \\ &= \frac{1}{4\pi\epsilon_0} \cdot \frac{\Delta Q}{x^2 + R^2} \cdot \frac{x}{\sqrt{x^2 + R^2}} \end{aligned}$$

so that:

$$E_{\text{a:ring}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{x}{(x^2 + R^2)^{3/2}} \sum \Delta Q$$

ΔQ is constant, so the sections can be summed without integration. At distance r, the axial field strength for the ring as a whole:

$$E_{\text{a:ring}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{rQ}{(r^2 + R^2)^{3/2}}$$

This can be extended to a **disk of charge**. By dividing the disk into a set of rings, each with radius s_i, the axial field strength:

$$E_{\text{a:disk}} = \frac{r}{4\pi\epsilon_0} \sum_i \frac{\Delta Q_i}{(r^2 + s_i^2)^{3/2}}$$

The **surface charge density** of area A:

$$\eta = \frac{Q}{A}$$

so $\Delta Q = \eta \Delta A$. Because $\Delta A = 2\pi s_i \Delta s_i$:

$$\Delta Q = 2\pi\eta s_i \Delta s_i$$

and:

$$E_{\text{a:disk}} = \frac{\eta r}{2\epsilon_0} \sum_i \frac{s_i \Delta s_i}{(r^2 + s_i^2)^{3/2}}$$

$$= \frac{\eta r}{2\epsilon_0} \int_0^R \frac{s}{(r^2+s^2)^{3/2}}\ \mathrm{d}s$$

If $u = r^2 + s^2$ and $\mathrm{d}u = 2s\,\mathrm{d}s$:

$$E_{\text{a:disk}} = \frac{\eta r}{4\epsilon_0} \int_{r^2}^{r^2+R^2} \frac{1}{u^{3/2}}\ \mathrm{d}u$$

$$= -\frac{\eta r}{2\epsilon_0} \cdot \frac{1}{\sqrt{u}}\Big|_{r^2}^{r^2+R^2}$$

$$= -\frac{\eta r}{2\epsilon_0}\Big(\frac{1}{\sqrt{r^2+R^2}} - \frac{1}{r}\Big)$$

Therefore, the field strength of a disk of charge, at distance r along the axis:

$$E_{\text{a:disk}} = \frac{\eta}{2\epsilon_0}\Big(1 - \frac{r}{\sqrt{r^2+R^2}}\Big)$$

It is expected that the field will approximate that of a point charge when r is large relative to the disk radius R. Often this can be verified by calculating the limit of the field strength as r approaches infinity, but in this case, that limit is zero. However, after factoring r from the denominator:

$$E_{\text{a:disk}} = \frac{\eta}{2\epsilon_0}\Big(1 - \frac{1}{\sqrt{1+R^2/r^2}}\Big)$$

The **binomial approximation** allows:

$$(1+x)^\alpha \approx 1 + \alpha x$$

when $|x| < 1$ and $|\alpha x| \ll 1$, so that:

$$E_{\text{a:disk}} \approx \frac{\eta}{2\epsilon_0}\Big(1 - \Big[1 + \big(-\frac{1}{2}\cdot\frac{R^2}{r^2}\big)\Big]\Big)$$

$$= \frac{\eta}{2\epsilon_0}\Big(\frac{1}{2}\cdot\frac{R^2}{r^2}\Big)$$

$$= \frac{\eta R^2}{4\epsilon_0 r^2}$$

when $r \gg R$. $\eta R^2 = Q/\pi$, so the field strength at a great distance:

$$E_{\text{a:disk}} \approx \frac{1}{4\pi\epsilon_0}\cdot\frac{Q}{r^2}$$

as expected.

The field strength for a **plane of charge** is found by letting the disk radius R approach infinity. At any point *outside* the plane:

$$E_{\text{plane}} = \frac{\eta}{2\epsilon_0}$$

Note that the strength is *constant* everywhere outside the plane.

A **parallel-plate capacitor** is constructed from two flat conductive surfaces with a non-conductive material between. As a charge forms on one plate, an equal and opposite charge forms on the other. Because the plates are very close, each can be modeled as a simple plane of charge. Each field radiates in both directions, overlapping inside *and* outside the capacitor. Inside, the fields point in the same direction, from the positive to the negative, so the field strength is twice that produced by a single plane:

$$E_{\text{cap}} = \frac{\eta}{\epsilon_0} = \frac{Q}{\epsilon_0 A}$$

Outside, the fields point in *opposite* directions, *away* from the positive and *toward* the negative. In an idealized capacitor, these cancel; in practice, a **fringe field** is produced, especially near the edges.

18.4 Motion of charged objects

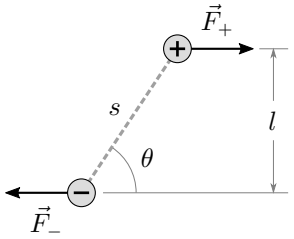
Given a particle with charge q and mass m, its acceleration within electric field $\vec{E}$:

$$\vec{a} = \frac{\vec{F}}{m} = \frac{q}{m}\vec{E}$$

The magnitude of the acceleration is determined by the **charge-to-mass ratio** q/m.

A field with the same direction and magnitude at every point is called a **uniform electric field**. This must be distinguished from a uniform *charge* distribution, which *could* produce a uniform field, but often won't. As expected, a charged particle experiences constant acceleration in a uniform field.

Because there is no gradient, no *net* force is exerted on a dipole in a uniform field, but it *is* subject to a polarizing torque. Equal and opposite forces apply to the charges, so the dipole acts as a force couple. If the dipole's length is s, and if the angle between that length and the lines of action is θ, then the distance between the lines $l = s\sin\theta$:



Therefore, the magnitude of the torque:

$$\tau = lF = s\sin\theta \cdot qE$$

The magnitude of the dipole moment $p = qs$, so:

$$\tau = pE\sin\theta$$

In general, if $\vec{r}$ is the displacement from a pivot to the point of application, then $\vec{\tau} = \vec{r} \times \vec{F}$, so that:

$$\vec{\tau} = q\vec{r} \times \frac{\vec{F}}{q} = \vec{p} \times \vec{E}$$

19 Gauss' law

The **electric flux** Φ is the total strength of the electric field passing through some surface; it is measured in Nm^2/C. For a given field, the flux varies with the area and orientation of the surface; larger surfaces collect more of the field, as do those that are more *perpendicular* to it, since these have larger cross-sectional areas relative to its direction.

A uniform electrical field $\vec{E}$ can be decomposed into two components, one *parallel* to the surface, and one *perpendicular*. The parallel component does not pass through the surface at all. If θ is the angle between $\vec{E}$ and unit vector $\hat{n}$ that is normal to the surface, the perpendicular component:

$$E_\perp = E\cos\theta$$

Therefore, the flux in a *uniform* field:

$$\Phi = EA\cos\theta$$

If the surface is represented by **area vector** $\vec{A} = A\hat{n}$:

$$\Phi = \vec{E} \cdot \vec{A}$$

In a *non-uniform* field, if $\vec{E}_i$ is the strength at a point on the surface, and if $(\delta\vec{A})_i$ is the infinitesimal area at that point, the flux through the point:

$$\Phi_i = \vec{E}_i \cdot (\delta\vec{A})_i$$

The flux through the entire surface is given by the **surface integral**:

$$\Phi = \sum_i \vec{E}_i \cdot (\delta\vec{A})_i = \int \vec{E} \cdot \mathrm{d}\vec{A}$$

A **closed surface** is one that completely divides an *inside* volume from the *outside*. A **Gaussian surface** is a closed surface with an electric field passing through it. The flux through such a surface is given by the **closed surface integral**:

$$\Phi = \oint \vec{E} \cdot \mathrm{d}\vec{A}$$

This is calculated like any other integral; the circle over the integral sign merely indicates that the surface is closed. $\mathrm{d}\vec{A}$ is assumed to point from inside to outside.

If a point charge is centered within a spherical Gaussian surface, its field will be constant in strength everywhere on the surface, and normal to the surface as well. Therefore:

$$\Phi = \oint \vec{E} \cdot \mathrm{d}\vec{A} = E\oint \mathrm{d}A = EA$$

with A being the total area of the surface. If Q is the source charge, and r the sphere's radius:

$$E = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2} \qquad A = 4\pi r^2$$

E decreases as r increases, but A increases by a like amount, so that:

$$\Phi = \oint \vec{E} \cdot \mathrm{d}\vec{A} = \frac{Q}{\epsilon_0}$$

This is called **Gauss' law**.

The flux through a small area is constant for any radius, so the law applies to closed surfaces of *any shape*, since these can always be modeled as a collection of narrow radial sections of varying length. Complex surfaces may require that sections exit and re-enter the volume one or more times. These also conform to Gauss' law, since exiting produces outward-pointing $\mathrm{d}\vec{A}$, entering produces *inward*-pointing $\mathrm{d}\vec{A}$, and each section necessarily exits the volume once more than it enters.

Similarly, charges *outside* a closed surface produce flux at specific intersection points, but they contribute nothing to the *total* flux, since the number of entrances always equals the number of exits. Therefore, the total flux through a closed surface that contains no net charge must be *zero*.

Finally, because the shape of the surface can vary, a charge's location can vary as well. Gauss' law therefore applies to *collections* of charges within the surface, regardless of their distribution, with the total charge being given by Q.

19.1 Symmetric charge distributions

A shape is **symmetric** if it is unchanged after one or more geometric transformations. In particular:

- It has **translation symmetry** along a given axis if it is unchanged after being moved along that axis;
- It has **rotation symmetry** about an axis if it is unchanged after being rotated about that axis;
- It has **reflection symmetry** relative to a plane if it is unchanged after each point is moved to the same distance from the other side of the plane.

A symmetric charge distribution produces a field with the same symmetry. No vector in a symmetric field can have a component that is inconsistent with the field's symmetry; therefore, the field produced by a spherical distribution must point *toward* or *away* from the center, since a tangential component would change direction when reflected.

Because it shows the total flux through a surface, Gauss' law can be used to find the field strength *through each point*, if it can be shown that the field has the same strength everywhere on the surface, and if it is always normal to the surface.

If a charge distribution has *spherical symmetry*, a sphere centered on the distribution will meet both these criteria. Consider a **spherical shell of charge**. For a Gaussian surface *inside* the shell, the total flux must be zero. Because it is required to have the same direction and magnitude everywhere on the surface, *the field must also be zero* everywhere inside a spherical shell.

Note that, although charges *outside* a Gaussian surface do not add to the *total* flux, they *do* produce non-zero field values at specific points if the charge distribution lacks the necessary symmetry. Gauss' law merely requires that such charges produce offsetting values at other points so that the flux as a whole is zero.

More generally, the field produced by a symmetric distribution is determined by dividing the total flux through the surface by its area. The area of a sphere is $4\pi r^2$, so if Q is the charge contained by that sphere:

$$E_{\text{sph}} = \frac{\Phi}{A} = \frac{Q/\epsilon_0}{4\pi r^2}$$

Therefore, the field produced by a **sphere of charge**, or any other distribution with spherical symmetry:

$$\vec{E}_{\text{sph}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r^2}\hat{r}$$

For a given r, this matches the field produced by a point charge, though Q will vary with r if the radius is not outside the distribution as a whole.

Given a *uniform* sphere of charge with radius R, the field *inside* the sphere at distance r varies with the amount of charge contained by r. The spherical shell outside r produces no inside field, as already demonstrated. If Q_r and V_r are the charge and volume contained by radius r, and if Q and V are the charge and volume of the distribution as a whole:

$$\frac{Q_r}{Q} = \frac{V_r}{V} = \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi R^3}$$

so that:

$$Q_r = \frac{r^3}{R^3}Q$$

By Gauss' law:

$$E_{\text{in:sph}} \cdot 4\pi r^2 = \frac{(r^3/R^3)Q}{\epsilon_0}$$

so the field *inside* the uniform sphere of charge:

$$\vec{E}_{\text{in:sph}} = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{R^3} r \cdot \hat{r}$$

As r increases, the charge near the center has less effect, but much more charge is encompassed by the new volume. Therefore, though the strength of this field *outside* the distribution decreases with the square of the distance, it *increases linearly* with distance *inside*.

As already seen, the field produced by a *line of charge* can be derived from Coulomb's law after a difficult integral substitution. With Gauss' law, it is calculated more easily. If the line has linear charge density λ, and if a section of length L is enclosed by a cylinder, the total charge within $Q = \lambda L$. The field is normal to the outside surface and equally strong at all points, while the flat surfaces at the ends are parallel to the field, allowing them to be ignored. Because the outside area $A = 2\pi rL$:

$$E_{\text{line}} = \frac{Q/\epsilon_0}{A} = \frac{\lambda L/\epsilon_0}{2\pi rL} = \frac{\lambda}{2\pi\epsilon_0 r}$$

as expected.

The field produced by a *plane of charge* is found in like manner. If a cylinder passes halfway through the plane, and if its axis is perpendicular to it, a disk of charge will be found within. Given cylinder radius R and surface charge density η, the contained charge $Q = 2\pi R^2\eta$. The ends of

the cylinder are normal to the field, and their total surface area is $4\pi R^2$. The walls of the cylinder are parallel, so they add no flux. The field strength:

$$E_{\text{plane}} = \frac{Q/\epsilon_0}{A} = \frac{2\pi R^2 \eta/\epsilon_0}{4\pi R^2} = \frac{\eta}{2\epsilon_0}$$

as expected.

19.2 Conductors in electrostatic equilibrium

When a conductor reaches electrostatic equilibrium, the field strength everywhere inside drops to zero. Conductors contain an abundance of charge carriers, so if the field *were* non-zero, some charges would move; in particular, if the field strength were *positive*, some positive charges would follow the field, or some negative charges would move into it, and the positive concentration would be dispersed or canceled. In this sense, every electric field directs charges toward an equilibrium that *dissipates the field*, if the charges are mobile. Because there is no field, the flux through all points of a Gaussian surface that is anywhere inside the conductor is also zero, implying that the conductor's interior is electrically neutral. Any excess charge must therefore be found *on the surface*. Furthermore, the electric field on the surface must at every point be directed *outward*, and it must be *normal* to the surface, since any tangential component would produce a current that would rearrange the charges.

The same reasoning applies to fields produced by *external* charges. Such a field must be normal to the surface of the conductor, and it must end at that surface. A metal box can therefore be used to **screen** an external electric field from its interior. Wire cages are also reasonably effective.

Though it is likely to vary at different points, the field strength at the surface can be related to the surface charge density η at each point. If a small cylinder passes halfway through the surface, and if its axis is perpendicular to the surface, a disk of area A and charge $Q = \eta A$ will be enclosed. This resembles the process used to analyze a plane of charge, but in that case, the field pointed outward on *both* sides. This field has been shown to point *outward only*, so the field strength at the surface of the conductor:

$$E_{\text{surf:conduct}} = \frac{\eta A/\epsilon_0}{A} = \frac{\eta}{\epsilon_0}$$

A Measurement

Accuracy describes the proximity of a measurement to the quantity being measured. **Precision** describes the proximity of repeated measurements to each other.

Systematic errors shift measurements in a consistent direction. These affect the accuracy of a measurement, but not its precision. **Random errors** impart no predictable bias. These affect precision, but over repeated trials, they have no effect on accuracy.

The **least count** of a measuring device is the smallest unit gradation offered by that device. When taking measurements, it is customary to record all decimal places up to the least count, plus an estimated digit. This establishes the number of **significant digits** in the measurement.

When reading values, zeroes to the left of the first non-zero digit are not counted as significant digits. If a decimal point is given, all trailing zeros are considered significant. If there is no decimal point, the significance of trailing zeroes is not defined, though they can generally be considered insignificant. Values that are *defined* rather than measured can be considered to have an unlimited number of significant digits.

The result of a calculation should not have more significant digits than the *least precise* measurement used, though it is customary to retain an extra digit if the result begins with one. It is permissible to retain one or two extra digits for intermediate calculations.

SI units take their name from *Le Système International d'Unités*. Common prefixes include:

Factor	Prefix	Symbol
10^{18}	exa	E
10^{15}	peta	P
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{2}	hecto	h
10^{1}	deka	da
.....		
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

B Vectors

By convention, vectors are not allowed to have negative magnitudes; if such a result is produced, the vector is instead made to point in the opposite direction.

In the rectangular coordinate system, a vector $\vec{A}$ may be decomposed into **component vectors** $\vec{A}_x$, $\vec{A}_y$, and $\vec{A}_z$ which are parallel to the axes. These may be described by their **components** A_x, A_y, and A_z, which *can* have negative values.

The **unit vectors** $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ have unit magnitudes that coincide with the positive x, y, and z axes. Therefore:

$$\vec{A} = A_x\hat{\imath} + A_y\hat{\jmath} + A_z\hat{k}$$

B.1 Dot products

If the counterclockwise angle from $\vec{A}$ to $\vec{B}$ is θ, the vectors' **dot product** or **scalar product**:

$$\vec{A} \cdot \vec{B} \equiv AB\cos\theta$$

Since the dot product of a unit vector with itself is one, and the dot product of any vector with an orthogonal vector is zero, this gives the sum of the products of the corresponding components:

$$\begin{aligned}\vec{A} \cdot \vec{B} &= (A_x\hat{\imath} + A_y\hat{\jmath} + A_z\hat{k}) \cdot (B_x\hat{\imath} + B_y\hat{\jmath} + B_z\hat{k}) \\ &= A_xB_x + A_yB_y + A_zB_z\end{aligned}$$

B.2 Cross products

If $\vec{A}$ and $\vec{B}$ are within the xy-plane, their **cross product** or **vector product**:

$$\vec{A} \times \vec{B} \equiv (AB\sin\theta)\hat{k}$$

with $\hat{k}$ pointing *toward* the viewer when θ is visible as a counterclockwise turn from $\vec{A}$ to $\vec{B}$. This product is perpendicular to both vectors and **normal** to the plane containing them. Its magnitude is greatest when A and B are perpendicular to each other, and it is zero when they point in the same or opposite directions. The cross product is *not* commutative:

$$\vec{A} \times \vec{B} \neq \vec{B} \times \vec{A}$$

While $\vec{B} \times \vec{A}$ does have the same magnitude, it points in the opposite direction.

The product rule can be applied to cross products, with:

$$\frac{\mathrm{d}}{\mathrm{d}u}(\vec{A} \times \vec{B}) = \left(\frac{\mathrm{d}\vec{A}}{\mathrm{d}u} \times \vec{B}\right) + \left(\vec{A} \times \frac{\mathrm{d}\vec{B}}{\mathrm{d}u}\right)$$

Sources

A Student's Guide to Waves
Daniel Fleisch, Laura Kinnaman
2016, Cambridge University Press

Fundamentals of Physics
David Halliday, Robert Resnick
1988, John Wiley & Sons

Physics for Scientists and Engineers
Randall D. Knight
2004, Pearson/Addison Wesley

(untitled draft of waves book)
David Morin
http://www.people.fas.harvard.edu/~djmorin/book.html

HyperPhysics
http://hyperphysics.phy-astr.gsu.edu/hbase/hph.html

Physics Stack Exchange
http://physics.stackexchange.com

Wikipedia
http://en.wikipedia.org

This document was typeset with LaTeX. Diagrams were created with Inkscape.