Modern Approach to Introductory Physics
Volume 1: Fundamental Principles

See Calculus.pdf on the website for basic calculus ideas that you will need for this class.

Chapter 1 Waves in One Dimension

The wave is a universal phenomenon which occurs in a multitude of physical contexts.

The purpose of this section is to describe the *kinematics* of waves, i. e., to provide tools for describing the form and motion of all waves irrespective of their underlying physical mechanisms.

Many examples of waves are well known to you.

You undoubtedly know about ocean waves and have probably played with a stretched slinky toy, producing undulations which move rapidly along the slinky.

Other examples of waves are sound, vibrations in solids, and light.

In this chapter we learn first about the basic properties of waves and introduce a special type of wave called the sine wave.

Examples of waves seen in the real world are presented.

We then learn about the superposition principle, which allows us to construct complex wave patterns by superimposing sine waves.

Using these ideas, we discuss the related ideas of beats and interferometry.

Finally, the ideas of wave packets and group velocity are introduced.

1.1 Transverse and Longitudinal Waves

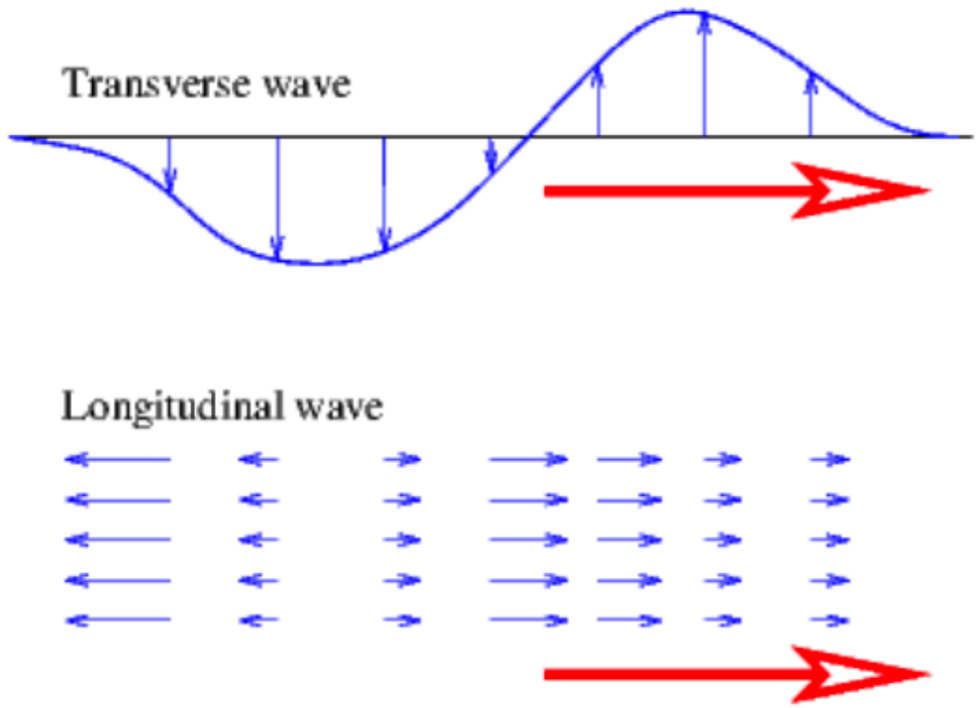


Figure 1.1: Example of displacements in transverse and longitudinal waves. The wave motion is to the right as indicated by the large arrows. The small arrows indicate the displacements at a particular instant.

With the exception of light, waves are undulations in a material medium.

For instance, ocean waves are (nearly) vertical undulations in the position of water parcels.

The oscillations in neighboring parcels are phased such that a *pattern* moves across the ocean surface.

Waves on a slinky are either *transverse*, in that the motion of the material of the slinky is perpendicular to the orientation of the slinky, or they are *longitudinal*, with material motion in the direction of the stretched slinky. (See figure 1.1.)

Some media support only longitudinal waves, others support only transverse waves, while yet others support both types.

Light waves are purely transverse, while sound waves are purely longitudinal.

Ocean waves are a peculiar mixture of transverse and longitudinal, with parcels of water moving in elliptical trajectories as waves pass.

Light is a form of electromagnetic radiation.

The undulations in an electromagnetic wave occur in the electric and magnetic fields.

These oscillations are perpendicular to the direction of motion of the wave (in a vacuum), which is why we call light a transverse wave.

1.2 Sine Waves

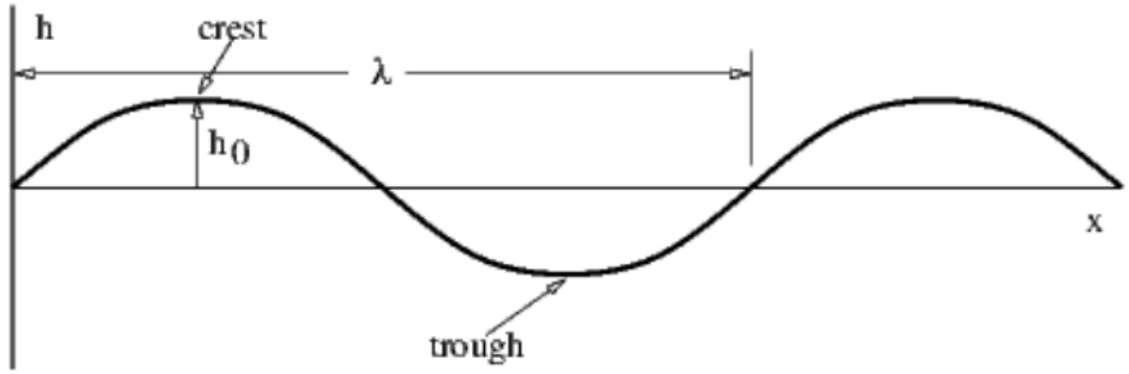


Figure 1.2: Definition sketch for a sine wave, showing the wavelength λ and the amplitude h_0 and he phase φ at various points.

A particularly simple kind of wave, the sine wave, is illustrated in figure 1.2.

This has the mathematical form

$$h(x) = h_0 \sin(2\pi x/\lambda), \tag{1.1}$$

where h is the *displacement* (which can be either longitudinal or transverse), h_0 is the maximum displacement, also called the *amplitude* of the wave, and λ is the *wavelength*.

The oscillatory behavior of the wave is assumed to carry on to infinity in both positive and negative x directions.

Notice that the wavelength is the distance through which the sine function completes one full cycle.

The *crest* and the *trough* of a wave are the locations of the maximum and minimum displacements, as seen in figure 1.2.

So far we have only considered a sine wave as it appears at a particular time.

All interesting waves move with time.

The movement of a sine wave to the right a distance d may be accounted for by replacing x in the above formula by $x - d$.

If this movement occurs in time t, then the wave moves at velocity $c = d/t$.

Solving this for d and substituting yields a formula for the displacement of a sine wave as a function of both distance x and time t:

$$h(x,t) = h_0 \sin[2\pi(x - ct)/\lambda]. \tag{1.2}$$

The time for a wave to move one wavelength is called the *period* of the wave: $T = \lambda/c$.

Thus, we can also write

$$h(x,t) = h_0 \sin[2\pi(x/\lambda - t/T)]. \tag{1.3}$$

Physicists actually like to write the equation for a sine wave in a slightly simpler form.

Defining the wavenumber as k = 2π/λ and the angular frequency as ω = 2π/T, we write

$$h(x,t) = h_0 \sin(kx - \omega t). \tag{1.4}$$

We normally think of the frequency of oscillatory motion as the number of cycles completed per second.

This is called the *rotational frequency*, and is given by $f = 1/T$.

It is related to the angular frequency by $\omega = 2\pi f$.

The rotational frequency is usually easier to measure than the angular frequency, but the angular frequency tends to be used more often in theoretical discussions.

As shown above, converting between the two is not difficult.

Rotational frequency is measured in units of hertz, abbreviated Hz; 1 Hz = 1 cycle s^{-1}.

Angular frequency also has the dimensions of inverse time, e. g., radian s^{-1}, but the term “hertz” is generally reserved only for rotational frequency.

The argument of the sine function is by definition an angle.

We refer to this angle as the *phase* of the wave, $\varphi = kx - \omega t$.

The difference in the phase of a wave at fixed time over a distance of one wavelength is 2π, as is the difference in phase at fixed position over a time interval of one wave period.

Since angles are dimensionless, we normally don't include this in the units for frequency.

However, it sometimes clarifies things to refer to the dimensions of rotational frequency as "rotations per second" or angular frequency as "radians per second".

As previously noted, we call h_0, the maximum displacement of the wave, the amplitude.

Often we are interested in the *intensity* of a wave, which is proportional to the square of the amplitude.

The intensity is often related to the amount of energy being carried by a wave.

The wave speed we have defined above, $c = \lambda / T$, is actually called the *phase speed*.

Since $\lambda = 2\pi/k$ and $T = 2\pi/\omega$, we can write the phase speed in terms of the angular frequency and the wavenumber:

$$c = \frac{\omega}{k} \quad \text{phase speed} \tag{1.5}$$

1.3 Types of Waves

In order to make the above material more concrete, we now examine the characteristics of various types of waves which may be observed in the real world.

1.3.1 Ocean Surface Waves

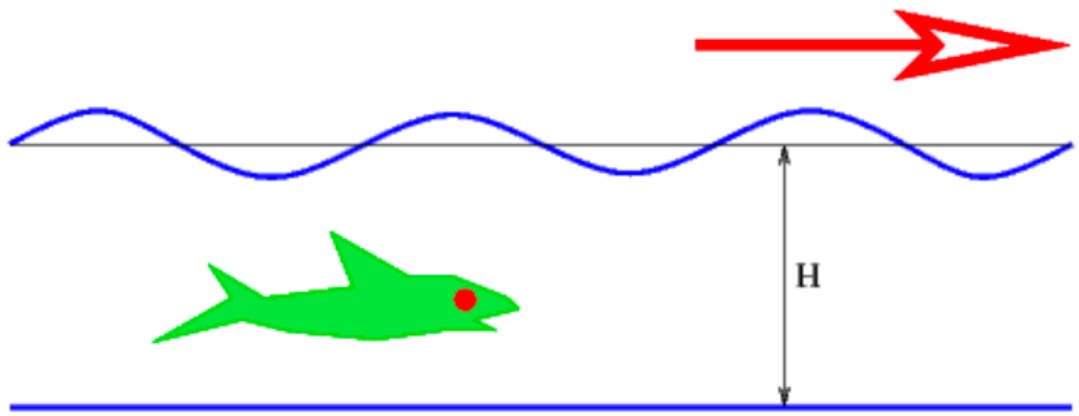


Figure 1.3: Wave on an ocean of depth H. The wave is moving to the right and the particles of water at the surface oscillate in elliptical trajectories as the wave crests and troughs pass.

These waves are manifested as undulations of the ocean surface as seen in figure 1.3.

The speed of ocean waves is given by the formula

$$c = \left(\frac{g \tanh(kH)}{k} \right)^{1/2}, \tag{1.6}$$

where $g = 9.8$ m s^{-2} is the earth's gravitational force per unit mass, H is the depth of the ocean, and the hyperbolic tangent is defined as

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}. \tag{1.7}$$

The equation for the speed of ocean waves comes from the theory for oscillations of a fluid surface in a gravitional field

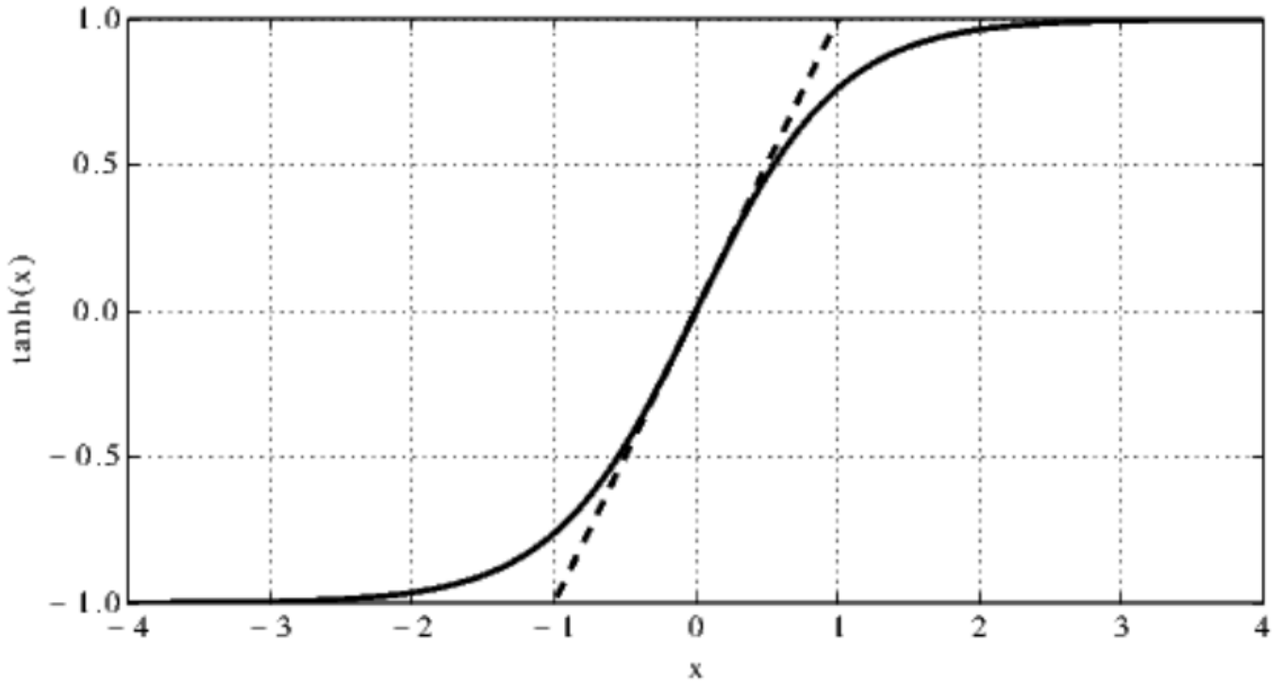


Figure 1.4: Plot of the function tanh(x). The dashed line shows our approximation tanh(x) ≈ x for $|x| \ll 1$.

As figure 1.4 shows, for $|x| \ll 1$, we can approximate the hyperbolic tangent by tanh(x) ≈ x, while for $|x| \gg 1$ it is +1 for $x > 0$ and -1 for $x < 0$.

This leads to two limits: Since $x = kH$, the *shallow water* limit, which occurs when $kH \ll 1$, yields a wave speed of

$$c \approx (gH)^{1/2}, \quad \text{(shallow water waves)}, \tag{1.8}$$

while the *deep water* limit, which occurs when $kH \gg 1$, yields

$$c \approx (g/k)^{1/2}, \quad \text{(deep water waves)}. \tag{1.9}$$

Notice that the speed of shallow water waves depends only on the depth of the water and on g.

In other words, all shallow water waves move at the same speed.

On the other hand, deep water waves of longer wavelength (and hence smaller wavenumber) move more rapidly than those with shorter wavelength.

Waves for which the wave speed varies with wavelength are called *dispersive*.

Thus, deep water waves are dispersive, while shallow water waves are non-dispersive.

For water waves with wavelengths of a few centimeters or less, surface tension becomes important to the dynamics of the waves.

In the deep water case, the wave speed at short wavelengths is given by the formula

$$c = (g/k + Ak)^{1/2} \tag{1.10}$$

where the constant A is related to an effect called surface tension.

For an air-water interface near room temperature, $A \approx 74$ cm^3 s^{-2}.

1.3.2 Sound Waves

Sound is a longitudinal compression-expansion wave in a fluid.

The wave speed for sound in an ideal gas is

$$c = (\gamma R T_{abs})^{1/2} \tag{1.11}$$

where γ and R are constants and T_{abs} is the *absolute temperature*.

The absolute temperature is measured in Kelvins and is numerically given by

$$T_{abs} = T_C + 273^\circ \tag{1.12}$$

where T_C is the temperature in Celsius degrees.

The angular frequency of sound waves is thus given by

$$\omega = ck = (\gamma R T_{abs})^{1/2} k. \tag{1.13}$$

The speed of sound in air at normal temperatures is about 340 m s^{-1}.

1.3.3 Light

Light moves in a vacuum at a speed of $c_{vac} = 3 \times 10_8$ m s_{-1}.

In transparent materials it moves at a speed less than c_{vac} by a factor n which is called the *refractive index* of the material:

$$c = c_{vac}/n. \tag{1.14}$$

Often the refractive index takes the form

$$n^2 \approx 1 + \frac{A}{1 - (k/k_R)^2}, \tag{1.15}$$

where k is the wavenumber and k_R and A are positive constants characteristic of the material.

The angular frequency of light in a transparent medium is thus

$$\omega = kc = kc_{vac}/n. \tag{1.16}$$

1.4 Superposition Principle

It is found empirically that as long as the amplitudes of waves in most media are small, two waves in the same physical location don't interact with each other.

Thus, for example, two waves moving in the opposite direction simply pass through each other without their shapes or amplitudes being changed.

When collocated, the total wave displacement is just the sum of the displacements of the individual waves.

This is called the *superposition principle*.

At sufficiently large amplitude the superposition principle often breaks down — interacting waves may scatter off of each other, lose amplitude, or change their form.

Interference is a consequence of the superposition principle.

When two or more waves are superimposed, the net wave displacement is just the algebraic sum of the displacements of the individual waves.

Since these displacements can be positive or negative, the net displacement can either be greater or less than the individual wave displacements.

The former case, which occurs when both displacements are of the same sign, is called *constructive interference*, while *destructive interference* occurs when they are of opposite sign

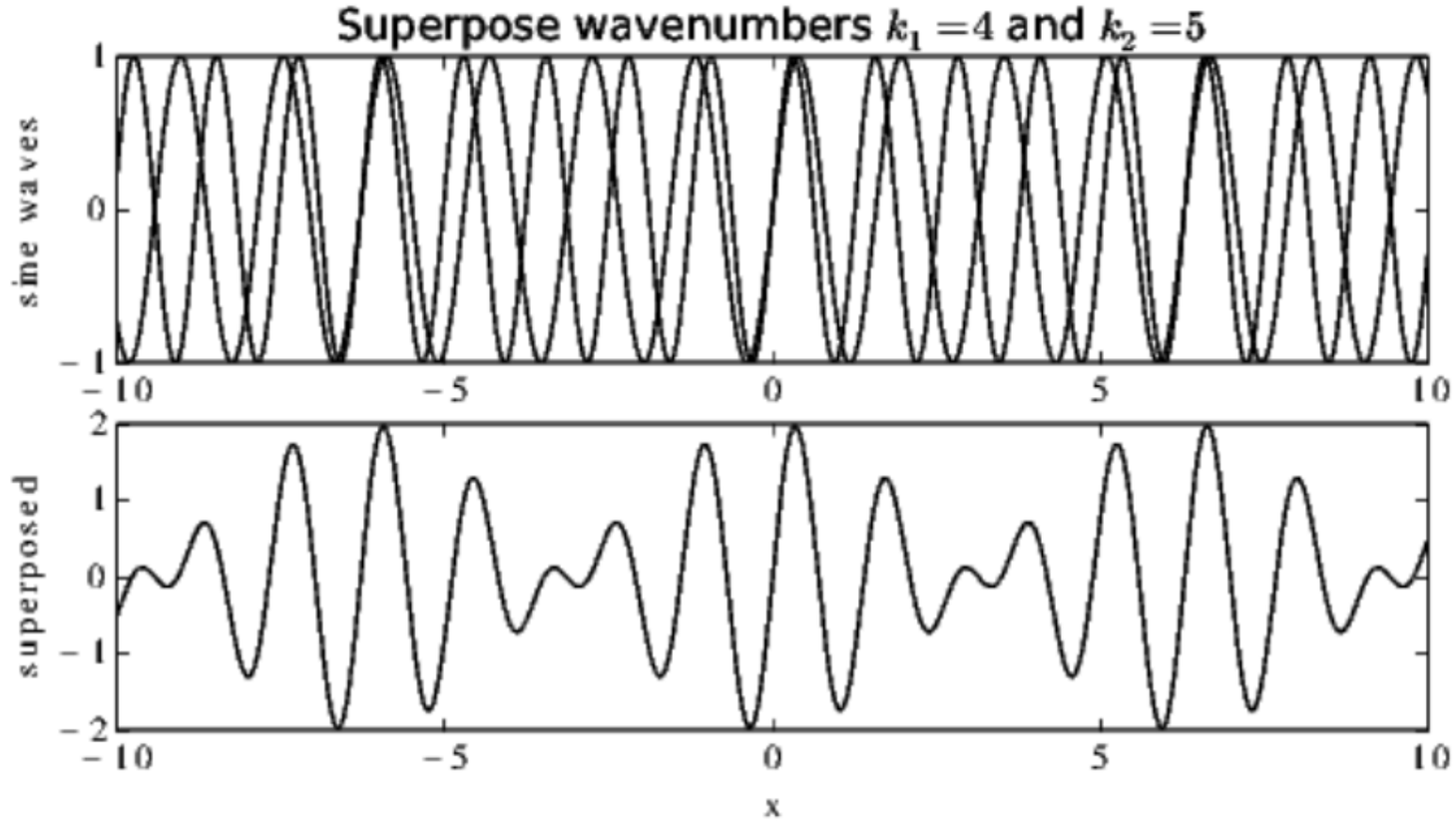


Figure 1.5: Superposition (lower panel) of two sine waves (shown individually in the upper panel) with equal amplitudes and wavenumbers $k_1 = 4$ and $k_2 = 5$.

Let us see what happens when we superimpose two sine waves with different wave numbers.

Figure 1.5 shows the superposition of two waves with wave numbers $k_1 = 4$ and $k_2 = 5$.

Notice that the result is a wave with about the same wavelength as the two initial waves, but which varies in amplitude depending on whether the two sine waves are interfering constructively or destructively.

We say that the waves are *in phase* if they are interfering constructively, and they are *out of phase* if they are interfering destructively.

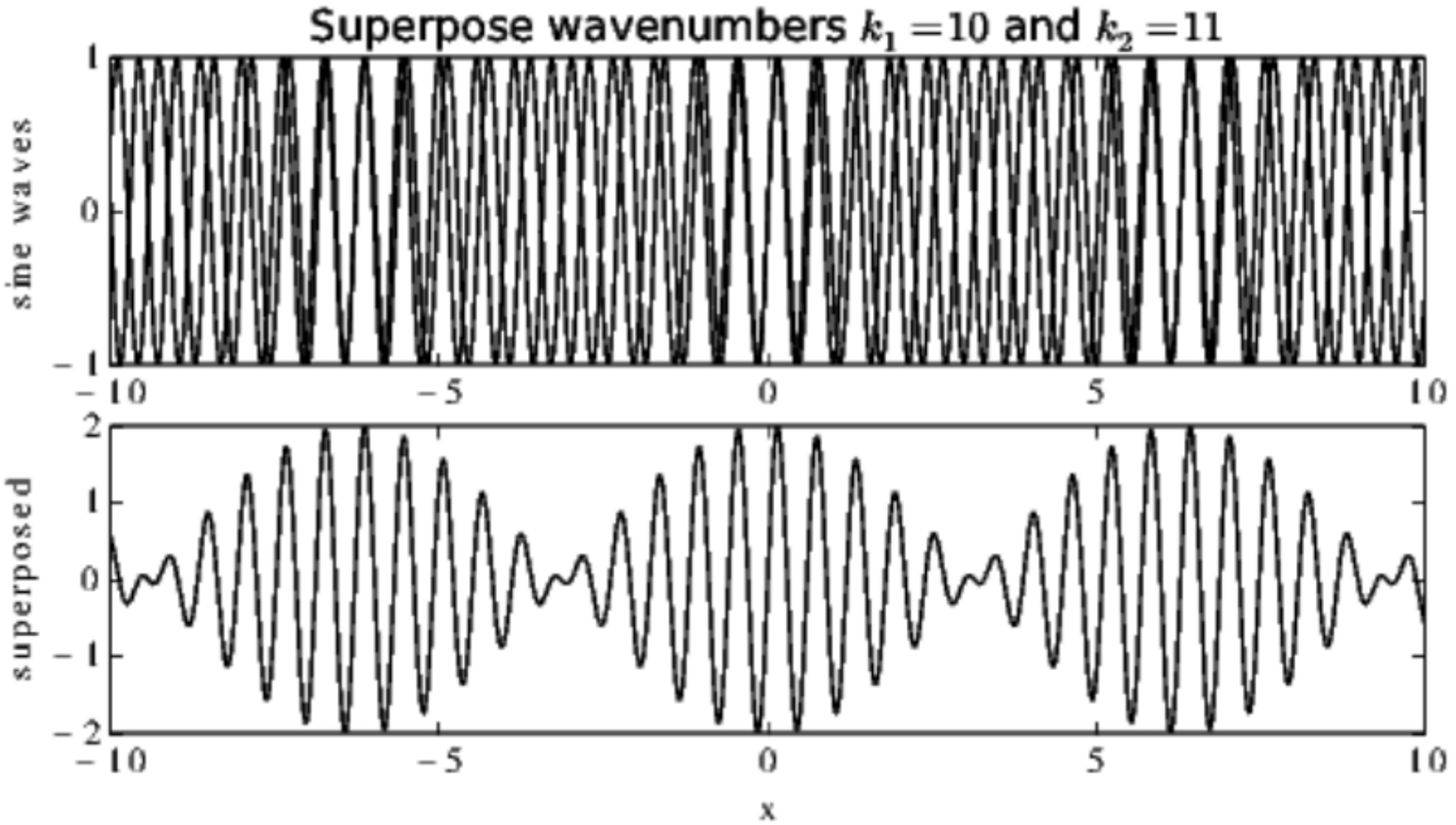


Figure 1.6: Superposition of two sine waves with equal amplitudes and wave numbers $k_1 = 10$ and $k_2 = 11$.

What happens when the wave numbers of the two sine waves are changed?

Figure 1.6 shows the result when $k_1 = 10$ and $k_2 = 11$.

Notice that though the wavelength of the resultant wave is decreased, the locations where the amplitude is maximum have the same separation in x as in figure 1.5.

If we superimpose waves with $k_1 = 10$ and $k_2 = 12$, as is shown in figure 1.7, we see that the x spacing of the regions of maximum amplitude has decreased by a factor of two.

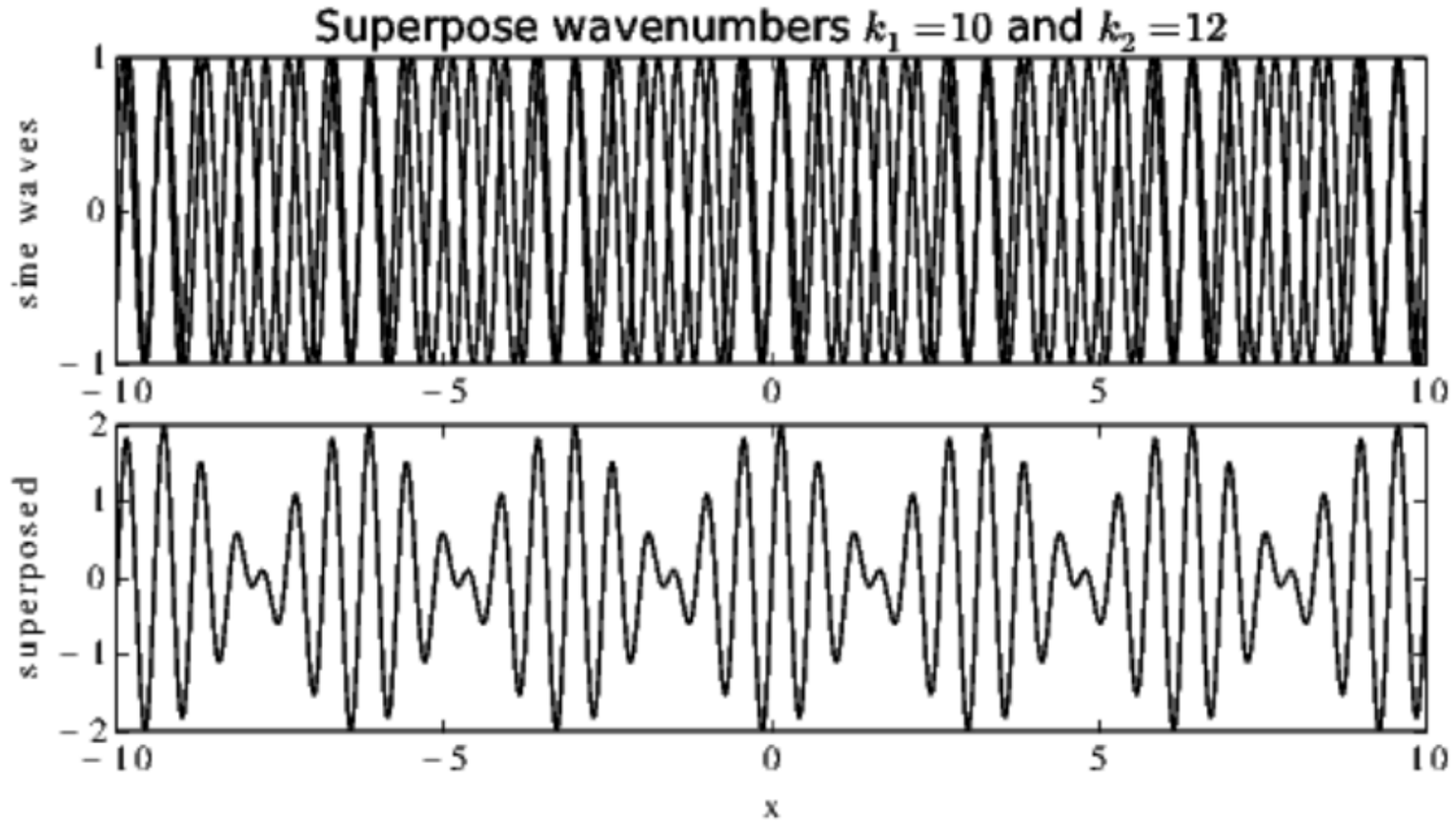


Figure 1.7: Superposition of two sine waves with equal amplitudes and wave numbers $k_1 = 10$ and $k_2 = 12$.

Thus, while the wave number of the resultant wave seems to be related to something like the *average* of the wave numbers of the component waves, the spacing between regions of maximum wave amplitude appears to go inversely with the *difference* of the wave numbers of the component waves.

In other words, if k_1 and k_2 are close together, the amplitude maxima are far apart and vice versa.

We can symbolically represent the sine waves that make up figures 1.5, 1.6, and 1.7 by a plot such as that shown in figure 1.8.

The amplitudes and wave numbers of each of the sine waves are indicated by vertical lines in this figure.

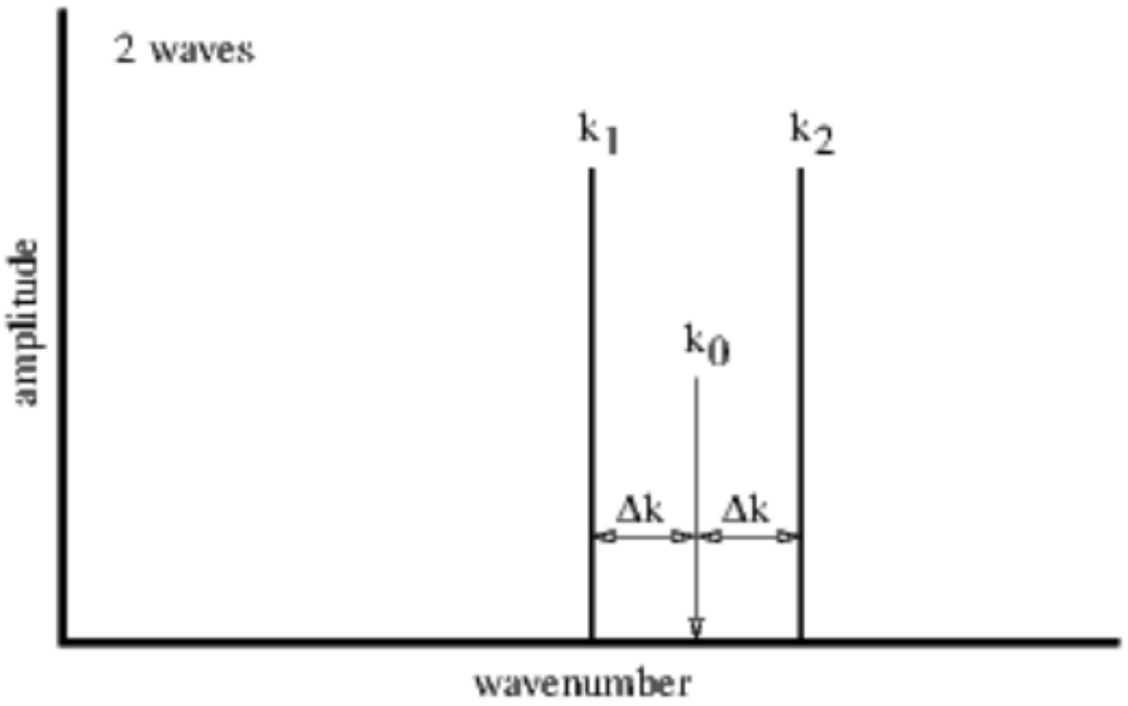


Figure 1.8: Representation of the wave numbers and amplitudes of two superimposed sine waves.

The regions of large wave amplitude are called *wave packets*.

Wave packets will play a central role in what is to follow, so it is important that we acquire a good understanding of them.

The wave packets produced by only two sine waves are not well separated along the x-axis.

However, if we superimpose many waves, we can produce an isolated wave packet.

For example, figure 1.9 shows the results of superimposing 20 sine waves with wave numbers $k = 0.4m$, $m = 1, 2, \ldots, 20$, where the amplitudes of the waves are largest for wave numbers near $k = 4$.

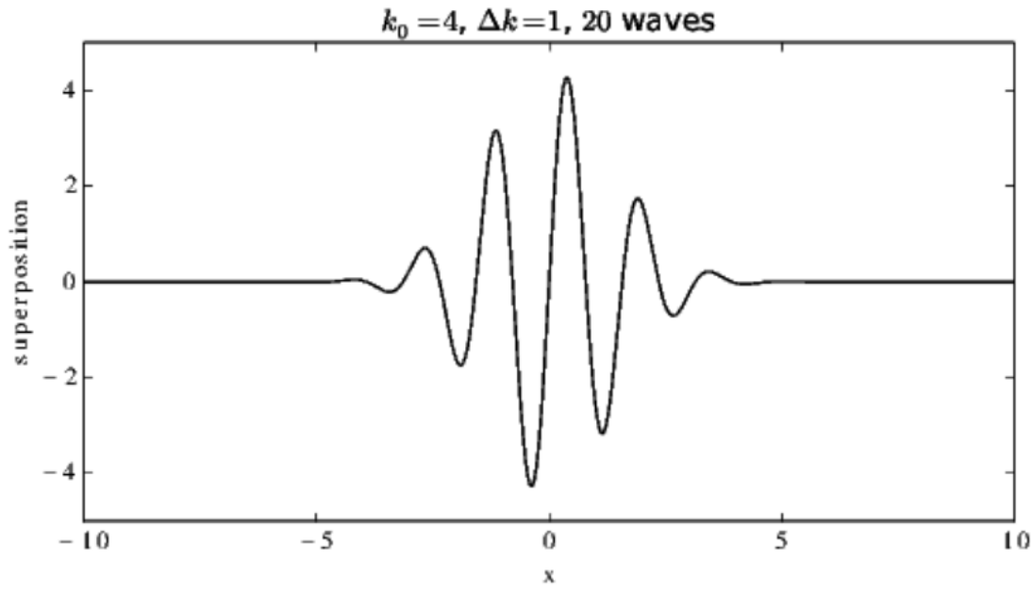


Figure 1.9: Superposition of twenty sine waves with $k_0 = 4$ and $\Delta k = 1$.

In particular, we assume that the amplitude of each sine wave is proportional to $\exp[-(k - k_2)^2 / \Delta k^2]$, where $k_0 = 4$ defines the maximum of the distribution of wave numbers and $\Delta k = 1$ defines the half-width of this distribution.

The amplitudes of each of the sine waves making up the wave packet in figure 1.9 are shown schematically in figure 1.10.

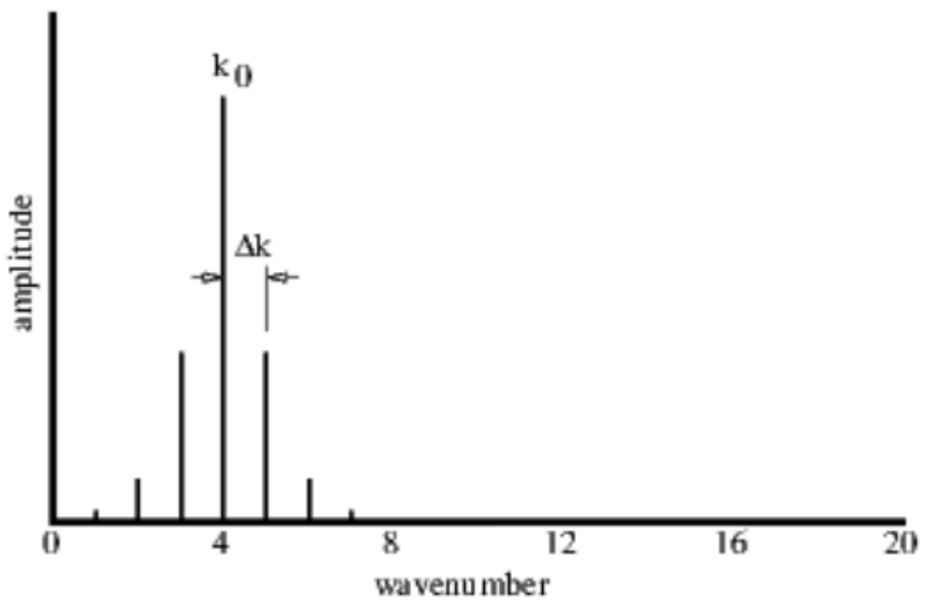


Figure 1.10: Representation of the distribution of wave numbers and amplitudes of 20 superimposed sine waves with maximum at $k_0 = 4$ and half-width $\Delta k = 1$.

The quantity Δk controls the distribution of the sine waves being superimposed — only those waves with a wave number k within approximately Δk of the *central wave number* k_0 of the wave packet, i. e., for $3 \leq k \leq 5$ in this case, contribute significantly to the sum.

If Δk is changed to 2, so that wave numbers in the range $2 \leq k \leq 6$ contribute significantly, the wave packet becomes narrower, as is shown in figures 1.11 and 1.12.

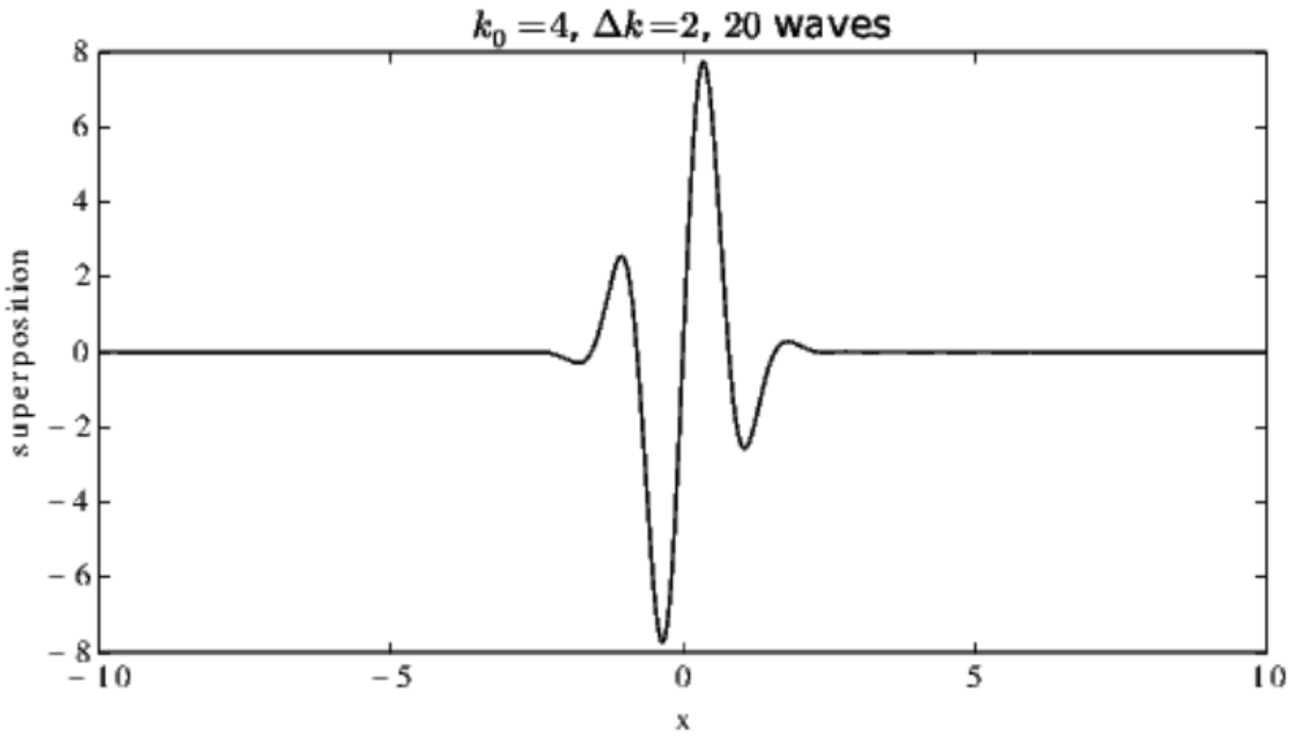


Figure 1.11: Superposition of twenty sine waves with $k_0 = 4$ and $\Delta k = 2$.

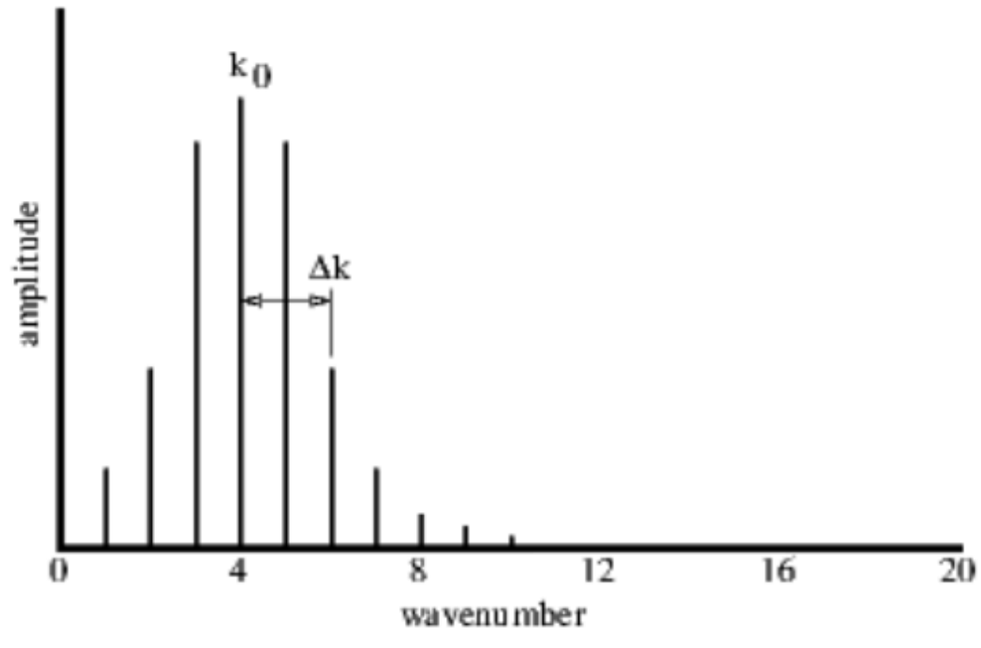


Figure 1.12: Representation of the distribution of wave numbers and amplitudes of 20 superimposed sine waves with maximum at $k_0 = 4$ and half-width $\Delta k = 2$.

Δk is called the *wave number spread* of the wave packet, and it evidently plays a role similar to the difference in wave numbers in the superposition of two sine waves — the larger the wave number spread, the smaller the physical size of the wave packet.

Furthermore, the wavenumber of the oscillations within the wave packet is given approximately by the central wavenumber.

We can better understand how wave packets work by mathematically analyzing the simple case of the superposition of two sine waves.

Let us define $k_0 = (k_1 + k_2)/2$ where k1 and k2 are the wavenumbers of the component waves.

Furthermore let us set $\Delta k = (k_2 - k_1)/2$.

The quantities k_0 and Δk are graphically illustrated in figure 1.8.

We can write $k_1 = k_0 - \Delta k$ and $k_2 = k_0 + \Delta k$ and use the trigonometric identity sin(a+ b) = sin(a) cos(b) + cos(a) sin(b) to find

$$\begin{aligned} \sin(k_1 x) + \sin(k_2 x) &= \sin[(k_0 - \Delta k)x] + \sin[(k_0 + \Delta k)x] \\ &= \sin(k_0 x)\cos(\Delta k x) - \cos(k_0 x)\sin(\Delta k x) + \\ &\quad \sin(k_0 x)\cos(\Delta k x) + \cos(k_0 x)\sin(\Delta k x) \\ &= 2\sin(k_0 x)\cos(\Delta k x). \end{aligned} \tag{1.17}$$

The sine factor on the bottom line of the above equation produces the oscillations within the wave packet, and as speculated earlier, this oscillation has a wavenumber k_0 equal to the average of the wavenumbers of the component waves.

The cosine factor modulates this wave with a spacing between regions of maximum amplitude of

$$\Delta x = \pi / \Delta k. \tag{1.18}$$

Thus, as we observed in the earlier examples, the length of the wave packet Δx is inversely related to the spread of the wave numbers Δk (which in this case is just the difference between the two wave numbers) of the component waves.

This relationship is central to the uncertainty principle of quantum mechanics.

1.5 Beats

Suppose two sound waves of different frequency but equal amplitude impinge on your ear at the same time.

The displacement perceived by your ear is the superposition of these two waves, with time dependence

$$h(t) = \sin(\omega_1 t) + \sin(\omega_2 t) = 2\sin(\omega_0 t)\cos(\Delta\omega t), \tag{1.19}$$

where we have used the above math trick, and where $\omega_0 = (\omega_1 + \omega_2)/2$ and $\Delta\omega = (\omega_2 - \omega_1)/2$.

What you actually hear is a tone with angular frequency ω_0 which fades in and out with period

$$T_{beat} = \pi/|\Delta\omega| = 2\pi/|\omega_2 - \omega_1| = 1/|f_2 - f_1|. \tag{1.20}$$

The beat frequency is simply

$$f_{beat} = 1/T_{beat} = |f_2 - f_1|. \tag{1.21}$$

Note how beats are the time analog of wave packets — the mathematics are the same except that frequency replaces wavenumber and time replaces space.

1.6 Interferometers

An interferometer is a device which splits a beam of light (or other wave) into two sub-beams, shifts the phase of one sub-beam with respect to the other, and then superimposes the sub-beams so that they interfere constructively or destructively, depending on the magnitude of the phase shift between them.

In this section we study the Michelson interferometer and interferometric effects in thin films.

1.6.1 The Michelson Interferometer

The American physicist Albert Michelson invented the optical interferometer illustrated in figure 1.13.

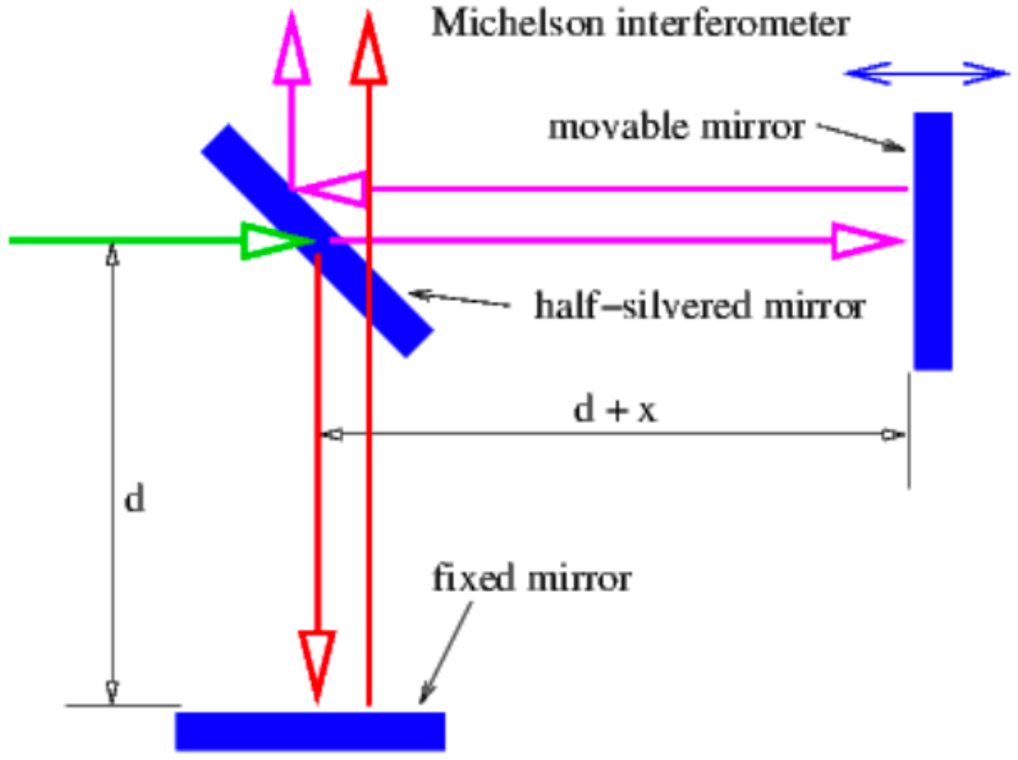


Figure 1.13: Sketch of a Michelson interferometer.

The incoming beam is split into two beams by the half-silvered mirror.

Each sub-beam reflects off of another mirror which returns it to the half-silvered mirror, where the two sub-beams recombine as shown.

One of the reflecting mirrors is movable by a sensitive micrometer device, allowing the path length of the corresponding sub-beam, and hence the phase relationship between the two sub-beams, to be altered.

As figure 1.13 shows, the difference in path length between the two sub-beams is $2x$ because the horizontal sub-beam traverses the path twice.

Thus, constructive interference occurs when this path difference is an integral number of wavelengths, i.e.,

$$2x = m\lambda, \quad m = 0, \pm 1, \pm 2, \ldots \quad \text{(Michelson interferometer)} \tag{1.22}$$

where λ is the wavelength of the wave and m is an integer.

Note that m is the number of wavelengths that fits evenly into the distance $2x$.

1.7 Thin Films

One of the most revealing examples of interference occurs when light interacts with a thin film of transparent material such as a soap bubble.

Figure 1.14 shows how a plane wave normally incident on the film is partially reflected by the front and rear surfaces.

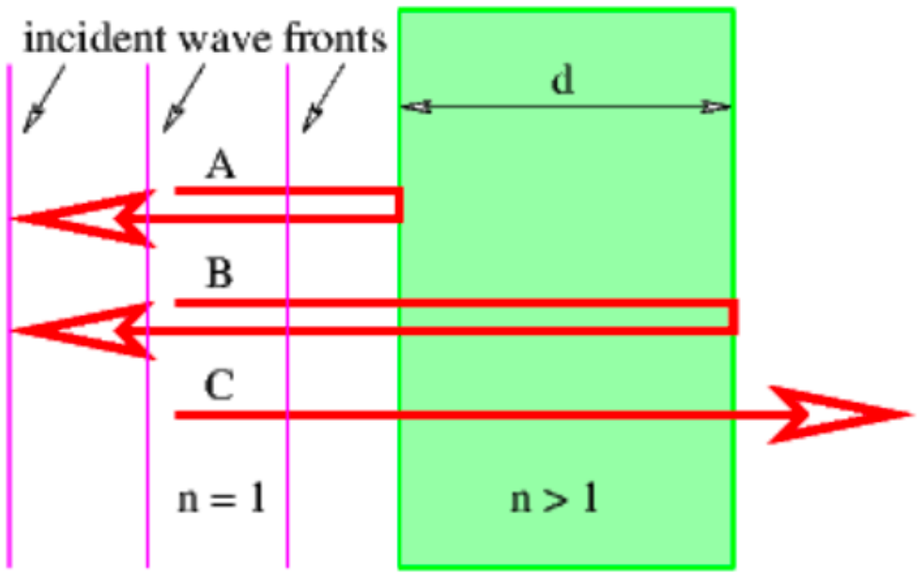


Figure 1.14: Plane light wave normally incident on a transparent thin film of thickness d and index of refraction $n > 1$. Partial reflection occurs at the front surface of the film, resulting in beam A, and at the rear surface, resulting in beam B. Much of the wave passes completely through the film, as with C.

The waves reflected off the front and rear surfaces of the film interfere with each other.

The interference can be either constructive or destructive depending on the phase difference between the two reflected waves.

If the wavelength of the incoming wave is λ, one would naively expect constructive interference to occur between the A and B beams if $2d$ were an integral multiple of λ.

Two factors complicate this picture.

First, the wavelength inside the film is *not* λ, but λ/n, where n is the index of refraction of the film.

Constructive interference would then occur if $2d = m\lambda/n$.

Second, it turns out that an additional phase shift of half a wavelength occurs upon reflection when the wave is incident on material with a higher index of refraction than the medium in which the incident beam is immersed.

This phase shift doesn't occur when light is reflected from a region with lower index of refraction than felt by the incident beam.

Thus beam B doesn't acquire any additional phase shift upon reflection.

As a consequence, constructive interference actually occurs when

$$2d = (m + 1/2)\lambda/n, \quad m = 0, 1, 2, \ldots \quad \text{(constructive interference)} \tag{1.23}$$

while destructive interference results when

$$2d = m\lambda/n, \quad m = 0, 1, 2, \ldots \quad \text{(destructive interference)}. \tag{1.24}$$

When we look at a soap bubble, we see bands of colors reflected back from a light source.

What is the origin of these bands?

Light from ordinary sources is generally a mixture of wavelengths ranging from roughly $\lambda = 4.5 \times 10^{-7}$ m (violet light) to $\lambda = 6.5 \times 10^{-7}$ m (red light).

In between violet and red we also have blue, green, and yellow light, in that order.

Because of the different wavelengths associated with different colors, it is clear that for a mixed light source we will have some colors interfering constructively while others interfere destructively.

Those undergoing constructive interference will be visible in reflection, while those undergoing destructive interference will not.

Another factor enters as well.

If the light is not normally incident on the film, the difference in the distances traveled between beams reflected off of the front and rear faces of the film will not be just twice the thickness of the film.

To understand this case quantitatively, we need the concept of refraction, which will be developed later in the context of geometrical optics.

However, it should be clear that different wavelengths will undergo constructive interference for different angles of incidence of the incoming light.

Different portions of the thin film will in general be viewed at different angles, and will therefore exhibit different colors under reflection, resulting in the colorful patterns normally seen in soap bubbles.

1.8 Math Review — Derivatives

This section provides a quick review of the idea of the derivative.

Often we are interested in the slope of a line tangent to a function $y(x)$ at some value of x.

This slope is called the *derivative* and is denoted dy/dx.

Since a tangent line to the function can be defined at any point x, the derivative itself is a function of x:

$$g(x) = \frac{dy(x)}{dx}. \tag{1.25}$$

As figure 1.15 illustrates, the slope of the tangent line at some point on the function may be approximated by the slope of a line connecting two points, A and B, set a finite distance apart on the curve:

$$\frac{dy}{dx} \approx \frac{\Delta y}{\Delta x}. \tag{1.26}$$

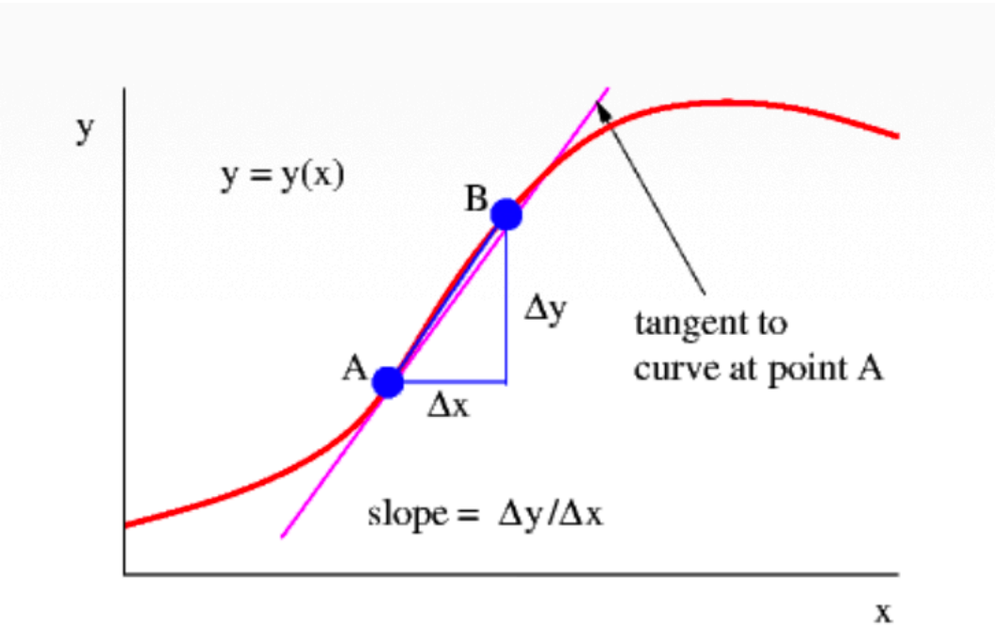


Figure 1.15: Estimation of the derivative, which is the slope of the tangent line. When point B approaches point A, the slope of the line AB approaches the slope of the tangent to the curve at point A.

As B is moved closer to A, the approximation becomes better.

In the limit when B moves infinitely close to A, it is exact.

Derivatives of some common functions are now given.

In each case a is a constant.

$$\frac{dx^a}{dx} = ax^{a-1} \tag{1.27}$$

$$\frac{d}{dx}\exp(ax) = a\exp(ax) \tag{1.28}$$

$$\frac{d}{dx}\log(ax) = \frac{1}{x} \tag{1.29}$$

$$\frac{d}{dx}\sin(ax) = a\cos(ax) \tag{1.30}$$

$$\frac{d}{dx}\cos(ax) = -a\sin(ax) \tag{1.31}$$

$$\frac{da\,f(x)}{dx} = a\frac{df(x)}{dx} \tag{1.32}$$

$$\frac{d}{dx}[f(x) + g(x)] = \frac{df(x)}{dx} + \frac{dg(x)}{dx} \tag{1.33}$$

$$\frac{d}{dx}f(x)g(x) = \frac{df(x)}{dx}g(x) + f(x)\frac{dg(x)}{dx} \quad \text{(product rule)} \tag{1.34}$$

$$\frac{d}{dx}f(y) = \frac{df}{dy}\frac{dy}{dx} \quad \text{(chain rule)} \tag{1.35}$$

The product and chain rules are used to compute the derivatives of complex functions.

For instance,

$$\frac{d}{dx}(\sin(x)\cos(x)) = \frac{d\sin(x)}{dx}\cos(x) + \sin(x)\frac{d\cos(x)}{dx} = \cos^2(x) - \sin^2(x)$$

and

$$\frac{d}{dx}\log(\sin(x)) = \frac{1}{\sin(x)}\frac{d\sin(x)}{dx} = \frac{\cos(x)}{\sin(x)}.$$

1.9 Group Velocity

We now ask the following question: How fast do wave packets move?

Surprisingly, we often find that wave packets move at a speed very different from the phase speed, $c = \omega/k$, of the wave composing the wave packet.

We shall find that the speed of motion of wave packets, referred to as the *group velocity*, is given by

$$u = \left.\frac{d\omega}{dk}\right|_{k=k_0} \qquad \text{(group velocity)}. \tag{1.36}$$

The derivative of $\omega(k)$ with respect to k is first computed and then evaluated at $k = k_0$, the central wavenumber of the wave packet of interest.

The relationship between the angular frequency and the wavenumber for a wave, $\omega = \omega(k)$, depends on the type of wave being considered.

Whatever this relationship turns out to be in a particular case, it is called the *dispersion relation* for the type of wave in question.

As an example of a group velocity calculation, suppose we want to find the velocity of deep ocean wave packets for a central wavelength of $\lambda_0 = 60$ m.

This corresponds to a central wavenumber of $k_0 = 2\pi/\lambda_0 \approx 0.1$ m_{-1}.

The phase speed of deep ocean waves is $c = (g/k)^{1/2}$.

However, since $c \equiv \omega/\ k$, we find the frequency of deep ocean waves to be $\omega = (gk)^{1/2}$.

The group velocity is therefore $u \equiv d\omega/\ dk = (g/k)^{1/2}/2 = c/2$.

For the specified central wavenumber, we find that $u \approx (9.8\ \text{m s}^{-2}/0.1\ \text{m}^{-1})^{1/2}/2 \approx 5\ \text{m s}^{-1}$.

By contrast, the phase speed of deep ocean waves with this wavelength is $c \approx 10\ \text{m s}^{-1}$.

Dispersive waves are waves in which the phase speed varies with wavenumber.

It is easy to show that dispersive waves have unequal phase and group velocities, while these velocities are equal for non-dispersive waves.

1.9.1 Derivation of Group Velocity Formula

We now derive equation (1.36).

It is easiest to do this for the simplest wave packets, namely those constructed out of the superposition of just two sine waves.

We will proceed by adding two waves with full space and time dependence:

$$h = \sin(k_1 x - \omega_1 t) + \sin(k_2 x - \omega_2 t) \tag{1.37}$$

After algebraic and trigonometric manipulations familiar from earlier sections, we find

$$h = 2\sin(k_0 x - \omega_0 t)\cos(\Delta k x - \Delta\omega t), \tag{1.38}$$

where as before we have $k_0 = (k_1 + k_2)/2$, $\omega_0 = (\omega_1 + \omega_2)/2$, $\Delta k = (k_2 - k_1)/2$, and $\Delta\omega = (\omega_2 - \omega_1)/2$.

Again think of this as a sine wave of frequency ω_0 and wavenumber k_0 modulated by a cosine function.

In this case the modulation pattern moves with a speed so as to keep the argument of the cosine function constant:

$$\Delta k x - \Delta\omega t = const. \tag{1.39}$$

Differentiating this with respect to t while holding Δk and $\Delta\omega$ constant yields

$$u \equiv \frac{dx}{dt} = \frac{\Delta\omega}{\Delta k}. \tag{1.40}$$

In the limit in which the deltas become very small, this reduces to the derivative

$$u = \frac{d\omega}{dk}, \tag{1.41}$$

which is the desired result.

1.9.2 Examples

We now illustrate some examples of phase speed and group velocity by showing the displacement resulting from the superposition of two sine waves, as given by equation (1.38), in the *x*-*t* plane.

This is an example of a spacetime diagram, of which we will see many examples later on.

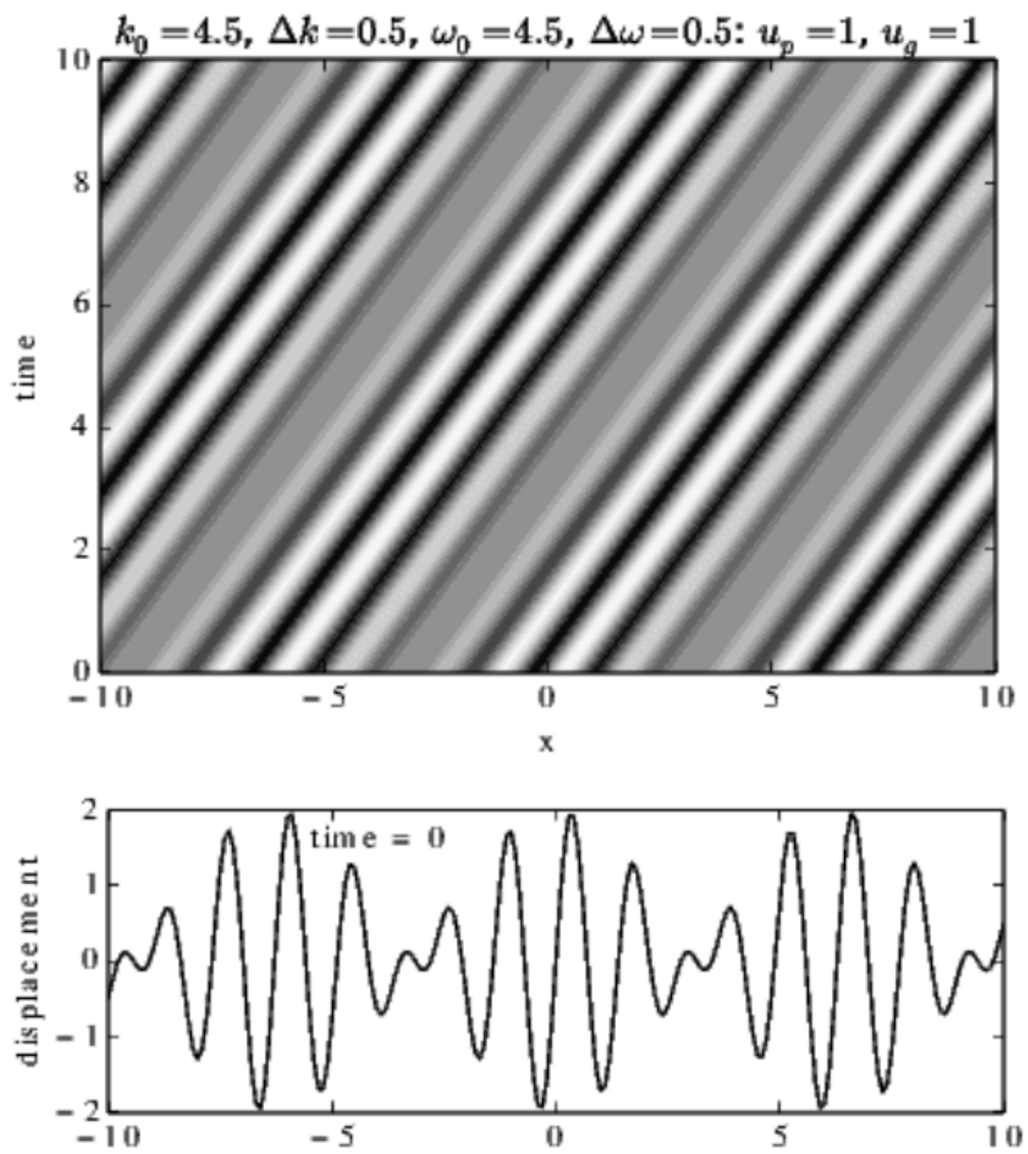


Figure 1.16: Upper panel: Net displacement of the sum of two traveling sine waves plotted in the *x*-*t* plane. The white indicates where the displacement is large and positive, while the black indicates where it is large and negative. Non-dispersive case. Lower panel: Plot of wave displacement as a function of *x* at time $t = 0$.

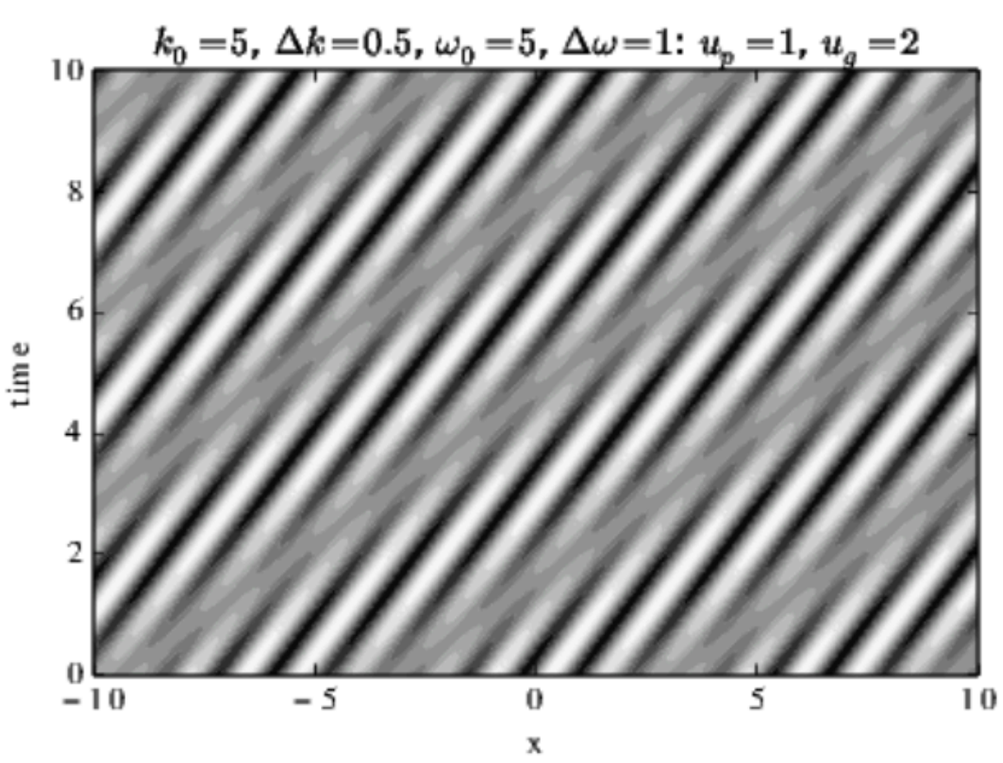


Figure 1.17: As in the upper panel of figure 1.16 except a dispersive case.

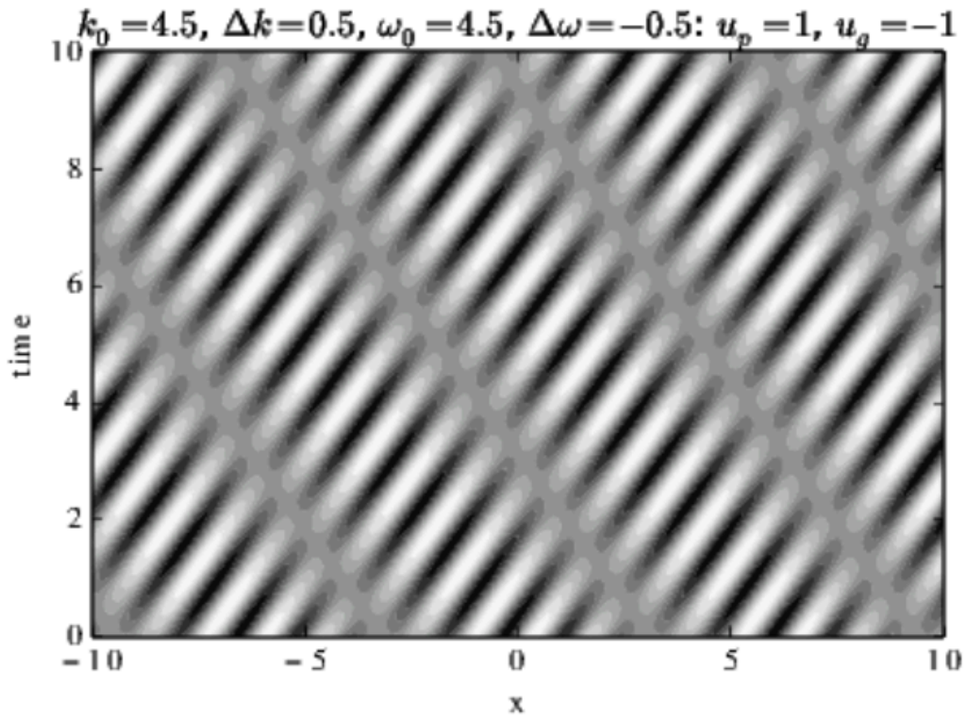


Figure 1.18: As in the upper panel of figure 1.16 except a dispersive case with phase and group velocities in opposite directions.

The upper panel of figure 1.16 shows a non-dispersive case in which the phase speed equals the group velocity.

The white and black regions indicate respectively strong wave crests and troughs (i. e., regions of large positive and negative displacements), with grays indicating a displacement near zero.

Regions with large displacements indicate the location of wave packets.

The positions of waves and wave packets at any given time may therefore be determined by drawing a horizontal line across the graph at the desired time and examining the variations in wave displacement along this line.

The lower panel of this figure shows the wave displacement as a function of x at time $t = 0$ as an aid to interpretation of the upper panel.

Notice that as time increases, the crests move to the right.

This corresponds to the motion of the waves within the wave packets.

Note also that the wave packets, i. e., the broad regions of large positive and negative amplitudes, move to the right with increasing time as well.

Since velocity is distance moved Δx divided by elapsed time Δt, the slope of a line in figure 1.16, $\Delta t / \Delta x$, is one over the velocity of whatever that line represents.

The slopes of lines representing crests are the same as the slopes of lines representing wave packets in this case, which indicates that the two move at the same velocity.

Since the speed of movement of wave crests is the phase speed and the speed of movement of wave packets is the group velocity, the two velocities are equal and the non-dispersive nature of this case is confirmed.

Figure 1.17 shows a dispersive wave in which the group velocity is twice the phase speed, while figure 1.18 shows a case in which the group velocity is actually opposite in sign to the phase speed.

See if you can confirm that the phase and group velocities seen in each figure correspond to the values for these quantities calculated from the specified frequencies and wave numbers.

Chapter 2

Waves in Two and Three Dimensions

In this chapter we extend the ideas of the previous chapter to the case of waves in more than one dimension.

The extension of the sine wave to higher dimensions is the *plane wave*.

Wave packets in two and three dimensions arise when plane waves moving in different directions are superimposed.

Diffraction results from the disruption of a wave which is impingent upon an object.

Those parts of the wave front hitting the object are scattered, modified, or destroyed.

The resulting *diffraction pattern* comes from the subsequent interference of the various pieces of the modified wave.

A knowledge of diffraction is necessary to understand the behavior and limitations of optical instruments such as telescopes.

Diffraction and interference in two and three dimensions can be manipulated to produce useful devices such as the *diffraction grating*.

2.1 Math Tutorial — Vectors

Before we can proceed further we need to explore the idea of a *vector*.

A vector is a quantity which expresses both magnitude and direction.

Graphically we represent a vector as an arrow.

In typeset notation a vector is represented by a boldface character, while in handwriting an arrow is drawn over the character representing the vector.

Figure 2.1 shows some examples of *displacement vectors*, i. e., vectors which represent the displacement of one object from another, and introduces the idea of vector addition.

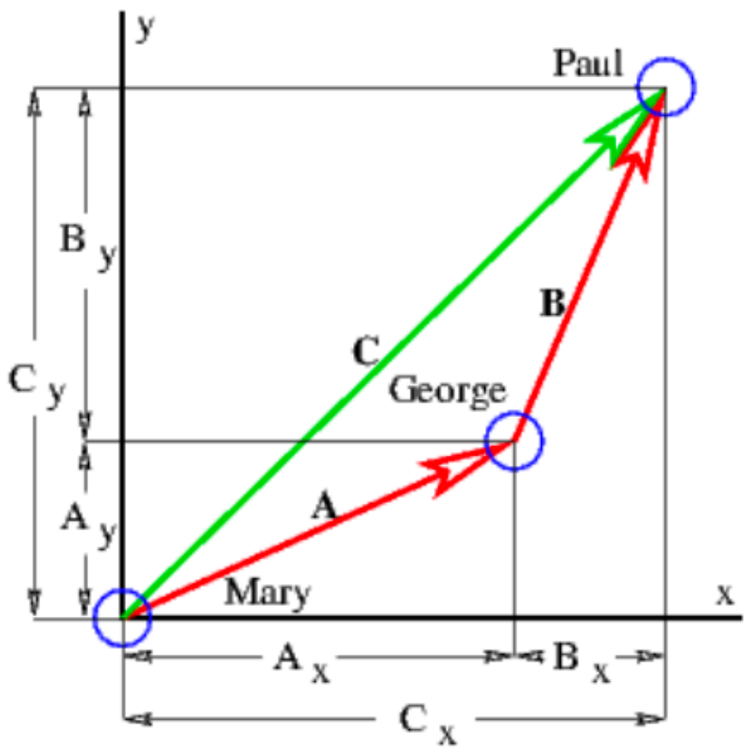


Figure 2.1: Displacement vectors in a plane. Vector **A** represents the displacement of George from Mary, while vector **B** represents the displacement of Paul from George. Vector **C** represents the displacement of Paul from Mary and **C** = **A** + **B**. The quantities A_x, A_y, etc., represent the Cartesian components of the vectors.

The tail of vector **B** is collocated with the head of vector **A**, and the vector which stretches from the tail of **A** to the head of **B** is the sum of **A** and **B**, called **C** in figure 2.1.

The quantities A_x, A_y, etc., represent the *Cartesian components* of the vectors in figure 2.1.

A vector can be represented either by its Cartesian components, which are just the projections of the vector onto the Cartesian coordinate axes, or by its direction and magnitude.

The direction of a vector in two dimensions is generally represented by the counterclockwise angle of the vector relative to the x axis. as shown in figure 2.2.

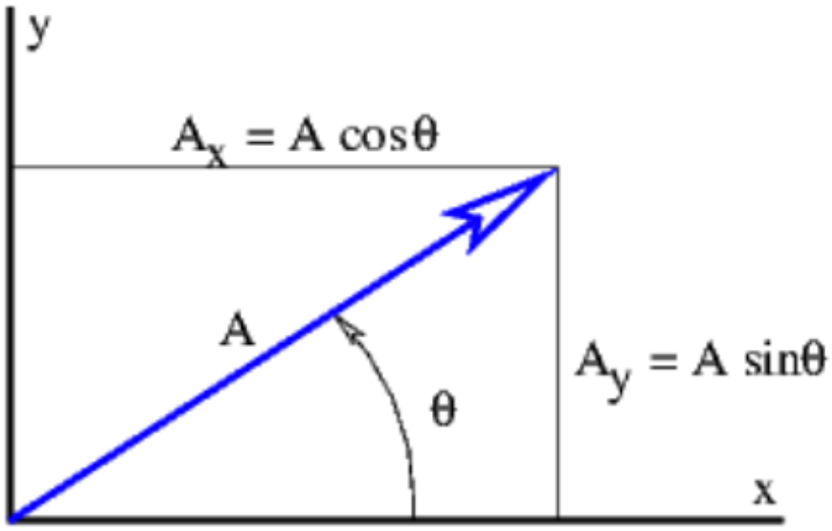


Figure 2.2: Definition sketch for the angle θ representing the orientation of a two dimensional vector.

Conversion from one form to the other is given by the equations

$$A = (A_x^2 + A_y^2)^{1/2} \qquad \theta = \tan^{-1}(A_y/A_x), \tag{2.1}$$

$$A_x = A\cos(\theta) \qquad A_y = A\sin(\theta), \tag{2.2}$$

where A is the magnitude of the vector.

A vector magnitude is sometimes represented by absolute value notation: $A \equiv |\mathbf{A}|$.

Notice that the inverse tangent gives a result which is ambiguous relative to adding or subtracting integer multiples of π.

Thus the quadrant in which the angle lies must be resolved by independently examining the signs of A_x and A_y and choosing the appropriate value of θ.

To add two vectors, **A** and **B**, it is easiest to convert them to Cartesian component form.

The components of the sum **C** = **A** + **B** are then just the sums of the components:

$$C_x = A_x + B_x \qquad C_y = A_y + B_y. \tag{2.3}$$

Subtraction of vectors is done similarly, e. g., if **A** = **C** - **B**, then

$$A_x = C_x - B_x \qquad A_y = C_y - B_y. \tag{2.4}$$

A *unit vector* is a vector of unit length.

One can always construct a unit vector from an ordinary (non-zero) vector by dividing the vector by its length: **n** = **A**/|**A**|.

This division operation is carried out by dividing each of the vector components by the number in the denominator.

Alternatively, if the vector is expressed in terms of length and direction, the magnitude of the vector is divided by the denominator and the direction is unchanged.

Unit vectors can be used to define a Cartesian coordinate system.

Conventionally, **i**, **j**, and **k** indicate the x, y, and z axes of such a system.

Note that **i**, **j**, and **k** are mutually perpendicular.

Any vector can be represented in terms of unit vectors and its Cartesian components: $\mathbf{A} = A\mathbf{i} + A_y\mathbf{j} + A\mathbf{k}$.

An alternate way to represent a vector is as a list of components: $\mathbf{A} = (A_x, A_y, A_z)$.

We tend to use the latter representation since it is somewhat more economical notation.

There are two ways to multiply two vectors, yielding respectively what are known as the *dot product* and the *cross product*.

The cross product yields another vector while the dot product yields a number.

Here we will discuss only the dot product.

The cross product will be presented later when it is needed.

Given vectors **A** and **B**, the dot product of the two is defined as

$$\mathbf{A} \cdot \mathbf{B} \equiv |\mathbf{A}||\mathbf{B}| \cos\theta, \tag{2.5}$$

where θ is the angle between the two vectors.

In two dimensions an alternate expression for the dot product exists in terms of the Cartesian components of the vectors:

$$\mathbf{A} \cdot \mathbf{B} = A_x B_x + A_y B_y. \tag{2.6}$$

It is easy to show that this is equivalent to the cosine form of the dot product when the x axis lies along one of the vectors, as in figure 2.3.

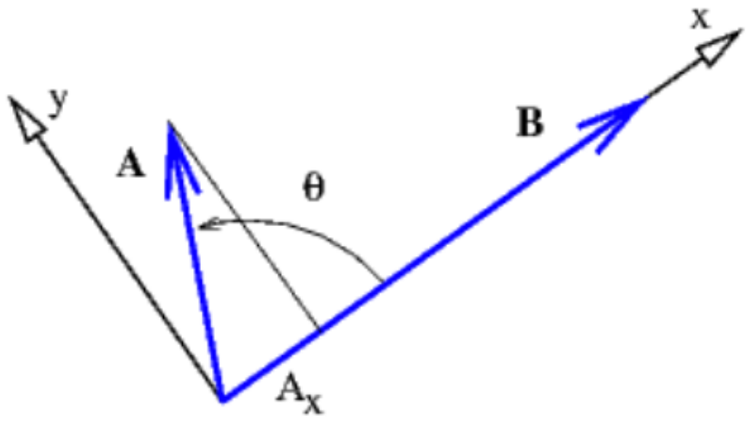


Figure 2.3: Definition sketch for dot product.

Notice in particular that $A_x = |\mathbf{A}| \cos\theta$, while $B_x = |\mathbf{B}|$ and $B_y = 0$.

Thus, $\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}| \cos\theta |\mathbf{B}|$ in this case, which is identical to the form given in equation (2.5).

All that remains to be proven for equation (2.6) to hold in general is to show that it yields the same answer regardless of how the Cartesian coordinate system is oriented relative to the vectors.

To do this, we must show that $A_x B_x + A_y B_y = A'_x B'_x + A'_y B'_y$, where the primes indicate components in a coordinate system rotated from the original coordinate system.

Figure 2.4 shows the vector **R** resolved in two coordinate systems rotated with respect to each other.

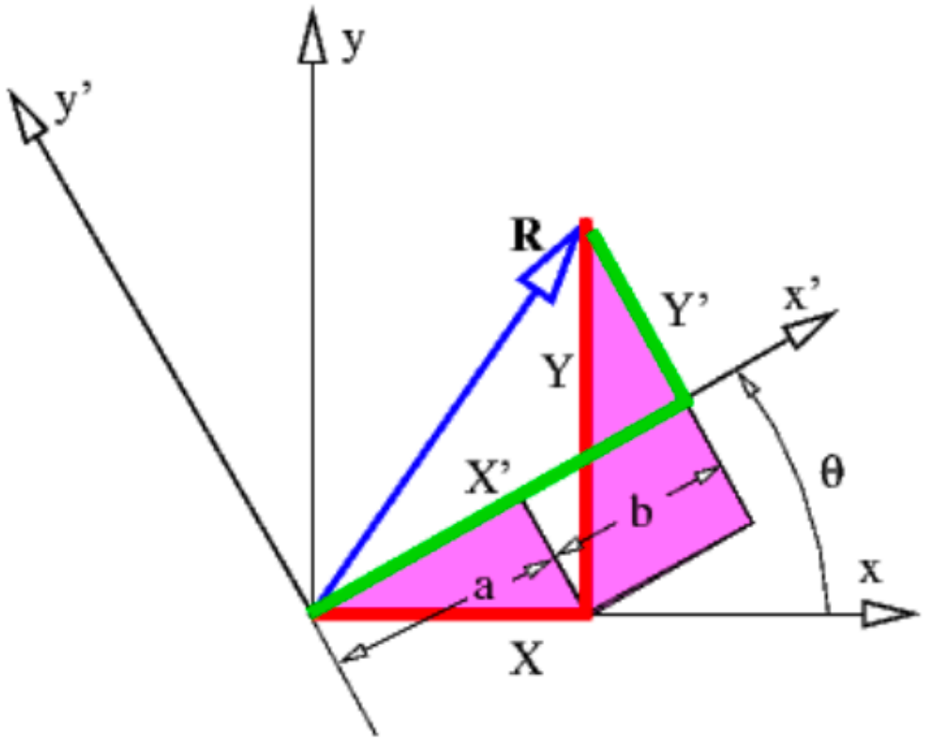


Figure 2.4: Definition figure for rotated coordinate system. The vector **R** has components X and Y in the unprimed coordinate system and components X' and Y' in the primed coordinate system.

From this figure it is clear that $X' = a + b$.

Focusing on the shaded triangles, we see that $a = X \cos\theta$ and $b = Y \sin\theta$.

Thus, we find $X' = X \cos\theta + Y \sin\theta$.

Similar reasoning shows that $Y' = -X \sin\theta + Y \cos\theta$.

Substituting these and using the trigonometric identity $\cos^2\theta + \sin^2\theta = 1$ results in

$$\begin{aligned} A'_x B'_x + A'_y B'_y &= (A_x \cos\theta + A_y \sin\theta)(B_x \cos\theta + B_y \sin\theta) \\ &+ (-A_x \sin\theta + A_y \cos\theta)(-B_x \sin\theta + B_y \cos\theta) \\ &= A_x B_x + A_y B_y \end{aligned} \tag{2.7}$$

thus proving the complete equivalence of the two forms of the dot product as given by equations (2.5) and (2.6). Multiply out the above expression to verify this.

A numerical quantity that doesn't depend on which coordinate system is being used is called a *scalar*.

The dot product of two vectors is a scalar.

However, the components of a vector, taken individually, are not scalars, since the components change as the coordinate system changes.

Since the laws of physics cannot depend on the choice of coordinate system being used, we insist that physical laws be expressed in terms of scalars and vectors, but not in terms of the components of vectors.

In three dimensions the cosine form of the dot product remains the same, while the component form is

$$\mathbf{A} \cdot \mathbf{B} = A_x B_x + A_y B_y + A_z B_z. \tag{2.8}$$

2.2 Plane Waves

A plane wave in two or three dimensions is like a sine wave in one dimension except that crests and troughs aren't points, but form lines (2-D) or planes (3-D) perpendicular to the direction of wave propagation.

Figure 2.5 shows a plane sine wave in two dimensions.

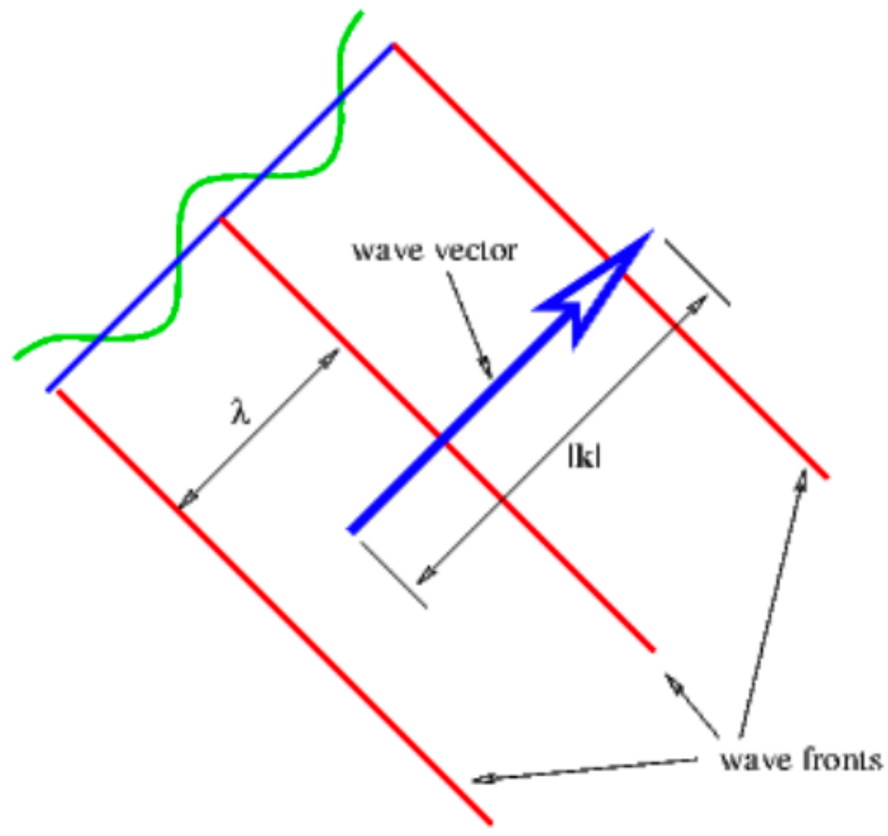


Figure 2.5: Definition sketch for a plane sine wave in two dimensions. The wave fronts are constant phase surfaces separated by one wavelength. The wave vector is normal to the wave fronts and its length is the wavenumber.

The large arrow is a vector called the *wave vector*, which defines (1) the direction of wave propagation by its orientation perpendicular to the wave fronts, and (2) the wavenumber by its length.

We can think of a wave front as a line along the crest of the wave.

The equation for the displacement associated with a plane sine wave (of unit amplitude) in three dimensions at some instant in time is

$$h(x, y, z) = \sin(\mathbf{k} \cdot \mathbf{x}) = \sin(k_x x + k_y y + k_z z). \tag{2.9}$$

Since wave fronts are lines or surfaces of constant phase, the equation defining a wave front is simply $\mathbf{k} \cdot \mathbf{x}$ = const.

In the two dimensional case we simply set $k_z = 0$.

Therefore, a wave front, or line of constant phase φ in two dimensions is defined by the equation

$$\mathbf{k} \cdot \mathbf{x} = k_x x + k_y y = \phi \quad \text{(two dimensions)}. \tag{2.10}$$

This can be easily solved for *y* to obtain the slope and intercept of the wave front in two dimensions.

As for one dimensional waves, the time evolution of the wave is obtained by adding a term -ωt to the phase of the wave.

In three dimensions the wave displacement as a function of both space and time is given by

$$h(x, y, z, t) = \sin(k_x x + k_y y + k_z z - \omega t). \tag{2.11}$$

The frequency depends in general on all three components of the wave vector.

The form of this function, $\omega = \omega(k_x,k_y,k_z)$, which as in the one dimensional case is called the *dispersion relation*, contains information about the physical behavior of the wave.

Some examples of dispersion relations for waves in two dimensions are as follows:

- Light waves in a vacuum in two dimensions obey

$$\omega = c(k_x^2 + k_y^2)^{1/2} \quad \text{(light)}, \tag{2.12}$$

  where c is the speed of light in a vacuum.

- Deep water ocean waves in two dimensions obey

$$\omega = g^{1/2}(k_x^2 + k_y^2)^{1/4} \quad \text{(ocean waves)}, \tag{2.13}$$

  where g is the strength of the Earth's gravitational field as before.

- Certain kinds of atmospheric waves confined to a vertical x-z plane called *gravity waves* (not to be confused with the gravitational waves of general relativity) obey

$$\omega = \frac{N k_x}{k_z} \quad \text{(gravity waves)}, \tag{2.14}$$

  where N is a constant with the dimensions of inverse time called the Brunt-Väisälä frequency.

Contour plots of these dispersion relations are plotted in the upper panels of figure 2.6.

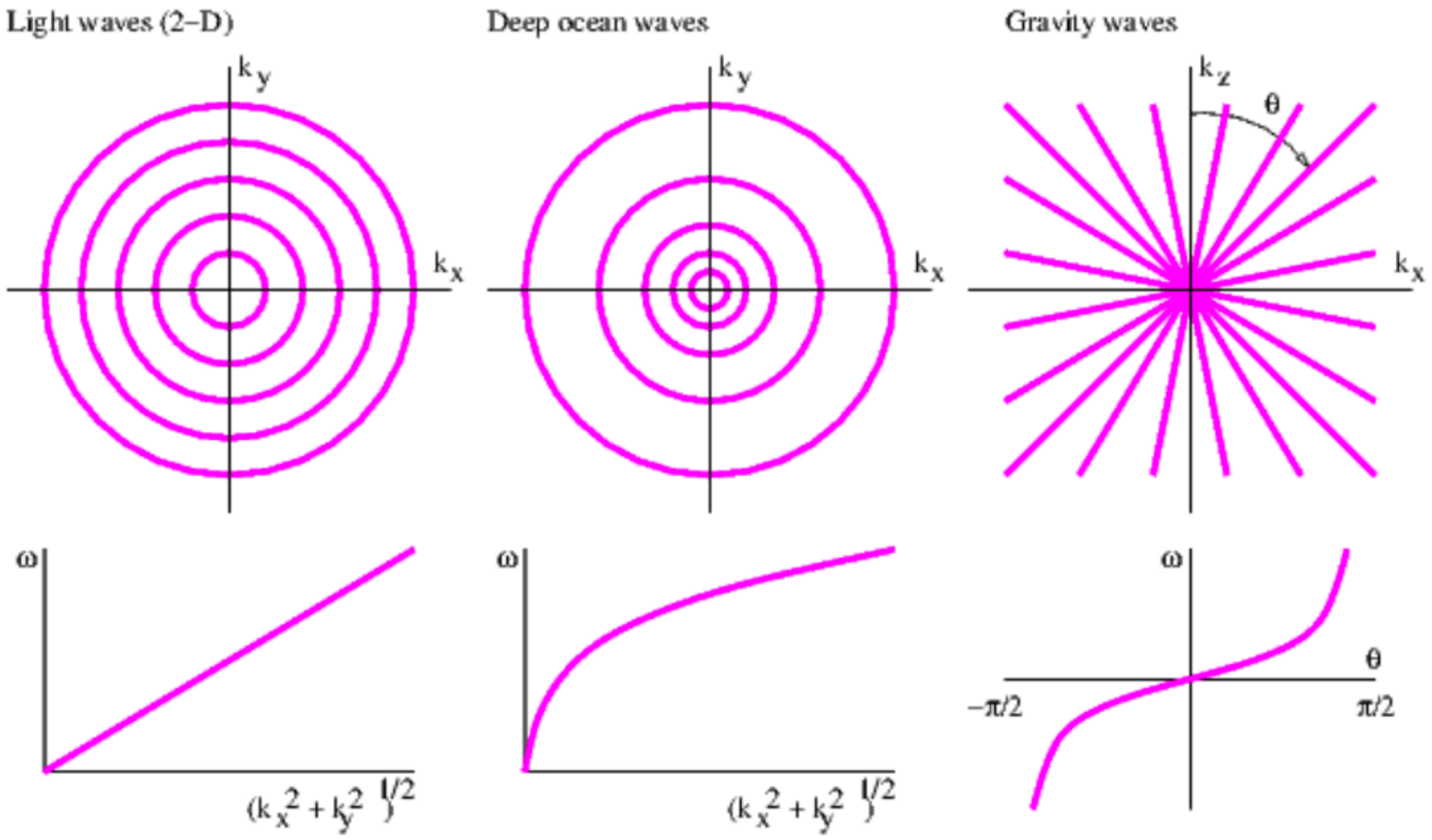


Figure 2.6: Contour plots of the dispersion relations for three kinds of waves in two dimensions. In the upper panels the curves show lines or contours along which the frequency ω takes on constant values. Contours are drawn for equally spaced values of ω. For light and ocean waves the frequency depends only on the magnitude of the wave vector, whereas for gravity waves it depends only on the wave vector's direction, as defined by the angle θ in the upper right panel. These dependences for each wave type are illustrated in the lower panels.

These plots are to be interpreted like topographic maps, where the lines represent contours of constant elevation.

In the case of figure 2.6, constant values of frequency are represented instead.

For simplicity, the actual values of frequency are not labeled on the contour plots, but are represented in the graphs in the lower panels.

This is possible because frequency depends only on wave vector magnitude $(k_x^2 + k_y^2)^{1/2}$ for the first two examples, and only on wave vector direction θ for the third.

2.3 Superposition of Plane Waves

We now study wave packets in two dimensions by asking what the superposition of two plane sine waves looks like.

If the two waves have different wave numbers, but their wave vectors point in the same direction, the results are identical to those presented in the previous chapter, except that the wave packets are indefinitely elongated without change in form in the direction perpendicular to the wave vector.

The wave packets produced in this case move in the direction of the wave vectors and thus appear to a stationary observer like a series of passing pulses with broad lateral extent.

Superimposing two plane waves which have the same frequency results in a stationary wave packet through which the individual wave fronts pass.

This wave packet is also elongated indefinitely in some direction, but the direction of elongation depends on the dispersion relation for the waves being considered.

These wave packets are in the form of steady beams, which guide the individual phase waves in some direction, but don't themselves change with time.

By superimposing multiple plane waves, all with the same frequency, one can actually produce a single stationary beam, just as one can produce an isolated pulse by superimposing multiple waves with wave vectors pointing in the same direction.

If the frequency of a wave depends on the magnitude of the wave vector, but not on its direction, the wave's dispersion relation is called isotropic; otherwise it is anisotropic.

In the isotropic case, two waves have the same frequency only if the lengths of their wave vectors, and hence their wavelengths, are the same.

The first two examples in figure 2.6 satisfy this condition, while the last example is anisotropic.

We now use the language of vectors to investigate the superposition of two plane waves with wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$:

$$h = \sin(\mathbf{k}_1 \cdot \mathbf{x} - \omega t) + \sin(\mathbf{k}_2 \cdot \mathbf{x} - \omega t). \tag{2.15}$$

Applying the trigonometric identity for the sine of the sum of two angles (as we have done previously), equation (2.15) can be reduced to

$$h = 2\sin(\mathbf{k}_0 \cdot \mathbf{x} - \omega t)\cos(\Delta\mathbf{k} \cdot x) \tag{2.16}$$

where

$$\mathbf{k}_0 = (\mathbf{k}_1 + \mathbf{k}_2)/2 \qquad \Delta\mathbf{k} = (\mathbf{k}_2 - \mathbf{k}_1)/2. \tag{2.17}$$

This is in the form of a sine wave moving in the $\mathbf{k}_0$ direction with phase speed $c_{phase} = \omega/|\mathbf{k}_0|$ and wavenumber $|\mathbf{k}_0|$, modulated in the $\Delta\mathbf{k}$ direction by a cosine function.

The lines of destructive interference are normal to $\Delta\mathbf{k}$.

The distance w between lines of destructive interference is the distance between successive zeros of the cosine function in equation (2.16), implying that $|\Delta\mathbf{k}|w = \pi$, which leads to

$$w = \pi/|\Delta\mathbf{k}|. \tag{2.18}$$

Thus, the smaller $|\Delta\mathbf{k}|$, the greater is the beam diameter.

2.3.1 Two Waves of Identical Wavelength

In this section we investigate the beams produced by superimposing isotropic waves of the same frequency.

Figure 2.7 illustrates what happens in such a superposition.

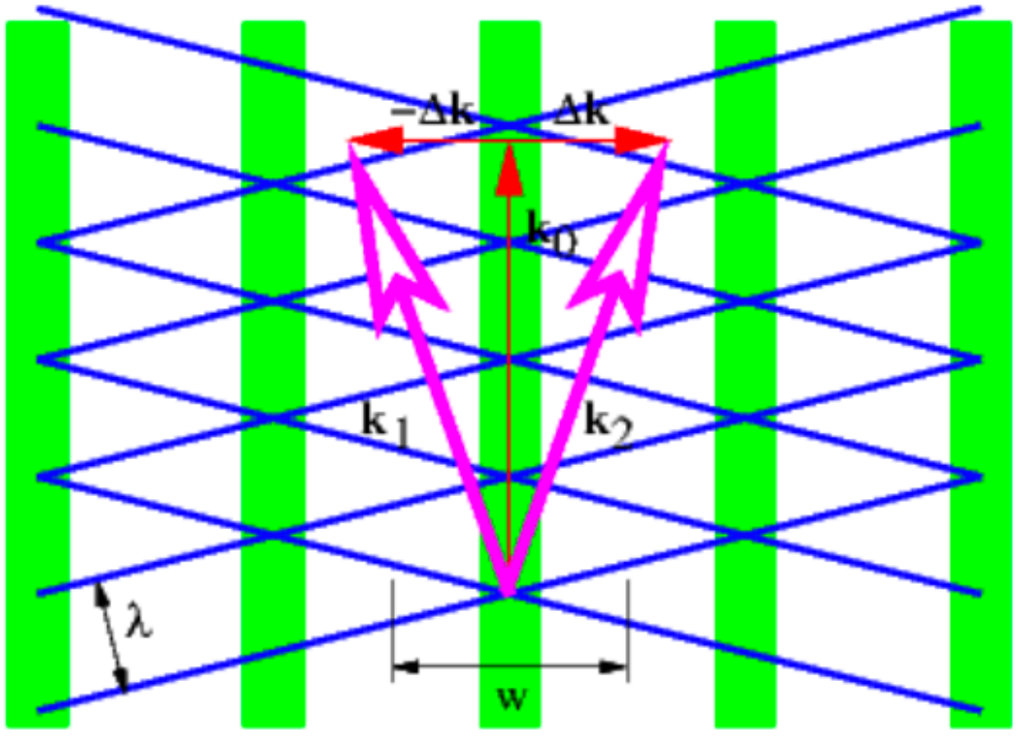


Figure 2.7: Wave fronts and wave vectors ($\mathbf{k}_1$ and $\mathbf{k}_2$) of two plane waves with the same wavelength but oriented in different directions. The vertical bands show regions of constructive interference where wave fronts coincide. The vertical regions in between have destructive interference, and hence define the lateral boundaries of the beams produced by the superposition. The quantities $\mathbf{k}_0$ and $\Delta\mathbf{k}$ are also shown.

Vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ of equal length give rise to a mean wave vector $\mathbf{k}_0$ and half the difference, $\Delta\mathbf{k}$.

As illustrated, the lines of constructive and destructive interference are perpendicular to $\Delta\mathbf{k}$.

Figure 2.8 shows a concrete example of the beams produced by superposition of two plane waves of equal wavelength oriented as in figure 2.7.

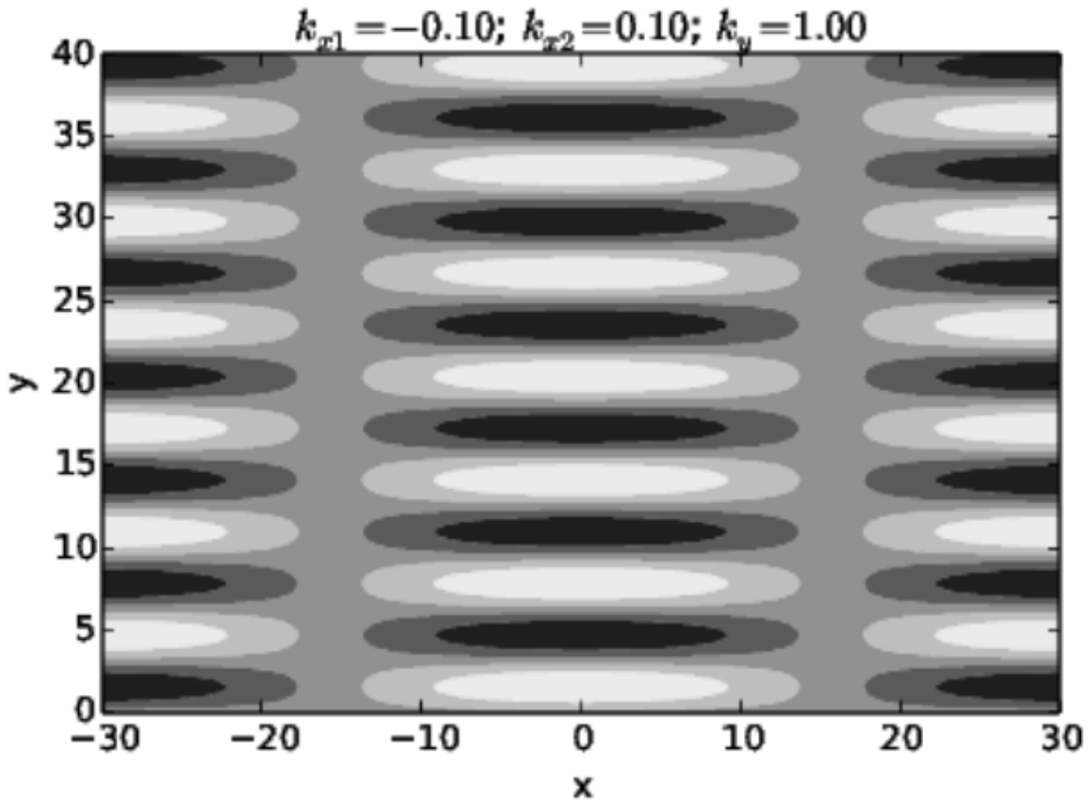


Figure 2.8: Example of beams produced by two plane waves with the same wavelength moving in different directions. The wave vectors of the two waves are $\mathbf{k} = (\pm 0.1, 1.0)$. Regions of positive displacement are lighter, while regions of negative displacement are darker.

The beams are aligned vertically, since $\Delta\mathbf{k}$ is horizontal, with the lines of destructive interference separating the beams located near $x = \pm 16$.

The transverse width of the beams of ≈ 32 satisfies equation (2.18) with $|\Delta\mathbf{k}| = 0.1$.

Each beam is made up of vertically propagating phase waves, with the crests and troughs indicated by the regions of white and black.

2.3.2 Two Waves of Differing Wavelength

In the third example of figure 2.6, the frequency of the wave depends only on the direction of the wave vector, independent of its magnitude, which is the reverse of the case for an isotropic dispersion relation.

In this highly anisotropic case, different plane waves with the same frequency have wave vectors which point in the same direction, but have different lengths.

More generally, one might have waves for which the frequency depends on *both* the direction and magnitude of the wave vector.

In this case, two different plane waves with the same frequency would typically have wave vectors which differ both in direction and magnitude.

Such an example is illustrated in figures 2.9 and 2.10.

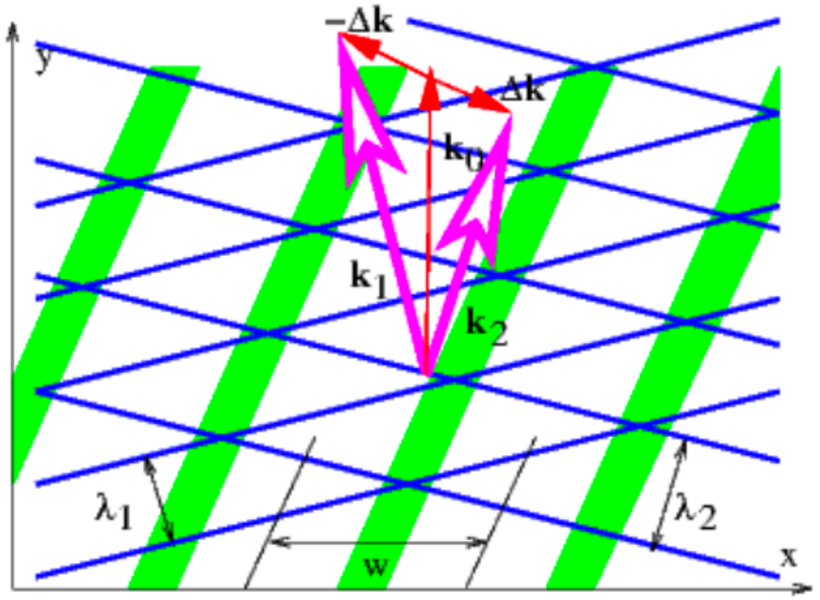


Figure 2.9: Wave fronts and wave vectors ($\mathbf{k}_1$ and $\mathbf{k}_2$) of two plane waves with different wavelengths oriented in different directions. The slanted bands show regions of constructive interference where wave fronts coincide. The slanted regions in between have destructive interference, and as mentioned previously, define the lateral limits of the beams produced by the superposition. The quantities $\mathbf{k}_0$ and $\Delta\mathbf{k}$ are also shown.

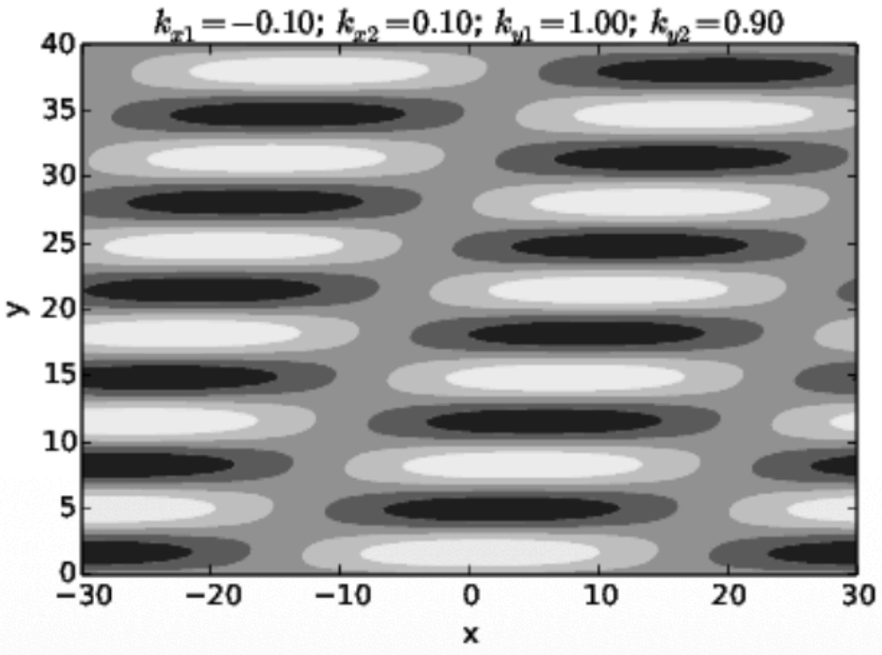


Figure 2.10: Example of beams produced by two plane waves with wave vectors differing in both direction and magnitude. The wave vectors of the two waves are $\mathbf{k}_1$ = (-0.1, 1.0) and $\mathbf{k}_2$ = (0.1, 0.9). Regions of positive displacement are lighter, while regions of negative displacement darker.

Figure 2.11 summarizes what we have learned about adding plane waves with the same frequency.

In general, the beam orientation (and the lines of constructive interference) are not perpendicular to the wave fronts.

This only occurs when the wave frequency is independent of wave vector direction.

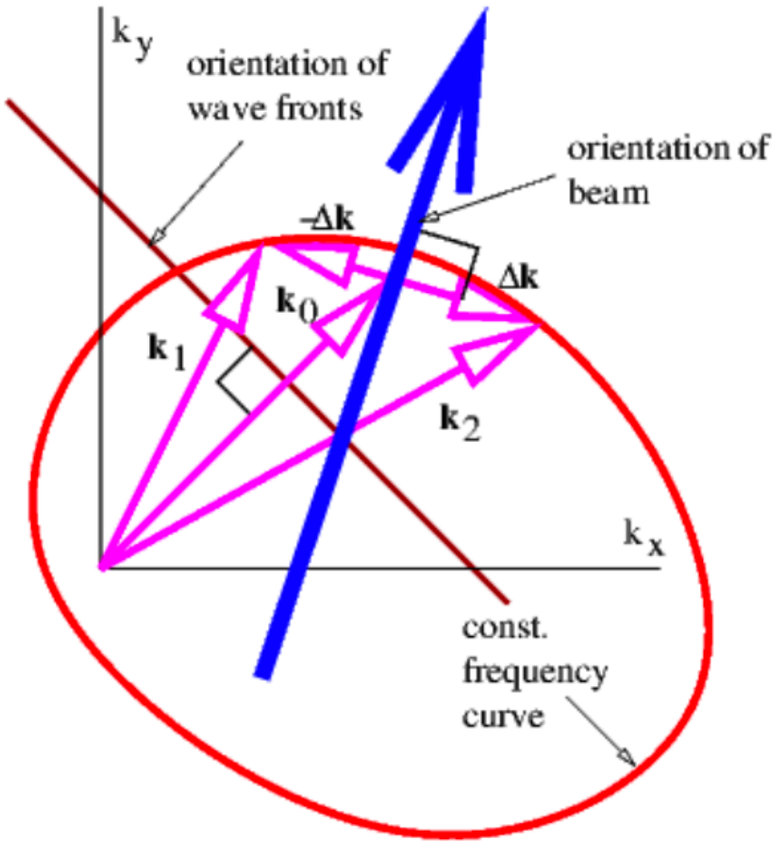


Figure 2.11: Illustration of factors entering the addition of two plane waves with the same frequency. The wave fronts are perpendicular to the vector average of the two wave vectors, $\mathbf{k}_0 = (\mathbf{k}_1 + \mathbf{k}_2)/\ 2$, while the lines of constructive interference, which define the beam orientation, are oriented perpendicular to the difference between these two vectors, $\Delta\mathbf{k} = (\mathbf{k}_1 - \mathbf{k}_1)/\ 2$.

2.3.3 Many Waves with the Same Wavelength

As with wave packets in one dimension, we can add together more than two waves to produce an isolated wave packet.

We will confine our attention here to the case of an isotropic dispersion relation in which all the wave vectors for a given frequency are of the same length.

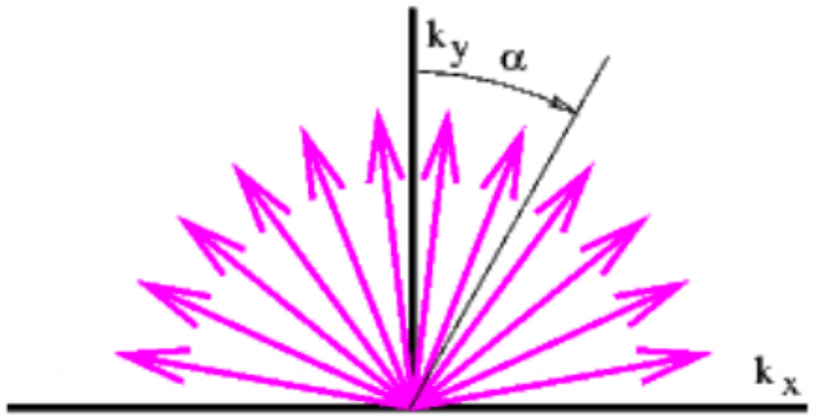


Figure 2.12: Illustration of wave vectors of plane waves which might be added together.

Figure 2.12 shows an example of this in which wave vectors of the same wavelength but different directions are added together.

Defining α_i as the angle of the ith wave vector clockwise from the vertical, as illustrated in figure 2.12, we could write the superposition of these waves at time $t = 0$ as

$$\begin{aligned} h &= \sum_i h_i \sin(k_{xi}x + k_{yi}y) \\ &= \sum_i h_i \sin[kx\sin(\alpha_i) + ky\cos(\alpha_i)] \end{aligned} \tag{2.19}$$

where we have assumed that $k_{xi} = k\sin(\alpha_i)$ and $k_{yi} = k\cos(\alpha_i)$.

The parameter $k = |\mathbf{k}|$ is the magnitude of the wave vector and is the same for all the waves.

Let us also assume in this example that the amplitude of each wave component decreases with increasing $|\alpha_i|$:

$$h_i = \exp[-(\alpha_i/\alpha_{max})^2]. \tag{2.20}$$

The exponential function decreases rapidly as its argument becomes more negative, and for practical purposes, only wave vectors with $|\alpha| \leq \alpha_{max}$ contribute significantly to the sum.

We call α_{max} the *spreading* angle.

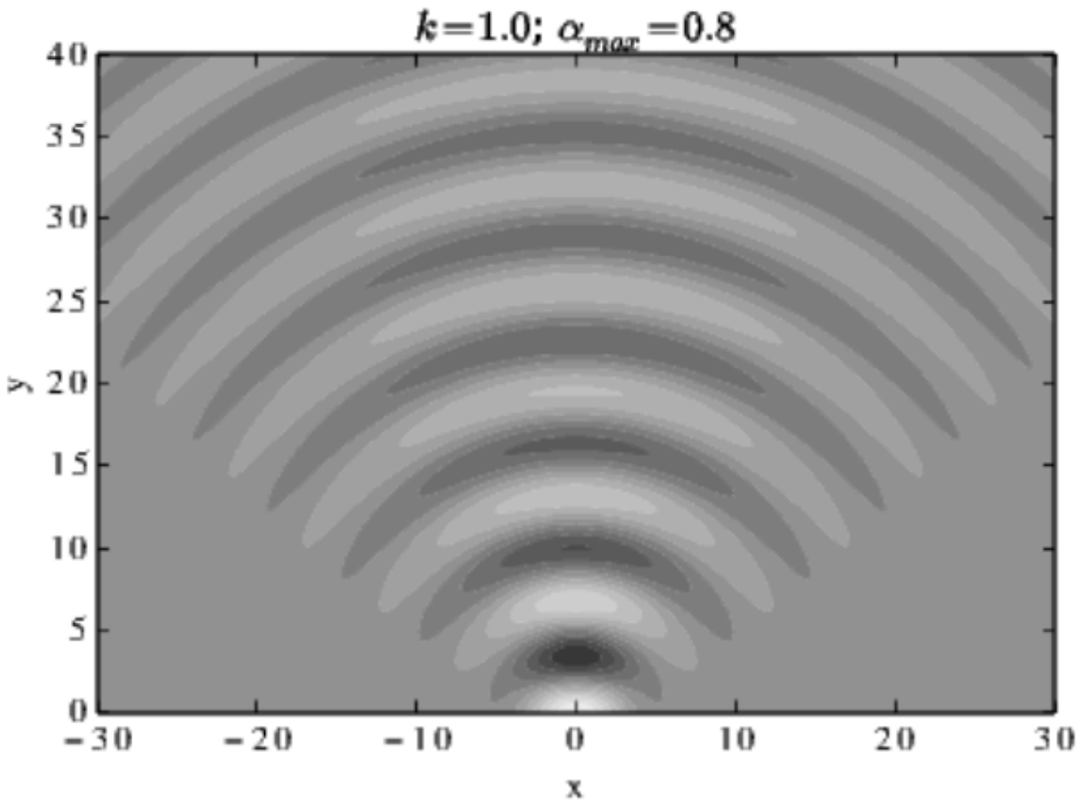


Figure 2.13: Plot of the displacement field $h(x,y)$ from equation (2.19) for $\alpha_{max} = 0.8$ and $k = 1$.

Figure 2.13 shows what $h(x,y)$ looks like when $\alpha_{max} = 0.8$ radians and $k = 1$.

Notice that for $y = 0$ the wave amplitude is only large for a small region in the range $-4 < x < 4$.

However, for $y > 0$ the wave spreads into a broad, semicircular pattern.

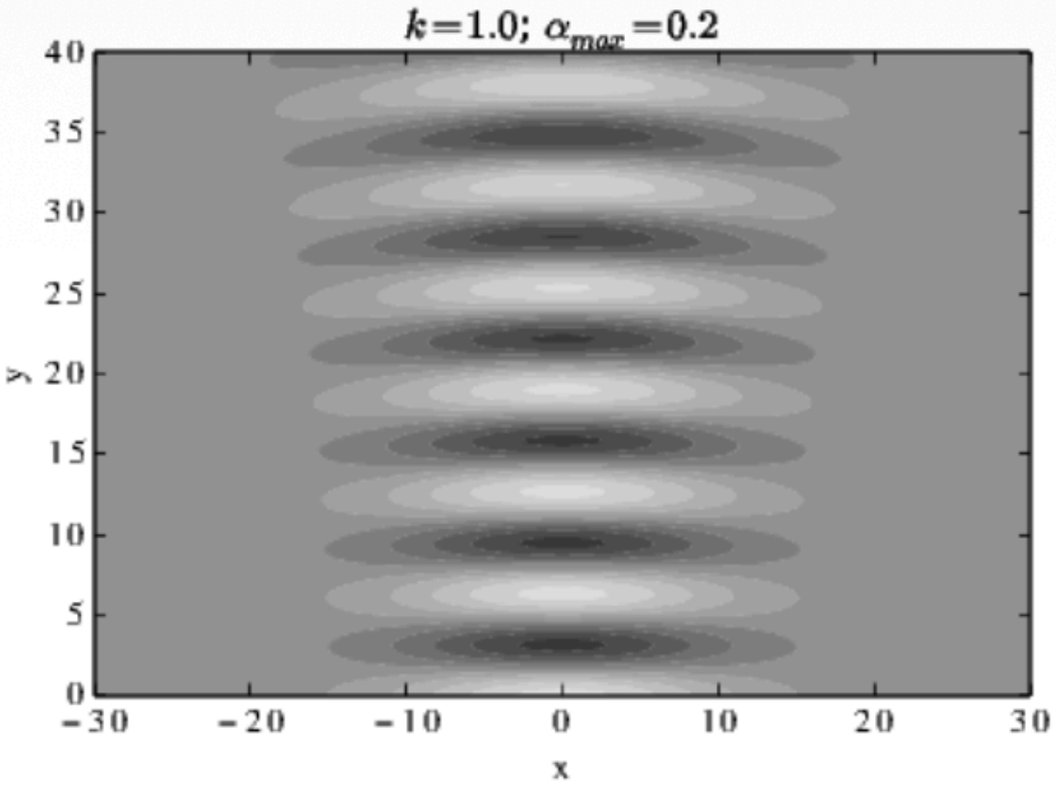


Figure 2.14: Plot of the displacement field $h(x,y)$ from equation (2.19) for $\alpha_{max} = 0.2$ and $k = 1$.

Figure 2.14 shows the computed pattern of $h(x,y)$ when the spreading angle $\alpha_{max} = 0.2$ radians.

The wave amplitude is large for a much broader range of x at $y = 0$ in this case, roughly $-12 < x < 12$.

On the other hand, the subsequent spread of the wave is much smaller than in the case of figure 2.13.

We conclude that a superposition of plane waves with wave vectors spread narrowly about a central wave vector which points in the y direction (as in figure 2.14) produces a beam which is initially broad in x but for which the breadth increases only slightly with increasing y.

However, a superposition of plane waves with wave vectors spread more broadly (as in figure 2.13) produces a beam which is initially narrow in x but which rapidly increases in width as y increases.

The relationship between the spreading angle α_{max} and the initial breadth of the beam is made more understandable by comparison with the results for the two-wave superposition discussed at the beginning of this section.

As indicated by equation (2.18), large values of k_x, and hence α, are associated with small wave packet dimensions in the x direction and vice versa.

The superposition of two waves doesn't capture the subsequent spread of the beam which occurs when many waves are superimposed, but it does lead to a rough quantitative relationship between α_{max} (which is just $\tan^{-1}(k_x/k_y)$ in the two wave case) and the initial breadth of the beam.

If we invoke the small angle approximation for $\alpha = \alpha_{max}$ so that $\alpha_{max} = \tan^{-1}(k_x/k_y) \approx k_x/k_y \approx k_x/k$, then $k_x \approx k\,\alpha_{max}$ and equation (2.18) can be written $w = \pi/k_x \approx \pi/(k\,\alpha_{max}) = \lambda/(2\,\alpha_{max})$.

Thus, we can find the approximate spreading angle from the wavelength of the wave λ and the initial breadth of the beam w:

$$\alpha_{max} \approx \lambda/(2w) \quad \text{(single slit spreading angle)}. \tag{2.21}$$

2.4 Diffraction Through a Single Slit

How does all of this apply to the passage of waves through a slit?

Imagine a plane wave of wavelength λ impingent on a barrier with a slit.

The barrier transforms the plane wave with infinite extent in the lateral direction into a beam with initial transverse dimensions equal to the width of the slit.

The subsequent development of the beam is illustrated in figures 2.13 and 2.14, and schematically in figure 2.15.

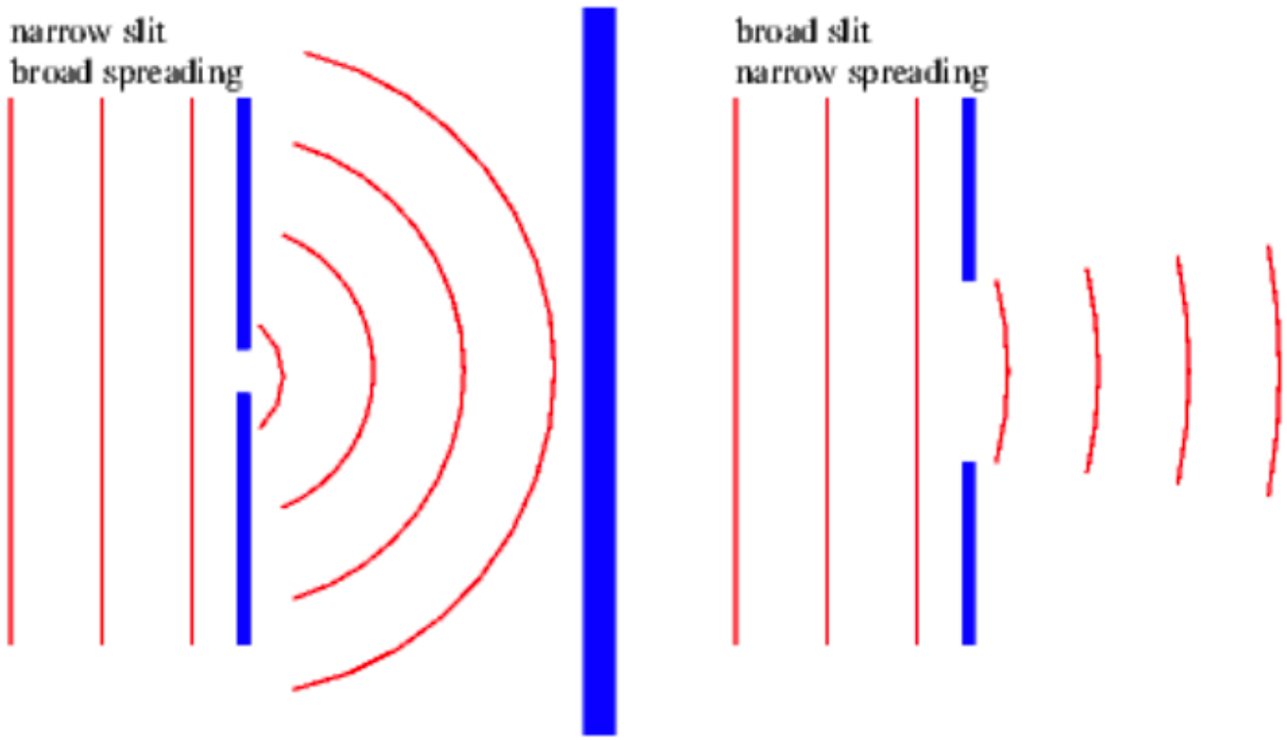


Figure 2.15: Schematic behavior when a plane wave impinges on a narrow slit and a broad slit.

In particular, if the slit width is comparable to the wavelength, the beam spreads broadly as in figure 2.13.

If the slit width is large compared to the wavelength, the beam doesn't spread as much, as figure 2.14 illustrates.

Equation (2.21) gives us an approximate quantitative result for the spreading angle if w is interpreted as the width of the slit.

One use of the above equation is in determining the maximum angular resolution of optical instruments such as telescopes.

The primary lens or mirror can be thought of as a rather large "slit" .

Light from a distant point source is essentially in the form of a plane wave when it arrives at the telescope.

However, the light passed by the telescope is no longer a plane wave, but is a beam with a tendency to spread.

The spreading angle α_{max} is given by equation (2.21), and the telescope cannot resolve objects with an angular separation less than α_{max}.

Replacing w with the diameter of the lens or mirror in equation (2.21) thus yields the telescope's angular resolution.

For instance, a moderate sized telescope with aperture 1 m observing red light with $\lambda \approx 6 \times 10^{-7}$ m has a maximum angular resolution of about 3×10^{-7} radians.

2.5 Two Slits

Let us now imagine a plane sine wave normally impingent on a screen with two narrow slits spaced by a distance d, as shown in figure 2.16.

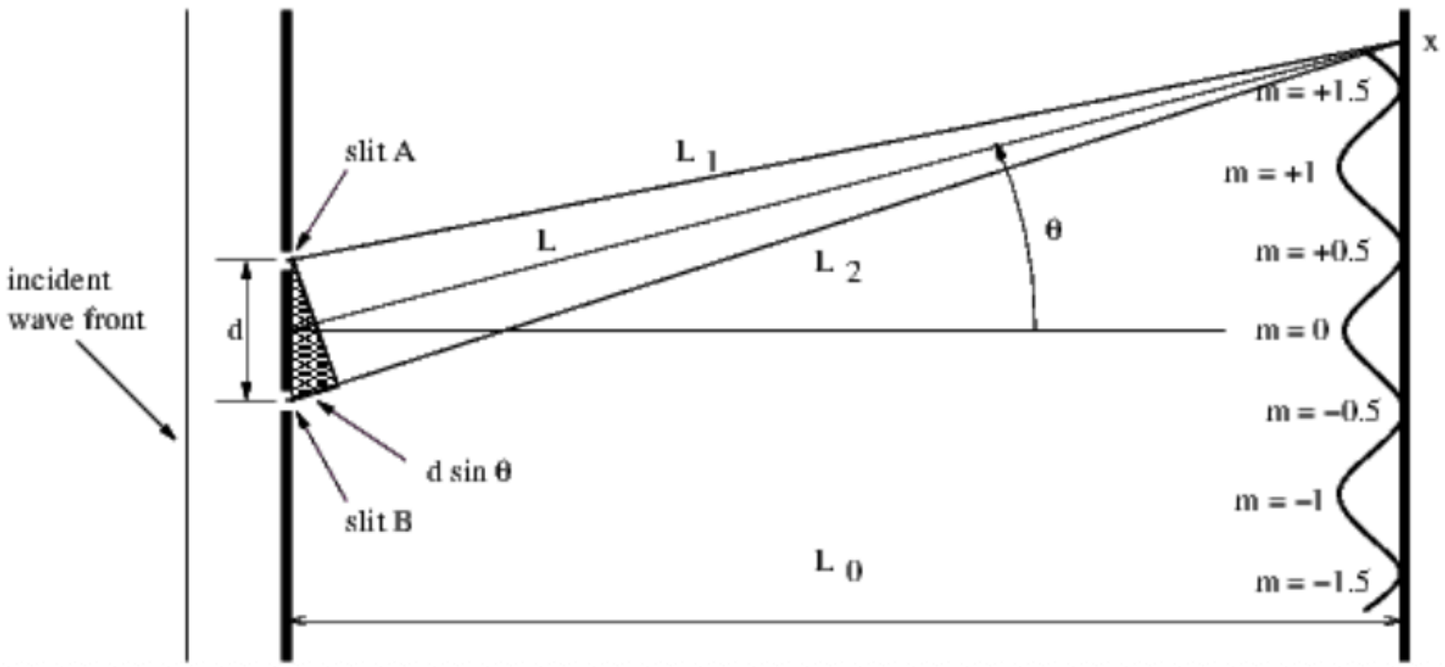


Figure 2.16: Definition sketch for the double slit. Light passing through slit B travels an extra

Since the slits are narrow relative to the wavelength of the wave impingent on them, the spreading angle of the beams is large and the diffraction pattern from each slit individually is a cylindrical wave spreading out in all directions, as illustrated in figure 2.13.

The cylindrical waves from the two slits interfere, resulting in oscillations in wave intensity at the screen on the right side of figure 2.16.

Constructive interference occurs when the difference in the paths traveled by the two waves from their originating slits to the screen, L_2 - L_1, is an integer multiple m of the wavelength λ: $L_2 - L_1 = m\lambda$.

If $L_0 \gg d$, the lines L_1 and L_2 are nearly parallel, which means that the narrow end of the dark triangle in figure 2.16 has an opening angle of θ.

Thus, the path difference between the beams from the two slits is $L_2 - L_1 = d \sin\theta$.

Substitution of this into the above equation shows that constructive interference occurs when

$$d\sin\theta = m\lambda, \quad m = 0, \pm 1, \pm 2, \ldots \qquad \text{(two slit interference)}. \tag{2.22}$$

Destructive interference occurs when m is an integer plus $1/2$.

The integer m is called the *interference order* and is the number of wavelengths by which the two paths differ.

2.6 Diffraction Gratings

Since the angular spacing $\Delta\theta$ of interference peaks in the two slit case depends on the wavelength of the incident wave, the two slit system can be used as a crude device to distinguish between the wavelengths of different components of a non-sinusoidal wave impingent on the slits.

However, if more slits are added, maintaining a uniform spacing d between slits, we obtain a more sophisticated device for distinguishing beam components.

This is called a *diffraction grating*.

Figures 2.17-2.19 show the intensity of the diffraction pattern as a function of position x on the display screen (see figure 2.16) for gratings with 2, 4, and 16 slits respectively, with the same slit spacing.

Notice how the interference peaks remain in the same place but increase in sharpness as the number of slits increases.

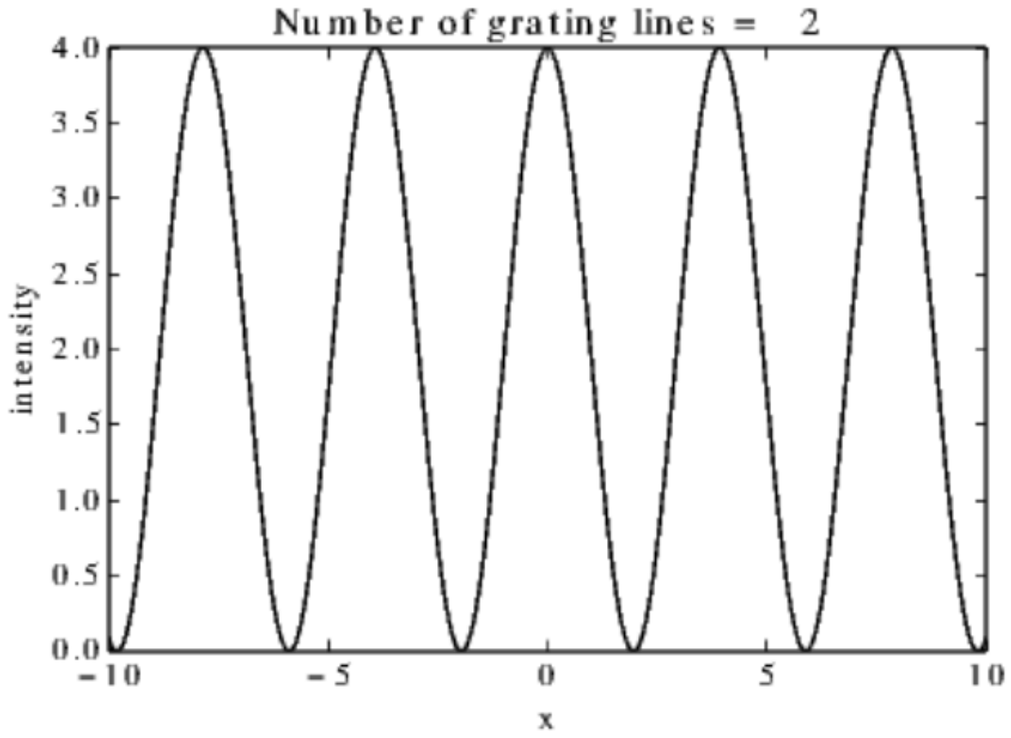


Figure 2.17: Intensity of interference pattern from a diffraction grating with 2 slits on the screen in figure 2.16. The position x on the screen is proportional to the angle θ in the small angle approximation.

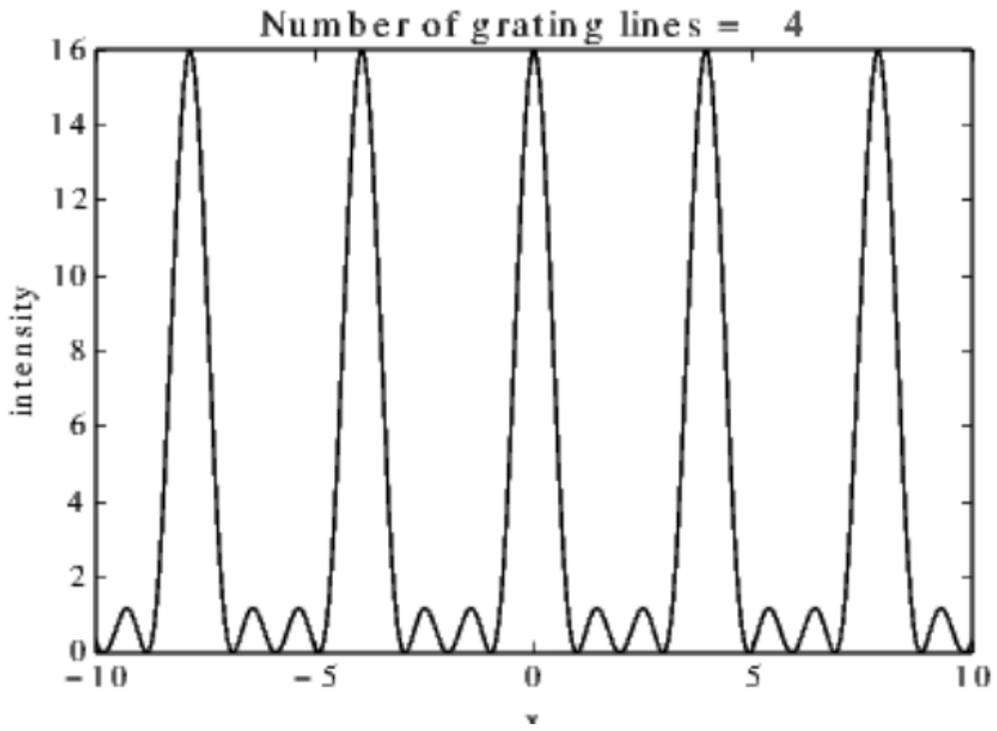


Figure 2.18: Intensity of interference pattern from a diffraction grating with 4 slits.

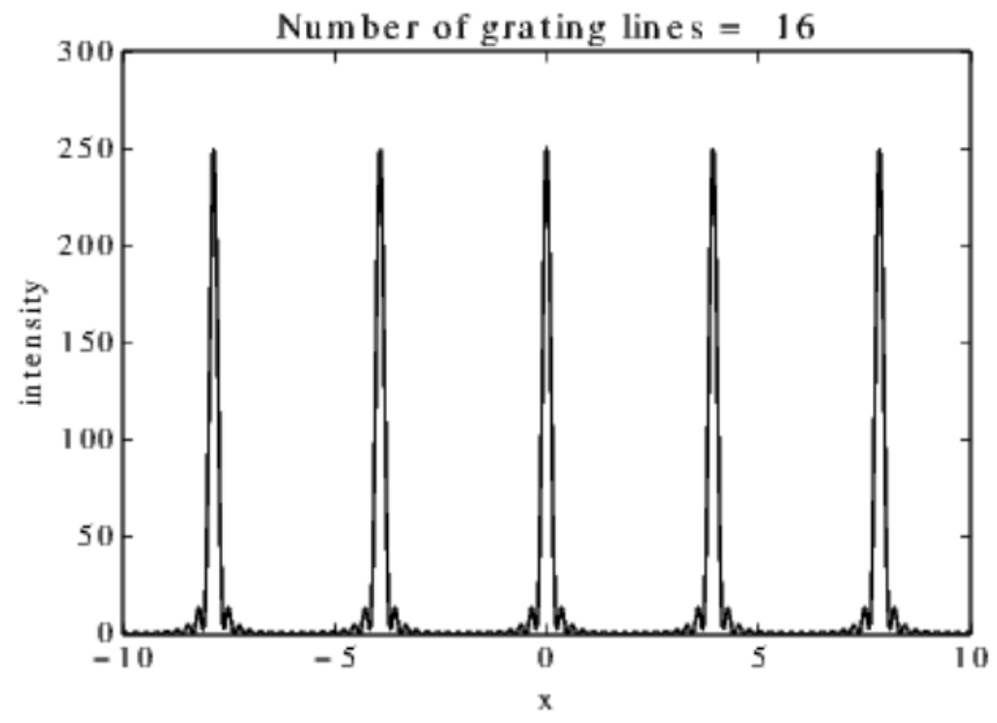


Figure 2.19: Intensity of interference pattern from a diffraction grating with 16 slits.

The width of the peaks is actually related to the overall width of the grating, $w = nd$, where n is the number of slits.

Thinking of this width as the dimension of a large single slit, the single slit equation, $\alpha_{max} = \lambda/(2w)$, tells us the angular width of the peaks.

Whereas the angular width of the interference peaks is governed by the single slit equation, their angular positions are governed by the two slit equation.

Let us assume for simplicity that $|\theta| \ll 1$ so that we can make the small angle approximation to the two slit equation, $m\lambda = d \sin\theta \approx d\theta$, and ask the following question: How different do two wavelengths differing by $\Delta\lambda$ have to be in order that the interference peaks from the two waves not overlap?

In order for the peaks to be distinguishable, they should be separated in θ by an angle $\Delta\theta = m\Delta\lambda/d$, which is greater than the angular width of each peak, α_{max}:

$$\Delta\theta > \alpha_{max}. \tag{2.23}$$

Substituting in the above expressions for $\Delta\theta$ and α_{max} and solving for $\Delta\lambda$, we get $\Delta\lambda > \lambda/(2mn)$, where λ is the average of the two wavelengths and $n = w/d$ is the number of slits in the diffraction grating.

Thus, the fractional difference between wavelengths which can be distinguished by a diffraction grating depends solely on the interference order m and the number of slits n in the grating:

$$\frac{\Delta\lambda}{\lambda} > \frac{1}{2mn}. \tag{2.24}$$

Chapter 3

Geometrical Optics

As was shown previously, when a plane wave is impingent on an aperture which has dimensions much greater than the wavelength of the wave, diffraction effects are minimal and a segment of the plane wave passes through the aperture essentially unaltered.

This plane wave segment can be thought of as a wave packet, called a *beam* or *ray*, consisting of a superposition of wave vectors very close in direction and magnitude to the central wave vector of the wave packet.

In most cases the ray simply moves in the direction defined by the central wave vector, i. e., normal to the orientation of the wave fronts.

However, this is not true when the medium through which the light propagates is optically anisotropic, i. e., light traveling in different directions moves at different phase speeds.

An example of such a medium is a calcite crystal.

In the anisotropic case, the orientation of the ray can be determined once the dispersion relation for the waves in question is known, by using the techniques developed in the previous chapter.

If light moves through some apparatus in which all apertures are much greater in dimension than the wavelength of light, then we can use the above rule to follow rays of light through the apparatus.

This is called the geometrical optics approximation.

3.1 Reflection and Refraction

Most of what we need to know about geometrical optics can be summarized in two rules, the laws of reflection and refraction.

These rules may both be inferred by considering what happens when a plane wave segment impinges on a flat surface.

If the surface is polished metal, the wave is *reflected*, whereas if the surface is an interface between two transparent media with differing indices of refraction, the wave is partially reflected and partially *refracted*.

Reflection means that the wave is turned back into the half-space from which it came, while refraction means that it passes through the interface, acquiring a different direction of motion from that which it had before reaching the interface.

Figure 3.1 shows the wave vector and wave front of a wave being reflected from a plane mirror.

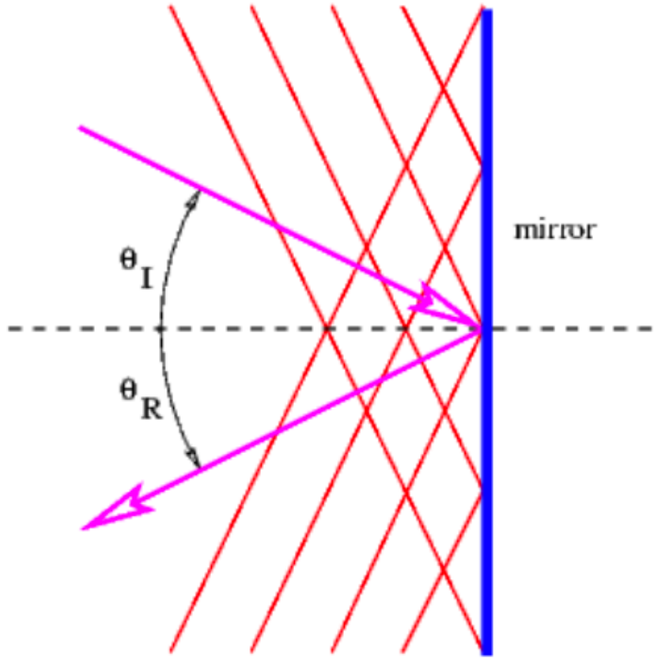


Figure 3.1: Sketch showing the reflection of a wave from a plane mirror. The law of reflection states that $\theta_I = \theta_R$.

The angles of incidence, θ_I, and reflection, θ_R, are defined to be the angles between the incoming and outgoing wave vectors respectively and the line normal to the mirror.

The law of reflection states that $\theta_R = \theta_I$.

This is a consequence of the need for the incoming and outgoing wave fronts to be in phase with each other all along the mirror surface.

This plus the equality of the incoming and outgoing wavelengths is sufficient to insure the above result.

Refraction, as illustrated in figure 3.2, is slightly more complicated.

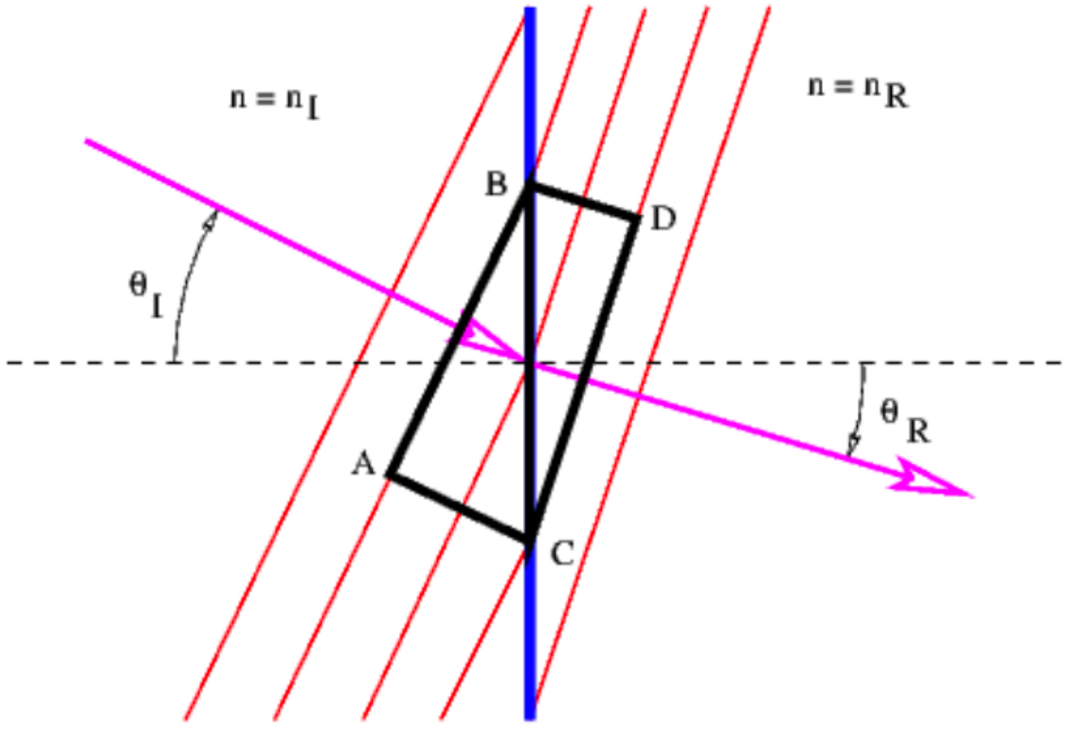


Figure 3.2: Sketch showing the refraction of a wave from an interface between two dielectric media with $n_2 > n_1$.

Since $n_R > n_I$, the speed of light in the right-hand medium is less than in the left-hand medium.

Recall that the speed of light in a medium with refractive index n is $c_{medium} = c_{vac}/n$.

The frequency of the wave packet doesn't change as it passes through the interface, so the wavelength of the light on the right side is less than the wavelength on the left side.

Let us examine the triangle ABC in figure 3.2.

The side AC is equal to the side BC times $\sin(\theta_I)$.

However, AC is also equal to $2\lambda_I$, or twice the wavelength of the wave to the left of the interface.

Similar reasoning shows that $2\lambda_R$, twice the wavelength to the right of the interface, equals BC times $\sin(\theta_R)$.

Since the interval BC is common to triangles ABC and DBC, we easily see that

$$\frac{\lambda_I}{\lambda_R} = \frac{\sin(\theta_I)}{\sin(\theta_R)}. \tag{3.1}$$

Since $\lambda_I = c_I T = c_{vac} T / n_I$ and $\lambda_R = c_R T = c_{vac} T / n_R$ where c_I and c_R are the wave speeds to the left and right of the interface, c_{vac} is the speed of light in a vacuum, and T is the (common) period, we can easily recast the above equation in the form

$$n_I \sin(\theta_I) = n_R \sin(\theta_R). \tag{3.2}$$

This is called *Snell's law*, and it governs how a ray of light bends as it passes through a discontinuity in the index of refraction. The angle θ_I is called the incident angle and θ_R is called the refracted angle.

Notice that these angles are measured from the normal to the surface, not the tangent.

3.2 Total Internal Reflection

When light passes from a medium of lesser index of refraction to one with greater index of refraction, Snell's law indicates that the ray bends *toward* the normal to the interface.

The reverse occurs when the passage is in the other direction.

In this latter circumstance a special situation arises when Snell's law predicts a value for the sine of the refracted angle greater than one.

This is physically untenable.

What actually happens is that the incident wave is *reflected* from the interface.

This phenomenon is called *total internal reflection*.

The minimum incident angle for which total internal reflection occurs is obtained by substituting $\theta_R = \pi/2$ into equation (3.2), resulting in

$$\sin(\theta_I) = n_R/n_I \quad \text{(total internal reflection)}. \tag{3.3}$$

3.3 Anisotropic Media

Notice that Snell's law makes the implicit assumption that rays of light move in the direction of the light's wave vector, i. e., normal to the wave fronts.

As the analysis in the previous chapter makes clear, this is valid only when the optical medium is isotropic, i. e., the wave frequency depends only on the magnitude of the wave vector, not on its direction.

Certain kinds of crystals, such as those made of calcite, are not isotropic — the speed of light in such crystals, and hence the wave frequency, depends on the orientation of the wave vector.

As an example, the angular frequency in an anisotropic medium might take the form

$$\omega = \left[\frac{c_1^2 (k_x + k_y)^2}{2} + \frac{c_2^2 (k_x - k_y)^2}{2} \right]^{1/2}, \tag{3.4}$$

where c_1 is the speed of light for waves in which $k_y = k_x$, and c_2 is its speed when $k_y = -k_x$.

Figure 3.3 shows an example in which a ray hits a calcite crystal oriented so that constant frequency contours are as specified in equation (3.4).

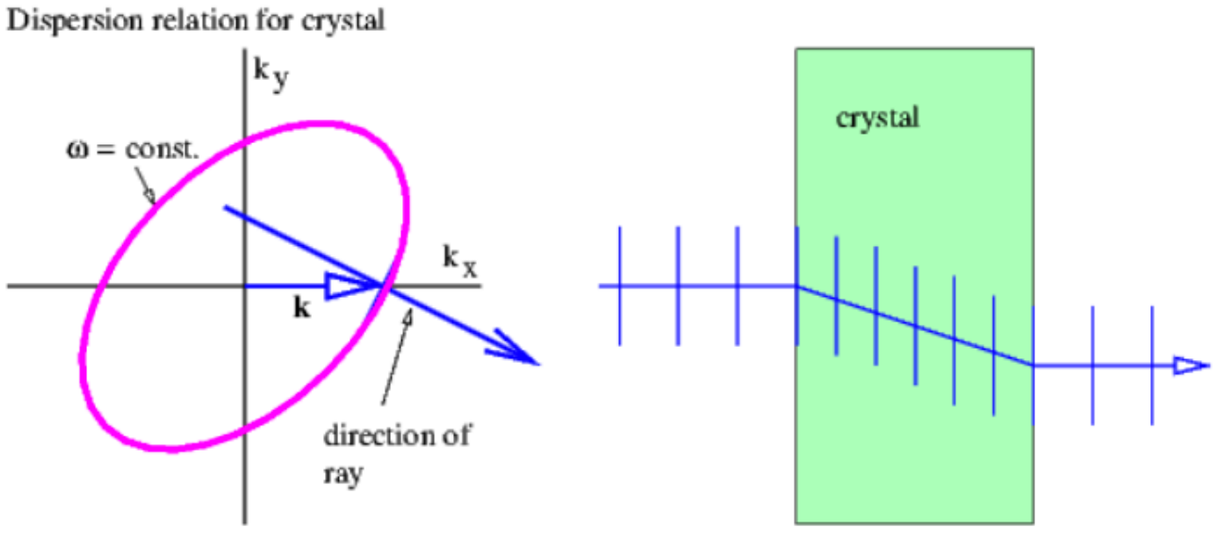


Figure 3.3: The right panel shows the fate of a light ray normally incident on the face of a properly cut calcite crystal. The anisotropic dispersion relation which gives rise to this behavior is shown in the left panel.

The wave vector is oriented normal to the surface of the crystal, so that wave fronts are parallel to this surface.

Upon entering the crystal, the wave front orientation must stay the same to preserve phase continuity at the surface.

However, due to the anisotropy of the dispersion relation for light in the crystal, the ray direction changes as shown in the right panel.

This behavior is clearly inconsistent with the usual version of Snell's law!

It is possible to extend Snell's law to the anisotropic case.

However, we will not present this here.

The following discussions of optical instruments will always assume that isotropic optical media are used.

3.4 Thin Lens Equation and Optical Instruments

Given the laws of reflection and refraction, one can see in principle how the passage of light through an optical instrument could be traced.

For each of a number of initial rays, the change in the direction of the ray at each mirror surface or refractive index interface can be calculated.

Between these points, the rays traces out a straight line.

Though simple in conception, this procedure can be quite complex in practice.

However, the procedure simplifies if a number of approximations, collectively called the *thin lens approximation*, are valid.

We begin with the calculation of the bending of a ray of light as it passes through a prism, as illustrated in figure 3.4.

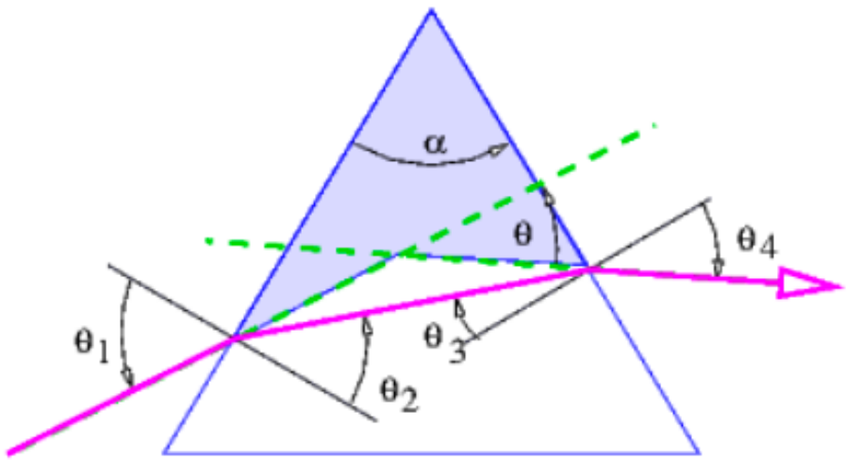


Figure 3.4: Bending of a ray of light as it passes through a prism.

The pieces of information needed to find θ, the angle through which the ray is deflected, are as follows: the geometry of the triangle defined by the entry and exit points of the ray and the upper vertex of the prism leads to

$$\alpha + (\pi/2 - \theta_2) + (\pi/2 - \theta_3) = \pi, \tag{3.5}$$

which simplifies to

$$\alpha = \theta_2 + \theta_3. \tag{3.6}$$

Snell's law at the entrance and exit points of the ray tell us that

$$n = \frac{\sin(\theta_1)}{\sin(\theta_2)} \qquad n = \frac{\sin(\theta_4)}{\sin(\theta_3)}, \tag{3.7}$$

where n is the index of refraction of the prism.

The index of refraction of the surroundings is assumed to be unity.

One can also infer that

$$\theta = \theta_1 + \theta_4 - \alpha. \tag{3.8}$$

This comes from the fact that the the sum of the internal angles of the shaded quadrangle in figure 3.4 is

$$(\pi/2 - \theta_1) + \alpha + (\pi/2 - \theta_4) + (\pi + \theta) = 2\pi. \tag{3.9}$$

Combining equations (3.6), (3.7), and (3.8) allows the ray deflection θ to be determined in terms of θ_1 and α, but the resulting expression is very messy.

However, great simplification occurs if the following conditions are met:

- The angle $\alpha \ll 1$.
- The angles $\theta_1, \theta_2, \theta_3, \theta_4 \ll 1$.

With these approximations it is easy to show that

$$\theta = \alpha(n-1) \quad \text{(small angles)}. \tag{3.10}$$

Generally speaking, lenses and mirrors in optical instruments have curved rather than flat surfaces.

However, we can still use the laws for reflection and refraction by plane surfaces as long as the segment of the surface on which the wave packet impinges is not curved very much on the scale of the wave packet dimensions.

This condition is easy to satisfy with light impinging on ordinary optical instruments.

In this case, the deflection of a ray of light is given by equation (3.10) if α is defined as the intersection of the tangent lines to the entry and exit points of the ray, as illustrated in figure 3.5.

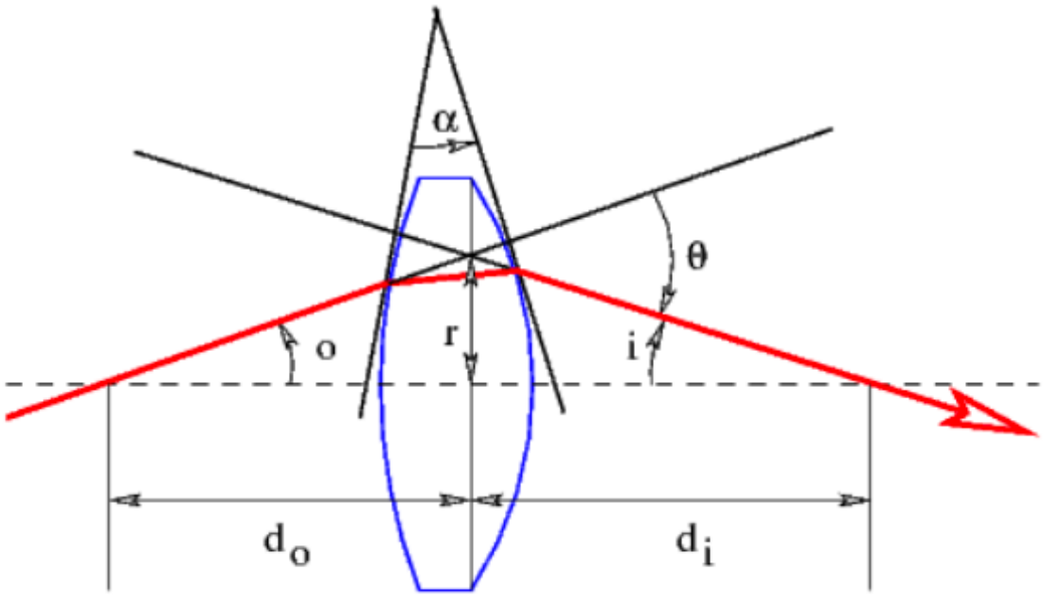


Figure 3.5: Light ray undergoing deflection through an angle θ by a lens. The angle α is the angle between the tangents to the entry and exit points of the ray on the lens.

A *positive lens* is thicker in the center than at the edges.

The angle α between the tangent lines to the two surfaces of the lens at a distance r from the central axis takes the form $\alpha = Cr$, where C is a constant.

The deflection angle of a beam hitting the lens a distance r from the center is therefore $\theta = Cr(n - 1)$, as indicated in figure 3.5.

The angles o and i sum to the deflection angle: $o + i = \theta = Cr(n - 1)$.

However, to the extent that the small angle approximation holds, $o = r/d_o$ and $i = r/d_i$ where d_o is the distance to the object and d_i is the distance to the image of the object.

Putting these equations together and cancelling the r results in the *thin lens equation*:

$$\frac{1}{d_o} + \frac{1}{d_i} = C(n-1) \equiv \frac{1}{f}. \qquad (3.11)$$

The quantity f is called the *focal length* of the lens.

Notice that $f = d_i$ if the object is very far from the lens, i. e., if d_o is extremely large.

Figure 3.6 shows how a positive lens makes an image.

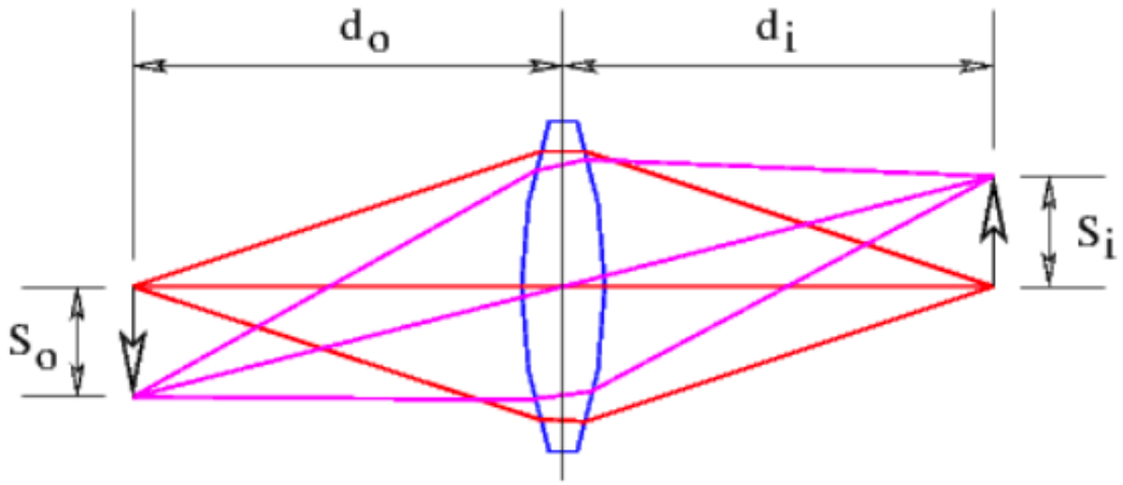


Figure 3.6: A positive lens producing an image on the right of the arrow on the left.

The image is produced by all of the light from each point on the object falling on a corresponding point in the image.

If the arrow on the left is an illuminated object, an *image* of the arrow will appear at right if the light coming from the lens is allowed to fall on a piece of paper or a ground glass screen.

The size of the object S_o and the size of the image S_i are related by simple geometry to the distances of the object and the image from the lens:

$$\frac{S_i}{S_o} = \frac{d_i}{d_o}. \tag{3.12}$$

Notice that a positive lens inverts the image.

An image will be produced to the right of the lens only if $d_o > f$.

If $d_o < f$, the lens is unable to converge the rays from the image to a point, as is seen in figure 3.7.

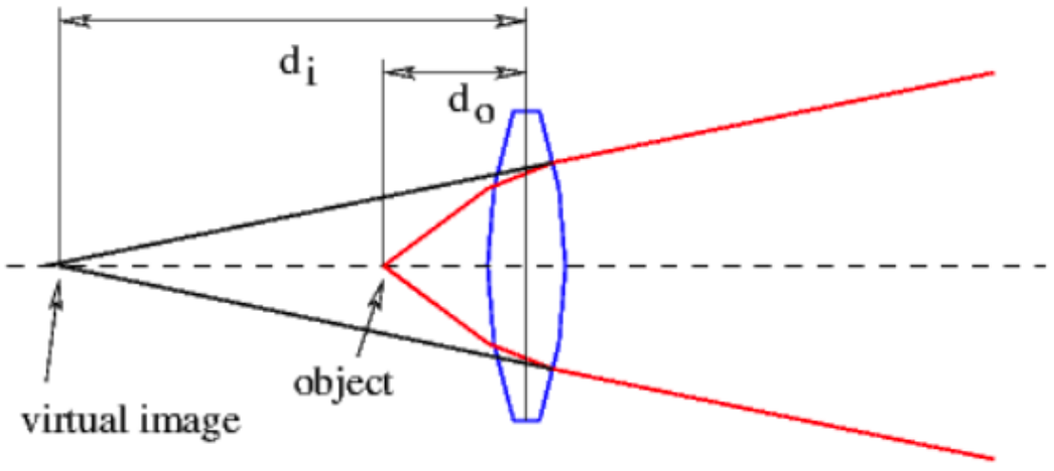


Figure 3.7: Production of a virtual image by a positive lens.

However, in this case the backward extension of the rays converge at a point called a *virtual image*, which in the case of a positive lens is always farther away from the lens than the object.

The image is called virtual because it does not appear on a ground glass screen placed at this point.

Unlike the real image seen in figure 3.6, the virtual image is not inverted.

The thin lens equation still applies if the distance from the lens to the image is taken to be negative.

A *negative lens* is thinner in the center than at the edges and produces only virtual images.

As seen in figure 3.8, the virtual image produced by a negative lens is closer to the lens than is the object.

Again, the thin lens equation is still valid, but both the distance from the image to the lens and the focal length must be taken as negative.

Only the distance to the object remains positive.

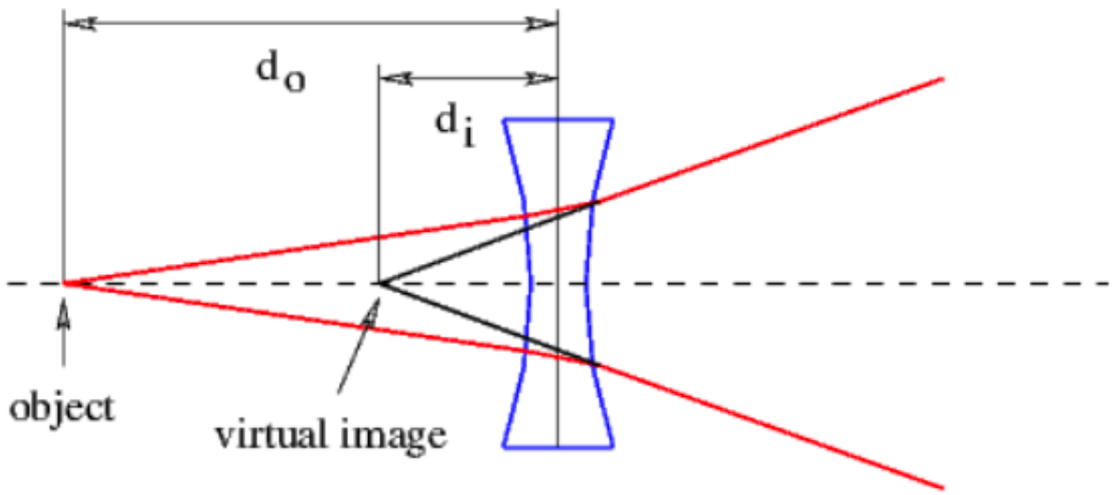


Figure 3.8: Production of a virtual image by a negative lens.

Curved mirrors also produce images in a manner similar to a lens, as shown in figure 3.9.

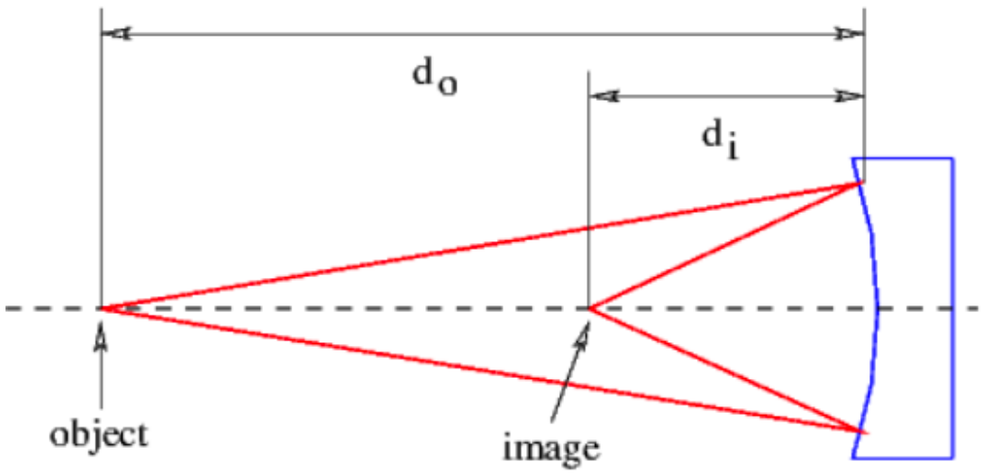


Figure 3.9: Production of a real image by a concave mirror.

A concave mirror, as seen in this figure, works in analogy to a positive lens, producing a real or a virtual image depending on whether the object is farther from or closer to the mirror than the mirror's focal length.

A convex mirror acts like a negative lens, always producing a virtual image.

The thin lens equation works in both cases as long as the angles are small.

3.5 Fermat's Principle

An alternate approach to geometrical optics can be developed from Fermat's principle.

This principle states (in its simplest form) that light waves of a given frequency traverse the path between two points which takes the least time.

The most obvious example of this is the passage of light through a homogeneous medium in which the speed of light doesn't change with position.

In this case the shortest time corresponds to the shortest distance between the points, which, as we all know, is a straight line.

Thus, Fermat's principle is consistent with light traveling in a straight line in a homogeneous medium.

Fermat's principle can also be used to derive the laws of reflection and refraction.

For instance, figure 3.10 shows a candidate ray for reflection in which the angles of incidence and reflection are not equal.

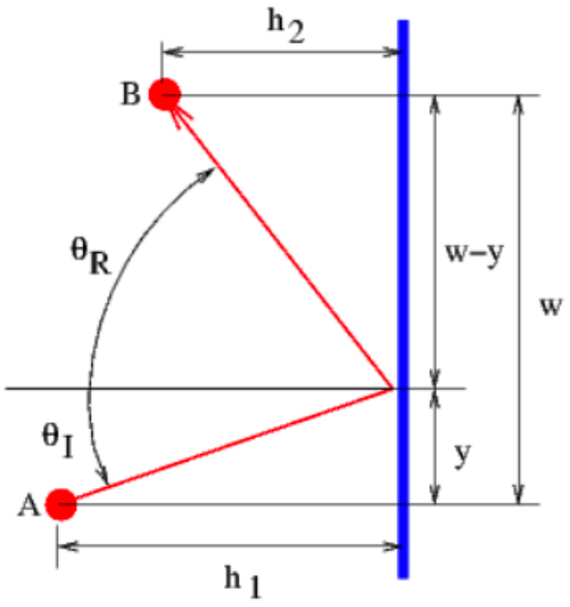


Figure 3.10: Definition sketch for deriving the law of reflection from Fermat's principle. θ_I is the angle of incidence and θ_R the angle of reflection as in figure 3.1.

The time required for the light to go from point A to point B is

$$t = ([h_1^2 + y^2]^{1/2} + [h_2^2 + (w - y)^2]^{1/2})/c \tag{3.13}$$

where c is the speed of light.

We find the minimum time by differentiating t with respect to y and setting the result to zero, with the result that

$$\frac{y}{[h_1^2 + y^2]^{1/2}} = \frac{w - y}{[h_2^2 + (w - y)^2]^{1/2}}. \tag{3.14}$$

However, we note that the left side of this equation is simply sin θ_I, while the right side is sin θ_R, so that the minimum time condition reduces to sin θ_I = sin θ_R or $\theta_I = \theta_R$, which is the law of reflection.

A similar analysis may be done to derive Snell's law of refraction.

The speed of light in a medium with refractive index n is c/n, where c is its speed in a vacuum.

Thus, the time required for light to go some distance in such a medium is n times the time light takes to go the same distance in a vacuum.

Referring to figure 3.11, the time required for light to go from A to B becomes

$$t = ([h_1^2 + y^2]^{1/2} + n[h_2^2 + (w - y)^2]^{1/2})/c. \tag{3.15}$$

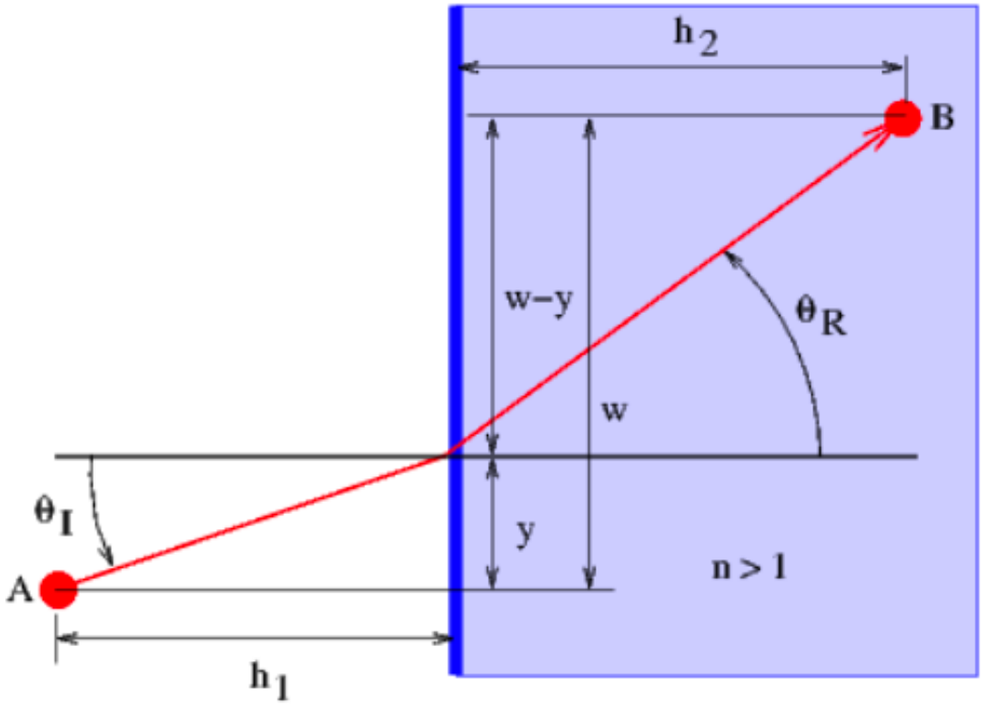


Figure 3.11: Definition sketch for deriving Snell's law of refraction from Fermat's principle. The shaded area has index of refraction $n > 1$.

This results in the condition

$$\sin\theta_I = n \sin\theta_R \tag{3.16}$$

where θ_R is now the refracted angle.

We recognize this result as Snell's law.

Notice that the reflection case illustrates a point about Fermat's principle: The minimum time may actually be a *local* rather than a global minimum — after all, in figure 3.10, the global minimum distance from A to B is still just a straight line between the two points!

In fact, light starting from point A will reach point B by *both* routes — the direct route and the reflected route.

It turns out that trajectories allowed by Fermat's principle don't strictly have to be minimum time trajectories.

They can also be *maximum* time trajectories, as illustrated in figure 3.12.

In this case light emitted at point O can be reflected back to point O from four points on the mirror, A, B, C, and D.

The trajectories O-A-O and O-C-O are minimum time trajectories while O-B-O and O-D-O are maximum time trajectories.

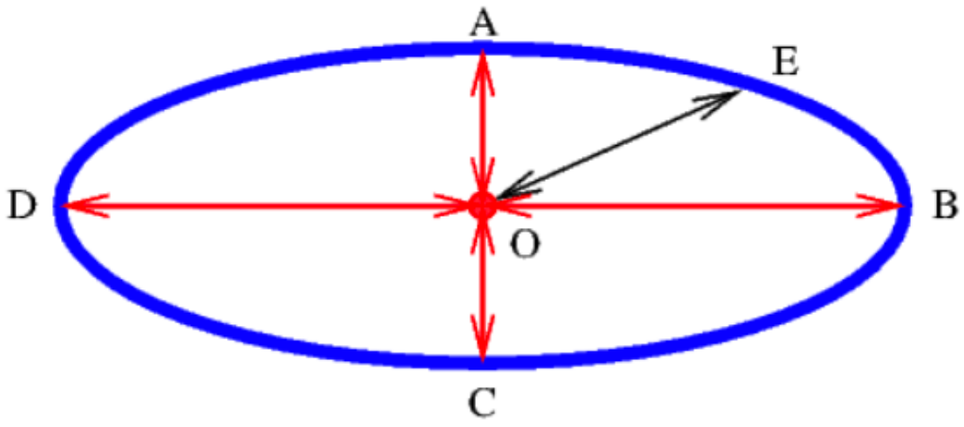


Figure 3.12: Ellipsoidal mirror showing minimum and maximum time rays from the center of the ellipsoid to the mirror surface and back again.

Fermat's principle seems rather mysterious.

However, the American physicist Richard Feynman made sense out of it by invoking an even more fundamental principle, as we now see.

If a light ray originates at point O in figure 3.12, reflects off of the ellipsoidal mirror surface at point A, and returns to point O, the elapsed time isn't much different from that experienced by a ray which reflects off the mirror a slight distance from point A and returns to O.

This is because at point A the beam from point O is perpendicular to the tangent to the surface of the mirror at point A.

In contrast, the time experienced by a ray going from point O to point E and back would differ by a much greater amount than the time experienced by a ray reflecting off the mirror a slight distance from point E.

This is because the tangent to the mirror surface at point E is not perpendicular to the beam from point O.

Technically, the change in the round trip time varies linearly with the deviation in the reflection point from point E, but quadratically with the deviation from point A.

If this deviation is small in the first place, then the change in the round trip time will be much smaller for the quadratic case than for the linear case.

It seems odd that we would speak of a beam reflecting back to point O if it hit the mirror at any point except A, B, C, or D, due to the requirements of the law of reflection.

However, recall that the law of reflection itself depends on Fermat's principle, so we cannot assume the validity of that law in this investigation.

Feynman postulated that light rays explore *all* possible paths from one point to another, but that the only paths realized in nature are those for which light taking closely neighboring paths experiences nearly the same elapsed time (or more generally, traverses nearly the same number of wavelengths) as the original path.

If this is true, then neighboring rays interfere constructively with each other, resulting in a much brighter beam than would occur in the absence of this constructive interference.

Thus, the round-trip paths O-A-O, O-B-O, O-C-O, and O-D-O in figure 3.12 actually occur, but not O-E-O.

Feynman explains Fermat's principle by invoking constructive and destructive interference!

Figure 3.13 illustrates a rather peculiar situation.

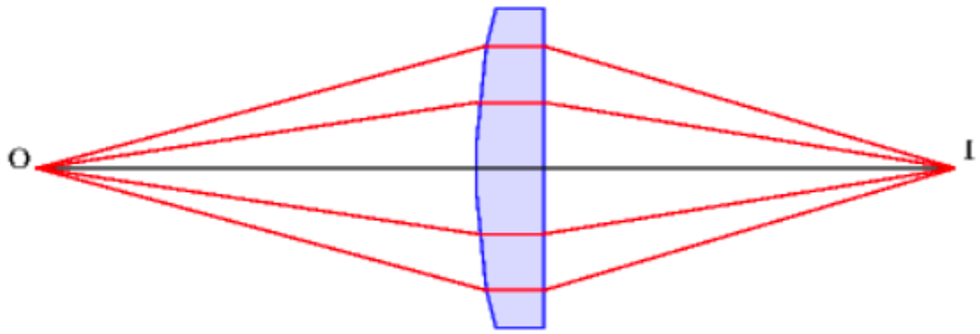


Figure 3.13: Ray trajectories from a point O being focused to another point I by a lens.

Notice that *all* the rays from point O which intercept the lens end up at point I. This would seem to contradict Fermat's principle, in that only the minimum (or maximum) time trajectories should occur.

However, a calculation shows that all the illustrated trajectories in this particular case take the *same* time.

Thus, the light cannot choose one trajectory over another using Fermat's principle and all of the trajectories are equally favored.

Note that this inference applies not to just any set of trajectories, but only those going from an object point to the corresponding image point.

Chapter 4

Special Relativity

Albert Einstein invented the special and general theories of relativity early in the 20th century, though many other people contributed to the intellectual climate which made these discoveries possible.

The special theory of relativity arose out of a conflict between the ideas of mechanics as developed by Galileo and Newton, and the theory of electromagnetism.

For this reason relativity is often discussed in textbooks after electromagnetism is developed.

However, special relativity is actually a valid extension to the Galilean world view which is needed when objects move at very high speeds, and it is only coincidentally related to electromagnetism.

For this reason we discuss relativity before electromagnetism.

The only fact from electromagnetism that we need is introduced now: There is a maximum speed at which objects can travel.

This is coincidentally equal to the speed of light in a vacuum, $c = 3 \times 10^8$ m s^{-1}.

Furthermore, a measurement of the speed of a particular light beam yields the same answer regardless of the speed of the light source or the speed at which the measuring instrument is moving.

This rather bizarre experimental result is in contrast to what occurs in Galilean relativity.

If two cars pass a pedestrian standing on a curb, one at 20 m s^{-1} and the other at 50 m s^{-1}, the faster car appears to be moving at 30 m s^{-1} relative to the slower car.

However, if a light beam moving at 3×10^8 m s^{-1} passes an interstellar spaceship moving at 2×10^8 m s^{-1}, then the light beam appears to occupants of the spaceship to be moving at 3×10^8 m s^{-1}, not 1×10^8 m s^{-1} .

Furthermore, if the spaceship beams a light signal forward to its (stationary) destination planet, then the resulting beam appears to be moving at 3×10^8 m s^{-1} to instruments at the destination, not 5×10^8 m s^{-1}.

The fact that we are talking about light beams is only for convenience.

Any other means of sending a signal at the maximum allowed speed would result in the same behavior.

We therefore cannot seek the answer to this apparent paradox in the special properties of light.

Instead we have to look to the basic nature of space and time.

4.1 Galilean Spacetime Thinking

In order to gain an understanding of both Galilean and Einsteinian relativity it is important to begin thinking of space and time as being different dimensions of a four-dimensional space called spacetime.

Actually, since we can't visualize four dimensions very well, it is easiest to start with only one space dimension and the time dimension.

Figure 4.1 shows a graph with time plotted on the vertical axis and the one space dimension plotted on the horizontal axis.

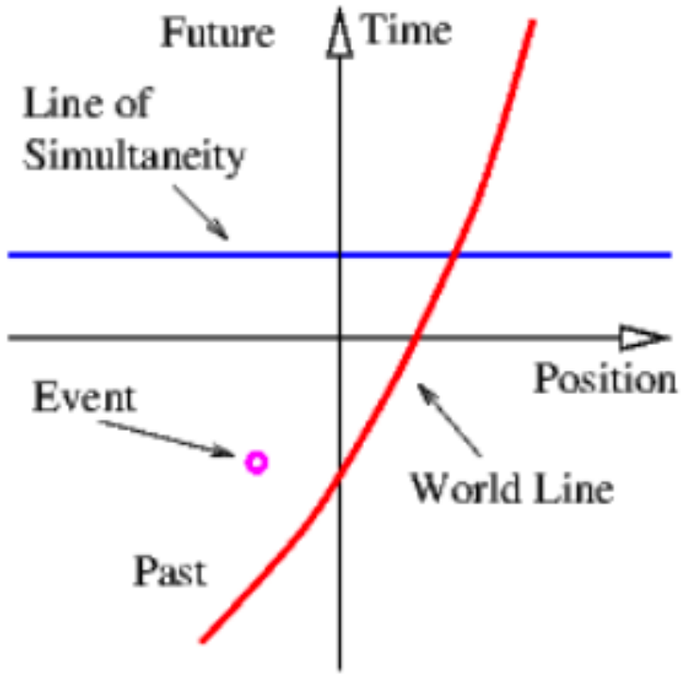


Figure 4.1: Spacetime diagram showing an *event*, a *world line*, and a *line of simultaneity*.

An *event* is something that occurs at a particular time and a particular point in space.

"Julius X. wrecks his car in Lemitar, NM on 21 June at 6:17 PM."

A *world line* is a plot of the position of some object as a function of time on a spacetime diagram, although it is conventional to put time on the vertical axis.

Thus, a world line is really a line in spacetime, while an event is a point in spacetime.

A horizontal line parallel to the position axis is a *line of simultaneity* in Galilean relativity — i. e., all events on this line occur simultaneously.

In a spacetime diagram the slope of a world line has a special meaning.

Notice that a vertical worldline means that the object it represents does not move — the velocity is zero.

If the object moves to the right, then the world line tilts to the right, and the faster it moves, the more the world line tilts.

Quantitatively, we say that

$$velocity = \frac{1}{slope\ of\ world\ line} \tag{4.1}$$

in Galilean relativity.

Notice that this works for negative slopes and velocities as well as positive ones.

If the object changes its velocity with time, then the world line is curved, and the instantaneous velocity at any time is the inverse of the slope of the tangent to the world line at that time.

The hardest thing to realize about spacetime diagrams is that they represent the past, present, and future all in one diagram.

Thus, spacetime diagrams don't change with time — the evolution of physical systems is represented by looking at successive horizontal slices in the diagram at successive times.

Spacetime diagrams represent evolution, but they don't evolve themselves.

The principle of relativity states that the laws of physics are the same in all inertial reference frames.

An inertial reference frame is one that is not accelerated.

Reference frames attached to a car at rest and to a car moving steadily down the freeway at 30 m s^{-1} are both inertial.

A reference frame attached to a car accelerating away from a stop light is not inertial.

The principle of relativity is an educated guess or hypothesis based on extensive experience.

If the principle of relativity weren't true, we would have to do all our calculations in some preferred reference frame.

This would be very annoying.

However, the more fundamental problem is that we have no idea what the velocity of this preferred frame might be.

Does it move with the earth?

That would be very earth-centric.

How about the velocity of the center of our galaxy or the mean velocity of all the galaxies?

Rather than face the issue of a preferred reference frame, physicists have chosen to stick with the principle of relativity.

If an object is moving to the left at velocity v relative to a particular reference frame, it appears to be moving at a velocity $v' = v - U$ relative to another reference frame which itself is moving at velocity U.

This is the Galilean velocity transformation law, and it is based on everyday experience.

If you are traveling 30 m s^{-1} down the freeway and another car passes you doing 40 m s^{-1}, then the other car moves past you at 10 m s^{-1} relative to your car.

Figure 4.2 shows how the world line of an object is represented differently in the unprimed (x, t) and primed (x', t') reference frames.

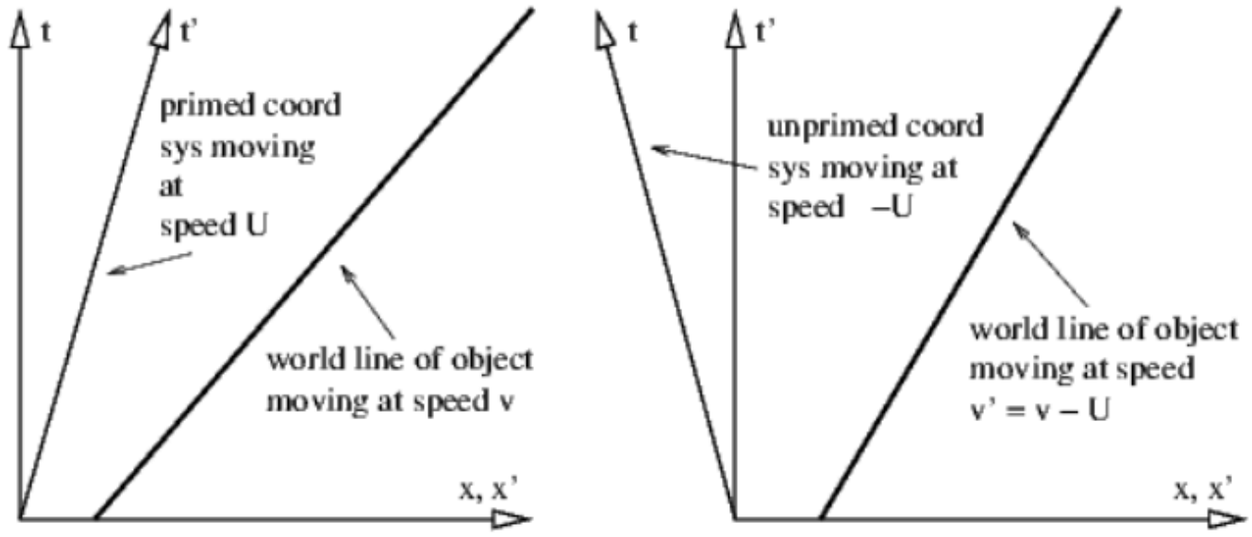


Figure 4.2: The left panel shows the world line in the unprimed reference frame, while the right panel shows it in the primed frame, which moves to the right at speed U relative to the unprimed frame. (The "prime" is just a label that allows us to distinguish the axes corresponding to the two reference frames.)

The difference between the velocity of the object and the velocity of the primed frame (i. e., the difference in the inverses of the slopes of the corresponding world lines) is the same in both reference frames in this Galilean case.

This illustrates the difference between a physical law independent of reference frame (the difference between velocities in Galilean relativity) and the different motion of the object in the two different reference frames.

4.2 Spacetime Thinking in Special Relativity

In special relativity we find that space and time "mix" in a way that they don't in Galilean relativity.

This suggests that space and time are different aspects of the same "thing", which we call *spacetime*.

If time and position are simply different dimensions of the same abstract space, then they should have the same units.

The easiest way to arrange this is to multiply time by the maximum speed, c, resulting in the kind of spacetime diagram shown in figure 4.3.

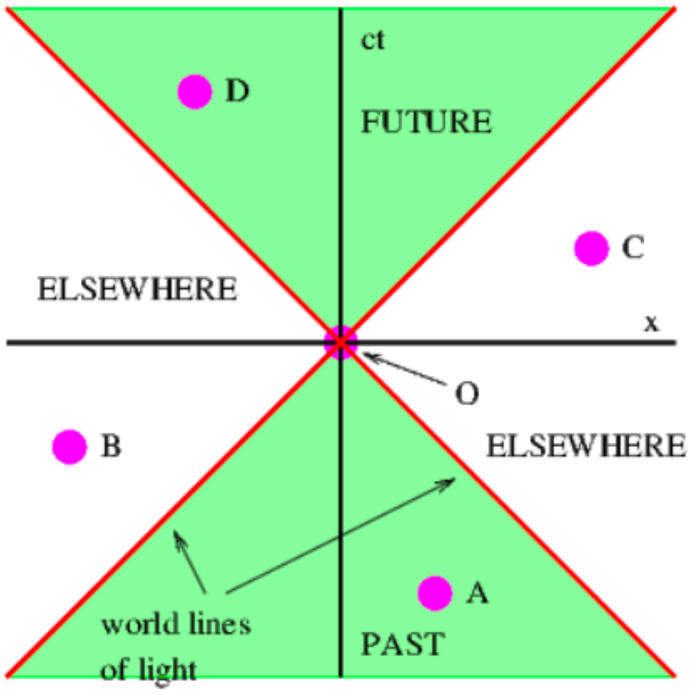


Figure 4.3: Scaled spacetime diagram showing world lines of light passing left and right through the origin.

Notice that world lines of light have slope ±1 when the time axis is scaled this way.

Furthermore, the relationship between speed and the slope of a world line must be revised to read

$$v = \frac{c}{slope} \quad \text{(world line)}. \tag{4.2}$$

Notice that it is physically possible for an object to have a world line which connects event O at the origin and the events A and D in figure 4.3, since the slope of the resulting world line would exceed unity, and thus represent a velocity less than the speed of light.

Events which can be connected by a world line are called *timelike* relative to each other.

On the other hand, event O cannot be connected to events B and C by a world line, since this would imply a velocity greater than the speed of light.

Events which cannot be connected by a world line are called *spacelike* relative to each other.

Notice the terminology in figure 4.3: Event A is in the *past* of event O, while event D is in the *future*. Events B and C are *elsewhere* relative to event O.

4.3 Postulates of Special Relativity

As we learned previously, the *principle of relativity* states that the laws of physics are the same in all inertial reference frames.

The principle of relativity applies to Einsteinian relativity just as it applies to Galilean relativity.

Notice that the constancy of the speed of light in all reference frames is consistent with the principle of relativity.

However, as noted above, it is inconsistent with our notions as to how velocities add, or alternatively, how we think the world should look from reference frames moving at different speeds.

We have called the classical way of understanding the view from different reference frames Galilean relativity.

The new way that reconciles the behavior of objects moving at very high speeds is called *Einsteinian relativity*.

Einstein's great contribution was to discover the laws that tell us how the world looks from reference frames moving at high speeds relative to each other.

These laws constitute a *geometry of spacetime*, and from them all of special relativity can be derived.

All of the observed facts about spacetime can be derived from two postulates:

- Whether two events are simultaneous depends on the reference frame from which they are viewed.
- Spacetime obeys a modified Pythagorean theorem, which gives the distance, I, in spacetime or spacetime interval as

$$I^2 = X^2 - c^2T^2, \tag{4.3}$$

where X, T, and I are defined in figure 4.4.

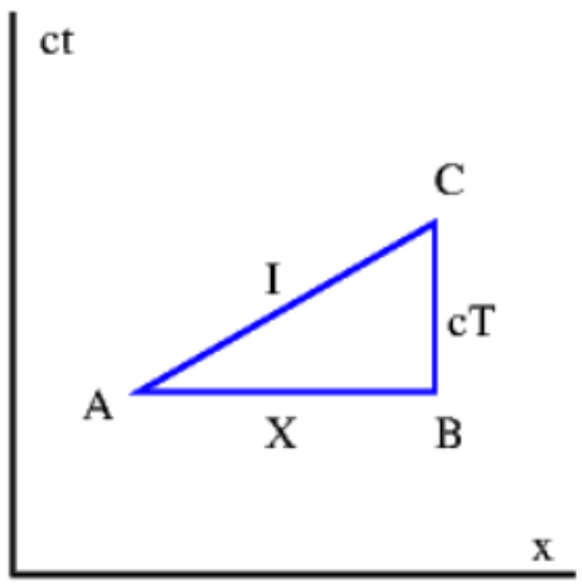


Figure 4.4: Triangle for Pythagorean theorem in spacetime.

Let us discuss these postulates in turn.

4.3.1 Simultaneity

The classical way of thinking about *simultaneity* is so ingrained in our everyday habits that we have a great deal of difficulty adjusting to what special relativity has to say about this subject.

Indeed, understanding how relativity changes this concept is the single most difficult part of the theory — once you understand this, you are well on your way to mastering relativity!

Before tackling simultaneity, let us first think about *collocation*.

Two events (such as A and E in figure 4.5) are collocated if they have the same x value.

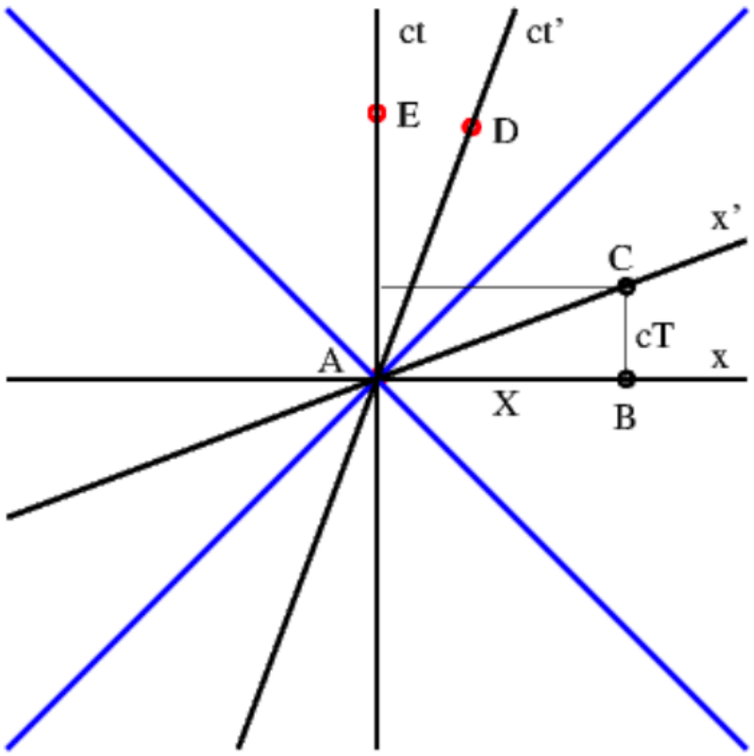


Figure 4.5: Sketch of coordinate axes for a moving reference frame, x', ct'. The meanings of the events A-E are discussed in the text. The lines tilted at ±45. are the world lines of light passing through the origin.

However, collocation is a concept that depends on the reference frame.

For instance, George is driving from Boston to Washington, with the line passing through events A and D being his world line.

Just as he passes New York he sneezes (event A in figure 4.5).

As he drives by Baltimore, he sneezes again (event D).

In the reference frame of the earth, these two sneezes are not collocated, since they are separated by many kilometers.

However, in the reference frame of George's car, they occur in the same place — assuming that George hasn't left the driver's seat!

Notice that any two events separated by a timelike interval are collocated in *some* reference frame.

The speed of the reference frame is given by equation (4.2), where the slope is simply the slope of the world line connecting the two events.

In Galilean relativity, if two events are simultaneous, we consider them to be simultaneous in all reference frames.

For instance, if two clocks, one in New York and one in Los Angeles, strike the hour at the same time in the earth reference frame, then in Galilean relativity these events also appear to be simultaneous to instruments in the space shuttle as it flies over the United States.

However, if the space shuttle is moving from west to east, i. e., from Los Angeles toward New York, careful measurements will show that the clock in New York strikes the hour *before* the clock in Los Angeles!

Thus, the Galilean point of view is not accurate.

Just as collocation depends on one's reference frame, this result shows that simultaneity also depends on the reference frame.

Figure 4.5 shows how this works.

In figure 4.5 events A and B are simultaneous in the rest or unprimed reference frame.

However, in the primed reference frame, events A and C are simultaneous, and event B occurs at an earlier time.

If A and B correspond to the clocks striking in Los Angeles and New York respectively, then it is clear that B must occur at an earlier time in the primed frame if indeed A and C are simultaneous in that frame.

The tilted line passing through events A and C in figure 4.5 is called the *line of simultaneity* for the primed reference frame.

Its slope is related to the speed, U, of the reference frame by

$$slope = U/c \quad (\text{line of simultaneity}). \tag{4.4}$$

Notice that this is the inverse of the slope of the world line attached to the primed reference frame.

There is thus a symmetry between the world line and the line of simultaneity of a moving reference frame — as the reference frame moves faster to the right, these two lines close like the blades of a pair of scissors on the 45° line.

In Galilean relativity it is fairly obvious what we mean by two events being simultaneous — it all boils down to coordinating portable clocks which are sitting next to each other, and then moving them to the desired locations.

Two events separated in space are simultaneous if they occur at the same time on clocks located near each event, assuming that the clocks have been coordinated in the above manner.

In Einsteinian relativity this doesn't work, because the very act of moving the clocks changes the rate at which the clocks run.

Thus, it is more difficult to determine whether two distant events are simultaneous.

An alternate way of experimentally determining simultaneity is shown in figure 4.6.

Since we know from observation that light travels at the same speed in all reference frames, the pulses of light emitted by the light sources in figure 4.6 will reach the two equidistant observers simultaneously in both cases.

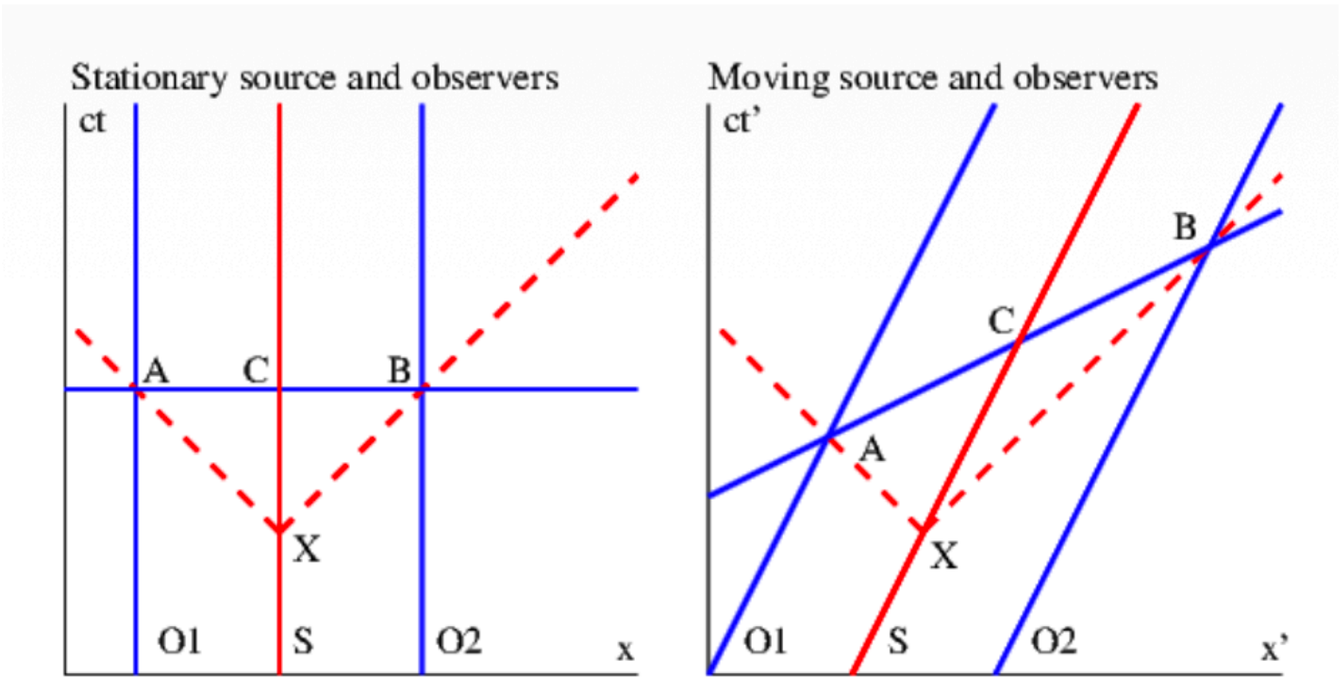


Figure 4.6: World lines of two observers (O1, O2) and a pulsed light source (S) equidistant between them. In the left frame the observers and the source are all stationary. In the right frame they are all moving to the right at half the speed of light. The dashed lines show pulses of light emitted simultaneously to the left and the right.

The line passing through these two events, A and B, defines a line of simultaneity for both stationary and moving observers.

For the stationary observers this line is horizontal, as in Galilean relativity.

For the moving observers the light has to travel farther in the rest frame to reach the observer receding from the light source, and it therefore takes longer in this frame.

Thus, event B in the right panel of figure 4.6 occurs later than event A in the stationary reference frame and the line of simultaneity is tilted.

We see that the postulate that light moves at the same speed in all reference frames leads inevitably to the dependence of simultaneity on reference frame.

4.3.2 Spacetime Pythagorean Theorem

TheThe Pythagorean theorem of spacetime differs from the usual Pythagorean theorem in two ways.

First, the vertical side of the triangle is multiplied by c.

This is a trivial scale factor that gives time the same units as space.

Second, the right side of equation (4.3) has a minus sign rather than a plus sign.

This highlights a fundamental difference between spacetime and the ordinary xyz space in which we live.

Spacetime is said to have a *non-Euclidean* (but not curved) geometry — in other words, the normal rules of geometry that we learn in high school don't always work for spacetime!

The main consequence of the minus sign in equation (4.3) is that I^2 can be negative and therefore I can be imaginary.

Furthermore, in the special case where $X = \pm cT$, we actually have $I = 0$ even though $X, T \neq 0$ — i. e., the "distance" between two well-separated events can be zero.

Clearly, spacetime has some weird properties!

The quantity I is usually called an *interval* in spacetime.

Generally speaking, if I_2 is positive, the interval is called *spacelike*, while for a negative I_2, the interval is called *timelike*.

A concept related to the spacetime interval is the *proper time* τ.

The proper time between the two events A and C in figure 4.4 is defined by the equation

$$\tau^2 = T^2 - X^2/c^2. \tag{4.5}$$

Notice that I and τ are related

$$\tau^2 = -I^2/c^2, \tag{4.6}$$

so the spacetime interval and the proper time are not independent concepts.

However, I has the dimensions of length and is real when the events defining the interval are spacelike relative to each other, whereas τ has the dimensions of time and is real when the events are timelike relative to each other.

Both equation (4.3) and equation (4.5) express the spacetime Pythagorean theorem.

If two events defining the end points of an interval have the same t value, then the interval is the ordinary space distance between the two events.

On the other hand, if they have the same x value, then the proper time is just the time interval between the events.

If the interval between two events is spacelike, but the events are not simultaneous in the initial reference frame, they can always be made simultaneous by choosing a reference frame in which the events lie on the same line of simultaneity.

Thus, the meaning of the interval in that case is just the distance between the events in the new reference frame.

Similarly, for events separated by a timelike interval, the proper time is just the time between two events in a reference frame in which the two events are collocated.

4.4 Time Dilation

Stationary and moving clocks run at different rates in relativity.

This is illustrated in figure 4.7.

The triangle ABC in the left panel of figure 4.7 can be used to illustrate this point.

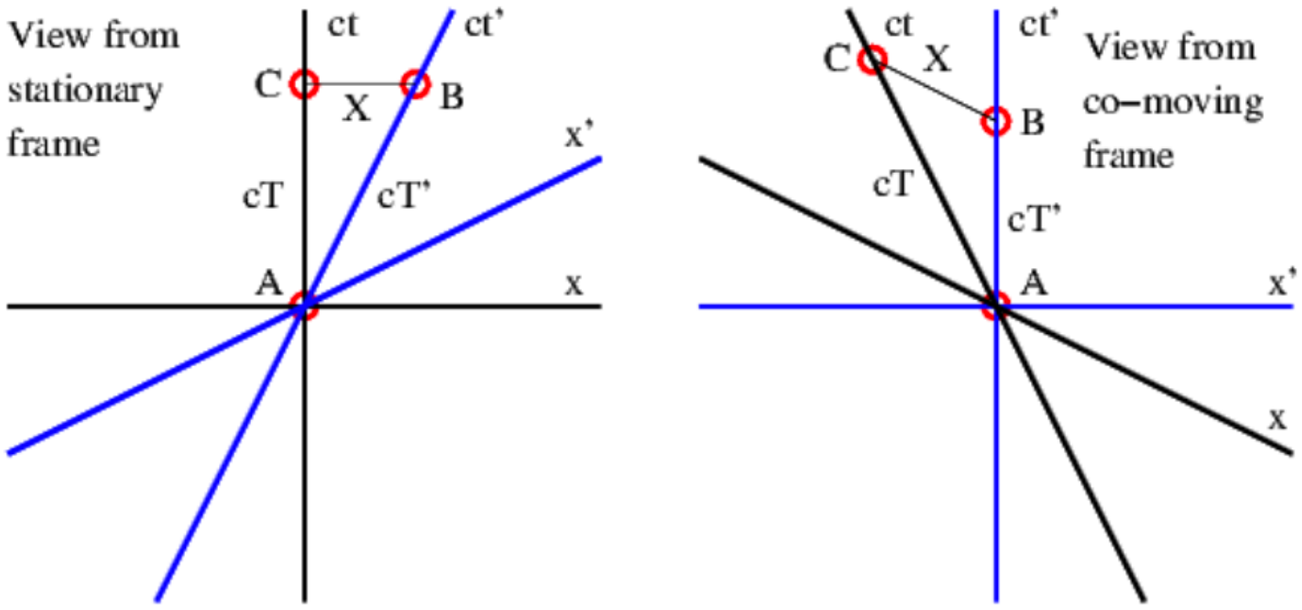


Figure 4.7: Two views of the relationship between three events, A, B, and C. The left panel shows the view from the unprimed reference frame, in which A and C are collocated, while the right panel shows the view from the primed frame, in which A and B are collocated.

Suppose that the line passing through the events A and C in this figure is the world line of a stationary observer.

At zero time another observer moving with velocity V passes the stationary observer.

The moving observer's world line passes through events A and B.

We assume that events B and C are simultaneous in the rest frame, so ABC is a right triangle.

Application of the spacetime Pythagorean theorem thus yields

$$c^2T'^2 = c^2T^2 - X^2. \tag{4.7}$$

Since the second observer is moving at velocity V , the slope of his world line is

$$\frac{c}{V} = \frac{cT}{X}, \tag{4.8}$$

where the right side of the above equation is the slope calculated as the rise of the world line cT over the run X between events A and B.

Eliminating X between the above two equations results in a relationship between T and T':

$$T' = T(1 - V^2/c^2)^{1/2} \equiv T/\gamma, \tag{4.9}$$

where

$$\gamma = \frac{1}{(1 - V^2/c^2)^{1/2}}. \tag{4.10}$$

The quantity γ occurs so often in relativistic calculations that we give it this special symbol.

Note by its definition that $\gamma \geq 1$.

Equation (4.9) tells us that the time elapsed for the moving observer is less than that for the stationary observer, which means that the clock of the moving observer runs more slowly.

This is called the *time* dilation effect.

Let us view this situation from the reference frame of the moving observer.

In this frame the moving observer becomes stationary and the stationary observer moves in the opposite direction, as illustrated in the right panel of figure 4.7.

By symmetric arguments, one infers that the clock of the initially stationary observer who is now moving to the left runs more slowly in this reference frame than the clock of the initially moving observer.

One might conclude that this contradicts the previous results.

However, examination of the right panel of figure 4.7 shows that this is not so.

The interval cT is still greater than the interval cT', because such intervals are relativistically invariant quantities.

However, events B and C are no longer simultaneous, so one cannot use these results to infer anything about the rate at which the two clocks run in this frame.

Thus, the relative nature of the concept of simultaneity saves us from an incipient paradox, and we see that the relative rates at which clocks run depends on the reference frame in which these rates are observed.

4.5 Lorentz Contraction

A similar argument can be made to show how the postulates of relativity result in the *Lorentz contraction*.

Figure 4.8 compares the length X' of a moving object measured in its own reference frame (left panel) with its length X as measured in a stationary reference frame (right panel).

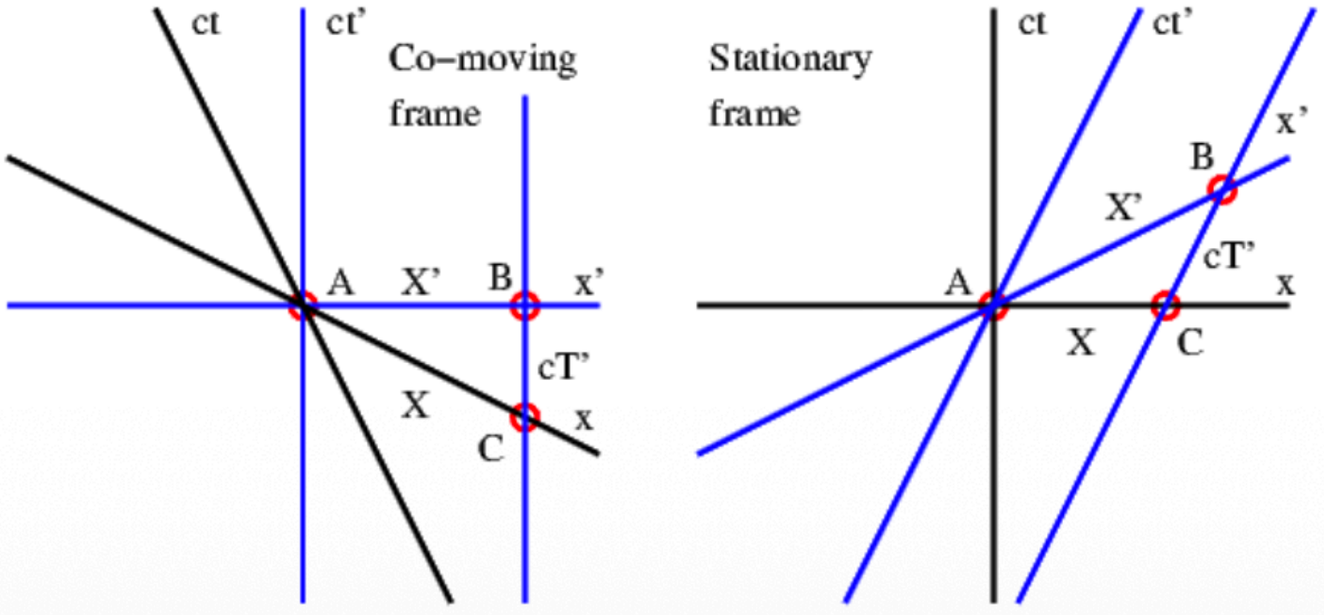


Figure 4.8: Definition sketch for understanding the Lorentz contraction. The parallel lines represent the world lines of the front and the rear of a moving object. The left panel shows a reference frame moving with the object, while the right panel shows a stationary reference frame.

The length of a moving object is measured by *simultaneously* measuring the positions of the front and the rear of the object and subtracting these two numbers.

Events A and C correspond to these position measurements for the stationary reference frame since they are respectively on the rear and front world lines of the object.

Thus, the interval AC, which is equal to X, is the length of the object as measured in the stationary frame.

In the left panel, X is the hypotenuse of a right triangle.

Therefore, by the Pythagorean theorem of spacetime, we have

$$X = I = (X'^2 - c^2T'^2)^{1/2}. \tag{4.11}$$

Now, the line passing through A and C in the left panel is the line of simultaneity of the stationary reference frame.

The slope of this line is $-V/c$, where V is the speed of the object relative to the stationary reference frame.

Geometrically in figure 4.8, the slope of this line is $-cT'/X'$, so we find by equating these two expressions for the slope that

$$T' = VX'/c^2. \tag{4.12}$$

Finally, eliminating T' between (4.11) and (4.12) results in

$$X = X'(1 - V^2/c^2)^{1/2} = X'/\gamma. \tag{4.13}$$

This says that the length of a moving object as measured in a stationary reference frame (X) is *less* than the actual length of the object as measured in its own reference frame (X').

This reduction in length is called the Lorentz contraction.

Note that the Lorentz contraction only occurs in the direction of motion.

The dimensions of a moving object perpendicular to the motion remain unchanged.

4.6 Twin Paradox

An interesting application of time dilation is the so-called *twin paradox*, which turns out not to be a paradox at all.

Two twins are initially the same age.

One twin travels to a distant star on an interstellar spaceship which moves at speed V , which is close to the speed of light.

Upon reaching the star, the traveling twin immediately turns around and heads home.

When reaching home, the traveling twin has aged less than the twin that stayed home.

This is easily explained by the time dilation effect, which shows that the proper time elapsed along the world line of the traveling twin is $(1 - V^2/c^2)^{1/2}$ times the proper time elapsed along the world line of the other twin.

The “paradox” part of the twin paradox arises from making the symmetric argument in which one assumes the reference frame of the traveling twin to be stationary.

The frame of the earth-bound twin must then travel in the sense opposite that of the erstwhile traveling twin, which means that the earthbound twin must age less rather than more.

This substitution is justified on the basis of the principle of relativity, which states that the laws of physics must be the same in all inertial reference frames.

However, the above argument is fallacious, because the reference frame of the traveling twin is *not* inertial throughout the entire trip, since at various points the spaceship has to accelerate and decelerate.

Thus, the principle of relativity cannot be used to assert the equivalence of the traveler's reference frame to the stationary frame.

Of particular importance is the period of deceleration and acceleration near the distant star.

During this interval, the line of simultaneity of the traveling twin rotates, as illustrated in figure 4.9, such that the twin staying at home rapidly ages in the reference frame of the traveler.

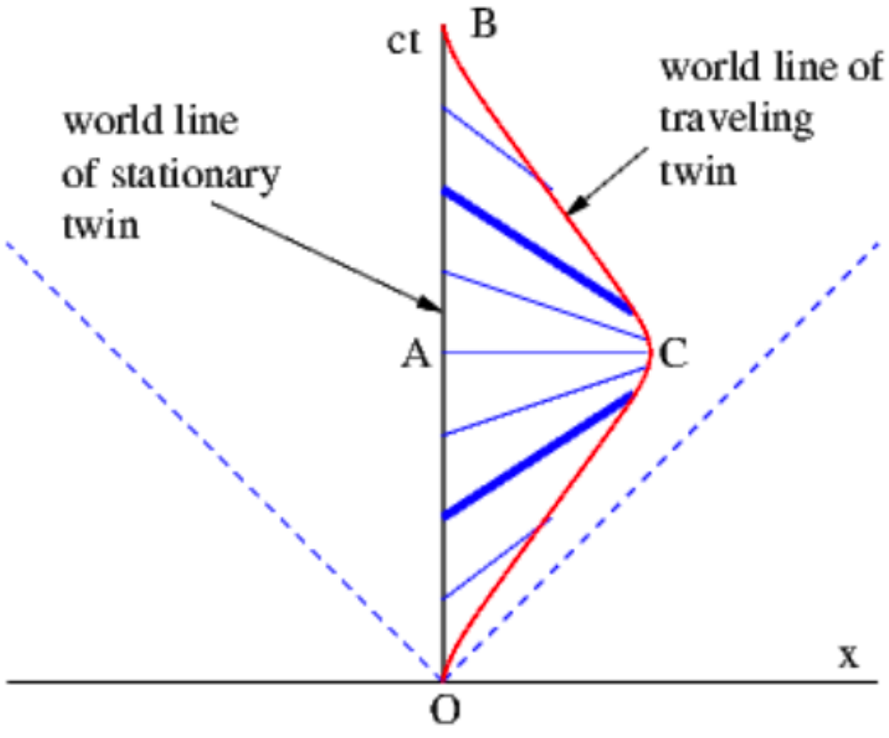


Figure 4.9: Definition sketch for the twin paradox. The vertical line is the world line of the twin that stays at home, while the traveling twin has the curved world line to the right. The slanted lines between the world lines are lines of simultaneity at various times for the traveling twin. The heavy lines of simultaneity bound the period during which the traveling twin is decelerating to a stop and accelerating toward home.

Thus, even though the acceleration of the traveling twin may occupy only a negligible segment of the twin's world line, the overall effect is not negligible.

In fact, the shorter and more intense the period of acceleration, the more rapidly the earth-bound twin ages in the traveling frame during this interval!

Chapter 5

Applications of Special Relativity

In this chapter we continue the study of special relativity.

Three important applications of the ideas developed in the previous chapter are made here.

First, we show how to describe waves in the context of spacetime.

We then see how waves which have no preferred reference frame (such as that of a medium supporting them) are constrained by special relativity to have a dispersion relation of a particular form.

This dispersion relation turns out to be that of the relativistic matter waves of quantum mechanics.

Second, we investigate the Doppler shift phenomenon, in which the frequency of a wave takes on different values in different coordinate systems.

Third, we show how to add velocities in a relativistically consistent manner.

A new mathematical idea is presented in the context of relativistic waves, namely the spacetime vector or four-vector.

Writing the laws of physics totally in terms of relativistic scalars and four-vectors insures that they will be valid in all inertial reference frames.

5.1 Waves in Spacetime

TheWe now look at the characteristics of waves in spacetime.

Recall that a sine wave moving to the right in one space dimension can be represented by

$$A(x,t) = A_0 \sin(kx - \omega t), \tag{5.1}$$

where A_0 is the (constant) amplitude of the wave, k is the wavenumber, and ω is the angular frequency, and that the quantity $\varphi = kx - \omega t$ is called the *phase* of the wave.

For a plane wave in three space dimensions, the wave is represented in a similar way,

$$A(\mathbf{x},t) = A_0 \sin(\mathbf{k}\cdot\mathbf{x} - \omega t), \tag{5.2}$$

where $\mathbf{x}$ is now the position vector and $\mathbf{k}$ is the wave vector.

The magnitude of the wave vector, $|\mathbf{k}| = k$ is just the wavenumber of the wave and the direction of this vector indicates the direction the wave is moving.

The phase of the wave in this case is $\varphi = \mathbf{k}\cdot\mathbf{x} - \omega t$.

In the case of a one-dimensional wave moving to the right $\varphi = kx - \omega t$.

A wave front has constant phase φ, so solving this equation for t and multiplying by c, the speed of light in a vacuum, gives us an equation for the world line of a wave front:

$$ct = \frac{ckx}{\omega} - \frac{c\phi}{\omega} = \frac{cx}{u_p} - \frac{c\phi}{\omega} \quad \text{(wave front)}. \tag{5.3}$$

The slope of the world line in a spacetime diagram is the coefficient of x, or c/u_p, where $u_p = \omega/k$ *is the* phase speed.

The world lines of the wave fronts of a wave are illustrated in figure 5.1.

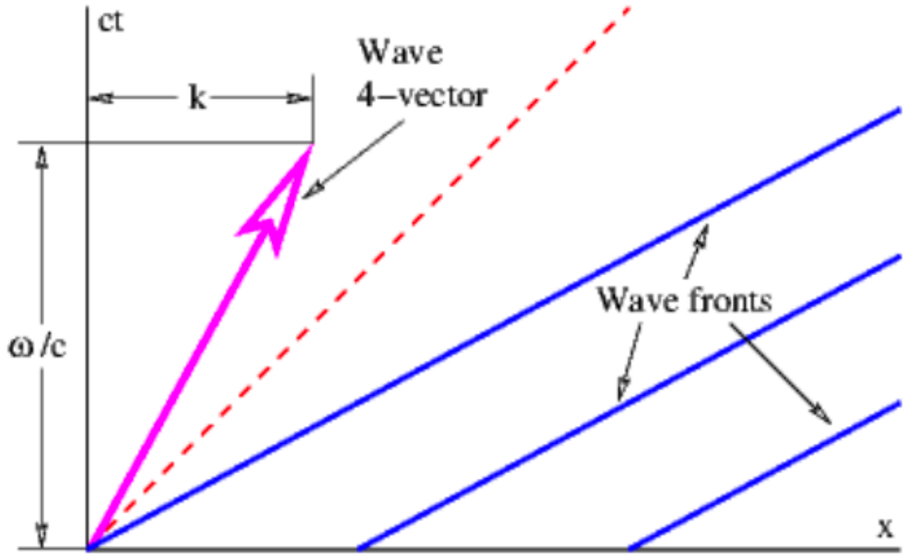


Figure 5.1: Sketch of wave fronts for a wave in spacetime. The large arrow is the associated wave four-vector, which has slope ω/ck. The slope of the wave fronts is the inverse, ck/ω. The phase speed of the wave is greater than c in this example. (Can you tell why?)

5.2 Math Tutorial – Four-Vectors

Also shown in figure 5.1 is a spacetime vector or *four-vector* which represents the frequency and wavenumber of the wave, which we refer to as the *wave four-vector*.

It is called a four-vector because it has 3 spacelike components and one timelike component when there are 3 space dimensions.

In the case shown, there is only a single space dimension.

The spacelike component of the wave four-vector is just k (or $\mathbf{k}$ when there are 3 space dimensions), while the timelike component is ω/c.

The c is in the denominator to give the timelike component the same dimensions as the spacelike component.

From figure 5.1 it is clear that the slope of the line representing the four-vector is ω/ck, which is just the inverse of the slope of the wave fronts.

Let us define some terminology.

We indicate a four-vector by underlining and write the components in the following way: $\underline{k} = (k,\omega/c)$, where $\underline{k}$ is the wave four-vector, k is its spacelike component, and ω/c is its timelike component.

For three space dimensions, where we have a wave vector rather than just a wavenumber, we write $\underline{k} = (\mathbf{k},\omega/c)$.

Another example of a four-vector is simply the position vector in spacetime, $\underline{x} = (x,ct)$, or $\underline{x} = (\mathbf{x},ct)$ in three space dimensions.

The *c multiplies* the timelike component in this case, because that is what is needed to give it the same dimensions as the spacelike component.

In three dimensions we define a vector as a quantity with magnitude and direction.

Extending this to spacetime, a four-vector is a quantity with magnitude and direction *in spacetime*.

Implicit in this definition is the notion that the vector's magnitude is a quantity independent of coordinate system or reference frame.

We have seen that the invariant interval in spacetime from the origin to the point (x,ct) is $I = (x^2 - c^2t^2)^{1/2}$, so it makes sense to identify this as the magnitude of the position vector.

This leads to a way of defining a dot product of four-vectors.

Given two four-vectors $\underline{A} = (\mathbf{A},A_t)$ and $\underline{B} = (\mathbf{B},B_t)$, the dot product is

$$\underline{A} \cdot \underline{B} = \mathbf{A} \cdot \mathbf{B} - A_t B_t \quad \text{(dot product in spacetime)}. \tag{5.4}$$

This is consistent with the definition of invariant interval if we set $\underline{A} = \underline{B} = \underline{x}$, since then $\underline{x} \cdot \underline{x} = x^2 - c^2t^2 = I^2$.

In the odd geometry of spacetime it is not obvious what "perpendicular" means.

We therefore *define* two four-vectors $\underline{A}$ and $\underline{B}$ to be perpendicular if their dot product is zero, $\underline{A} \cdot \underline{B} = 0$, in analogy with ordinary vectors.

The dot product of two four-vectors is a *scalar* result, i. e., its value is independent of coordinate system.

This can be used to our advantage on occasion.

For instance, consider the dot product of a four-vector $\underline{A}$ which resolves into (A_x,A_t) in the unprimed frame.

Let us further suppose that the spacelike component is zero in some primed frame, so that the components in this frame are $(0,A_t')$.

The fact that the dot product is independent of coordinate system means that

$$\underline{A}\cdot\underline{A} = A_x^2 - A_t^2 = -A_t'^2. \tag{5.5}$$

This constitutes an extension of the spacetime Pythagorean theorem to four-vectors other than the position four-vector.

Thus, for instance, the wavenumber for some wave may be zero in the primed frame, which means that the wavenumber and frequency in the unprimed frame are related to the frequency in the primed frame by

$$k^2 - \omega^2/c^2 = -\omega'^2/c^2. \tag{5.6}$$

5.3 Principle of Relativity Applied

Returning to the phase of a wave, we immediately see that

$$\phi = \mathbf{k}\cdot\mathbf{x} - \omega t = \mathbf{k}\cdot\mathbf{x} - (\omega/c)(ct) = \underline{k}\cdot\underline{x}. \tag{5.7}$$

Thus, a compact way to rewrite equation (5.2) is

$$A(\underline{x}) = A_0 \sin(\underline{k}\cdot\underline{x}). \tag{5.8}$$

\

Since $\underline{x}$ is known to be a four-vector and since the phase of a wave is known to be a scalar independent of reference frame, it follows that $\underline{k}$ is also a four-vector rather than just a set of numbers.

Thus, the square of the length of the wave four-vector must also be a scalar independent of reference frame:

$$\underline{k} \cdot \underline{k} = \mathbf{k} \cdot \mathbf{k} - \omega^2/c^2 = const. \tag{5.9}$$

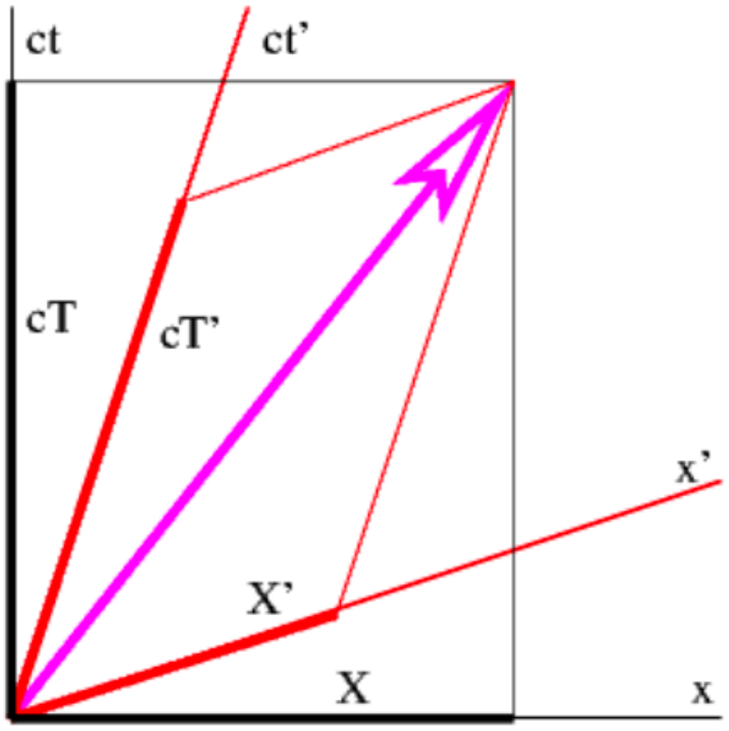


Figure 5.2: Resolution of a four-vector into components in two different reference frames.

The

Let us review precisely what this means.

As figure 5.2 shows, we can resolve a position four-vector into components in two different reference frames, $\underline{x} = (X,cT) = (X',cT')$.

However, even though $X \neq X'$ and $T \neq T'$, the vector lengths computed from these two sets of components are necessarily the same: $\underline{x} \cdot \underline{x} = X^2 - c^2T^2 = X'^2 - c^2T'^2$.

Applying this to the wave four-vector, we infer that

$$k^2 - \omega^2/c^2 = k'^2 - \omega'^2/c^2 = const., \tag{5.10}$$

where the unprimed and primed values of k and ω refer to the components of the wave four-vector in two different reference frames.

Up to now, this argument applies to *any* wave.

However, waves can be divided into two categories, those for which a "special" reference frame exists, and those for which there is no such special frame.

As an example of the former, sound waves look simplest in the reference frame in which the gas carrying the sound is stationary.

The same is true of light propagating through a material medium with an index of refraction not equal to unity.

In both cases the speed of the wave is the same in all directions *only* in the frame in which the material medium is stationary.

Suppose we have a machine that produces a wave with wavenumber k and frequency ω in its own rest frame.

If we observe the wave from a moving reference frame, the wavenumber and frequency will be different, say, k' and ω'.

However, these quantities will be related by equation (5.10).

Up to this point the argument applies to *any* wave whether a special reference frame exists or not; the observed changes in wavenumber and frequency have nothing to do with the wave itself, but are just consequences of how we have chosen to observe it.

However, if there is no special reference frame for the type of wave under consideration, then the same result can be obtained by keeping the observer stationary and moving the wave-producing machine in the opposite direction.

By moving it at various speeds, any desired value of k' can be obtained *in the initial reference frame* (as opposed to some other frame), and the resulting value of ω' can be computed using equation (5.10).

This is actually an amazing result.

We have shown on the basis of the principle of relativity that any wave type for which no special reference frame exists can be made to take on a full range of frequencies and wavenumbers in any given reference frame, and furthermore that these frequencies and wavenumbers obey

$$\omega^2 = k^2c^2 + \mu^2. \tag{5.11}$$

Equation (5.11) comes from solving equation (5.10) for ω_2 and the constant μ^2 equals the constant in equation (5.10) times $-c^2$.

Equation (5.11) relates frequency to wavenumber and therefore is the *dispersion relation* for such waves.

We call waves which have no special reference frame and therefore necessarily obey equation (5.11) *relativistic waves*.

The only difference in the dispersion relations between different types of relativistic waves is the value of the constant μ.

The meaning of this constant will become clear later.

5.4 Characteristics of Relativistic Waves

Light in a vacuum is an example of a wave for which no special reference frame exists.

For light, $\mu = 0$, and we have (taking the positive root) $\omega = ck$.

This simply states what we know already, namely that the phase speed of light in a vacuum is c.

If $\mu \neq 0$, waves of this type are dispersive.

The phase speed is

$$u_p = \frac{\omega}{k} = (c^2 + \mu^2/k^2)^{1/2}. \tag{5.12}$$

This phase speed always exceeds c, which at first seems like an unphysical conclusion.

However, the group velocity of the wave is

$$u_g = \frac{d\omega}{dk} = \frac{c^2 k}{(k^2 c^2 + \mu^2)^{1/2}} = \frac{kc^2}{\omega} = \frac{c^2}{u_p}, \tag{5.13}$$

which is always less than c.

Since wave packets and hence signals propagate at the group velocity, waves of this type are physically reasonable even though the phase speed exceeds the speed of light.

Another interesting property of such waves is that the wave four-vector is parallel to the world line of a wave packet in spacetime.

This is easily shown by the following argument.

As figure 5.1 shows, the spacelike component of a wave four-vector is k, while the timelike component is ω/c.

The slope of the four-vector on a spacetime diagram is therefore ω/kc.

However, the slope of the world line of a wave packet moving with group velocity u_g is $c/u_g = \omega/(kc)$, which is the same as the slope of the k four-vector.

Note that when $k = 0$ we have $\omega = \mu$.

In this case the group velocity of the wave is zero.

For this reason we call μ the *rest frequency* of the wave.

5.5 The Doppler Shift

You have probably heard how the pitch of a train horn changes as it passes you.

When the train is approaching, the pitch or frequency is higher than when it is moving away from you.

This is called the Doppler shift.

A similar, but distinct shift occurs if you are moving past a source of sound.

If a stationary whistle is blowing, the pitch perceived from a moving car is higher while moving toward the source than when moving away.

The first case thus has a *moving source*, while the second case has a *moving* observer.

In this section we compute the Doppler shift as it applies to light moving through a vacuum.

Figure 5.3 shows the geometry for computing the time between wave fronts of light for a stationary and a moving reference frame.

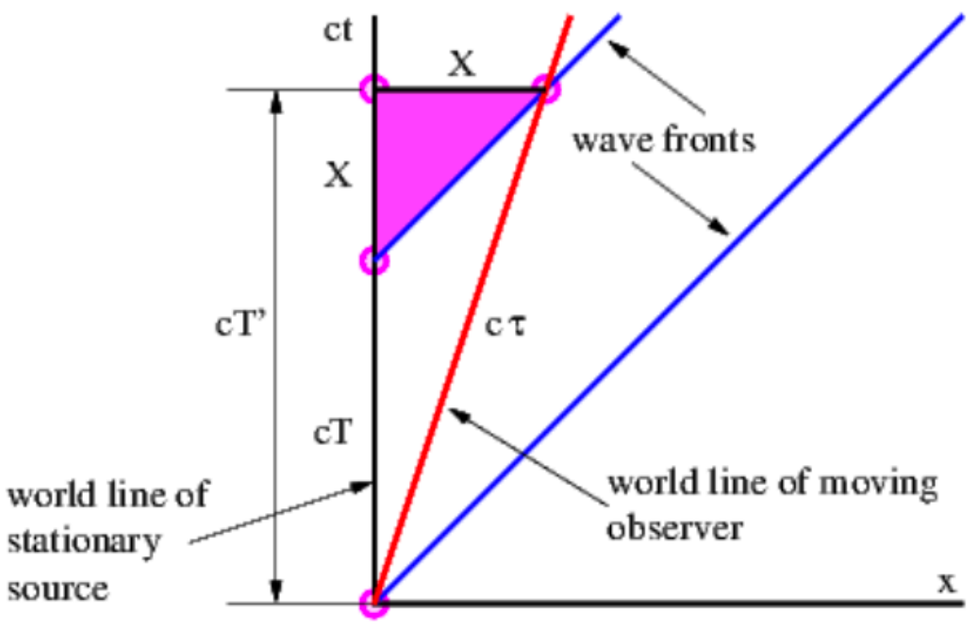


Figure 5.3: Definition sketch for computing the Doppler shift for light.

The time in the stationary frame is just T.

Since the world lines of the wave fronts have a slope of unity, the sides of the shaded triangle have the same value, X.

If the observer is moving at speed U, the slope of the observer's world line is c/U, which means that $c/U = (cT + X)/X$.

Solving this for X yields $X = UT/(1 - U/c)$, which can then be used to compute $T' = T + X/c = T/(1 - U/c)$.

This formula as it stands leads to the classical Doppler shift for a moving observer.

However, with relativistic velocities, one additional factor needs to be taken into account: The observer experiences time dilation since he or she is moving.

The actual time measured by the observer between wave fronts is actually

$$\tau = (T'^2 - X^2/c^2)^{1/2} = T\frac{(1 - U^2/c^2)^{1/2}}{1 - U/c} = T\left(\frac{1 + U/c}{1 - U/c}\right)^{1/2}, \qquad (5.14)$$

where the last step uses $1 - U^2/c^2 = (1 - U/c)(1 + U/c)$.

From this we infer the relativistic Doppler shift formula for light in a vacuum:

$$\omega' = \omega\left(\frac{1 - U/c}{1 + U/c}\right)^{1/2}, \qquad (5.15)$$

where the frequency measured by the moving observer is $\omega' = 2\pi/\tau$ and the frequency observed in the stationary frame is $\omega = 2\pi/T$.

We could go on to determine the Doppler shift resulting from a moving source.

However, by the principle of relativity, the laws of physics should be the same in the reference frame in which the observer is stationary and the source is moving.

Furthermore, the speed of light is still c in that frame.

Therefore, the problem of a stationary observer and a moving source is conceptually the same as the problem of a moving observer and a stationary source when the wave is moving at speed c.

This is unlike the case for, say, sound waves, where the stationary observer and the stationary source yield different formulas for the Doppler shift.

5.6 Addition of Velocities

Figure 5.4 shows the world line of a moving object from the point of view of two different reference frames, with the primed frame (left panel) moving to the right at speed U relative to the unprimed frame (right panel).

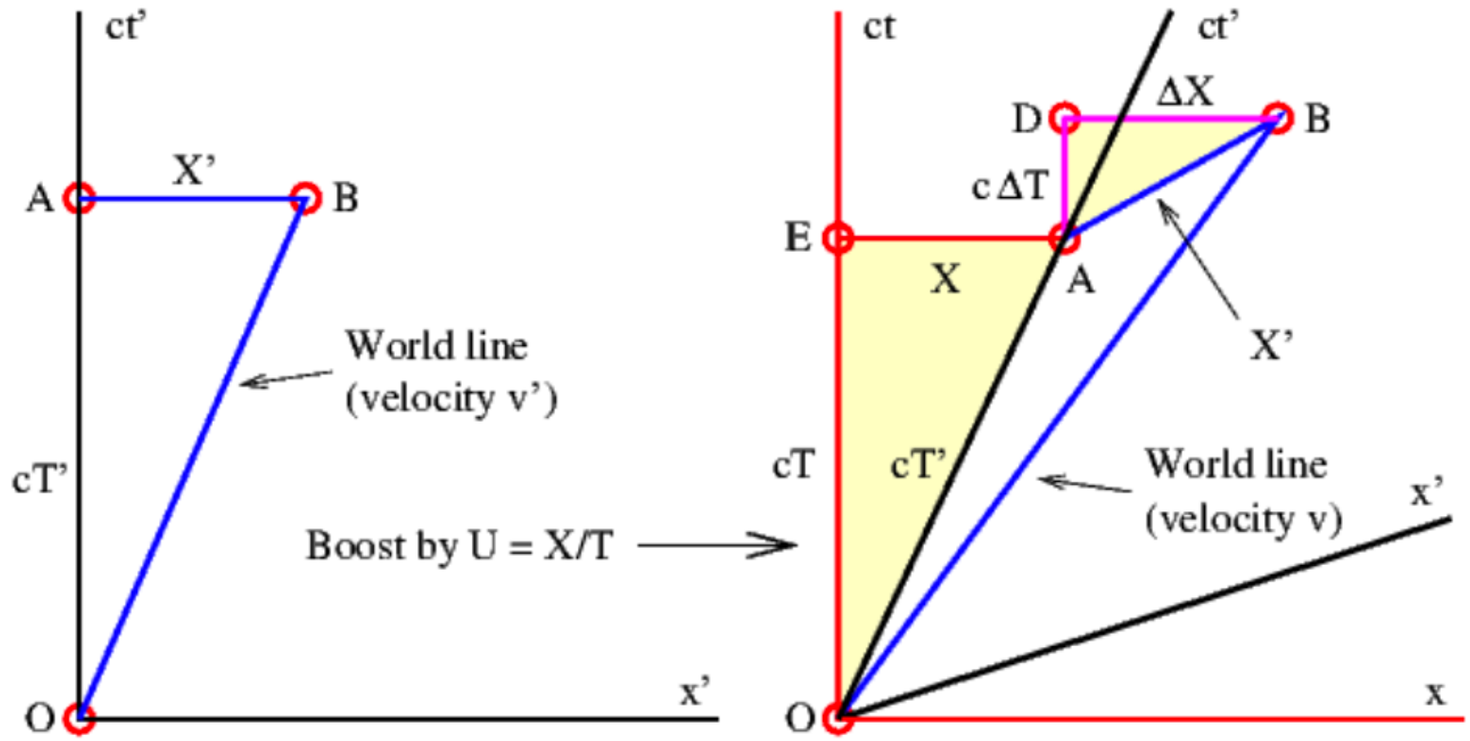


Figure 5.4: Definition sketch for relativistic velocity addition. The two panels show the world line of a moving object relative to two different reference frames moving at velocity U with respect to each other. The velocity of the world line in the left panel is v' while its velocity in the right panel is v.

The goal is to calculate the velocity of the object relative to the unprimed frame, v, assuming its velocity in the primed frame, v' is known.

The classical result is simply

$$v = U + v' \quad \text{(classical result)}. \tag{5.16}$$

However, this is inconsistent with the speed of light being constant in all reference frames, since if we substitute c for v', this formula predicts that the speed of light in the unprimed frame is $U + c$.

We can use the geometry of figure 5.4 to come up with the correct relativistic formula.

From the right panel of this figure we infer that

$$\frac{v}{c} = \frac{X + \Delta X}{cT + c\Delta T} = \frac{X/(cT) + \Delta X/(cT)}{1 + \Delta T/T}. \tag{5.17}$$

This follows from the fact that the slope of the world line of the object in this frame is c/v.

The slope is calculated as the ratio of the rise, $c(T + \Delta T)$, to the run, $X + \Delta X$.

From the left panel of figure 5.4 we similarly see that

$$\frac{v'}{c} = \frac{X'}{cT'}. \tag{5.18}$$

However, we can apply our calculations of Lorentz contraction and time dilation from the previous chapter to triangles ABD and OAE in the right panel.

The slope of AB is U⁄c because AB is horizontal in the left panel, so $X' = \Delta X(1 - U^2/c^2)^{1/2}$.

Similarly, the slope of OA is c⁄U since OA is vertical in the left panel, and $T' = T(1 - U^2/c^2)^{1/2}$.

Substituting these formulas into the equation for v′⁄c yields

$$\frac{v'}{c} = \frac{\Delta X}{cT}. \tag{5.19}$$

Again using what we know about the triangles ABD and OAE, we see that

$$\frac{U}{c} = \frac{c\Delta T}{\Delta X} = \frac{X}{cT}. \tag{5.20}$$

Finally, we calculate $\Delta T/T$ by noticing that

$$\frac{\Delta T}{T} = \frac{\Delta T}{T}\frac{c\Delta X}{c\Delta X} = \left(\frac{c\Delta T}{\Delta X}\right)\left(\frac{\Delta X}{cT}\right) = \frac{U}{c}\frac{v'}{c}. \tag{5.21}$$

Substituting equations (5.19), (5.20), and (5.21) into equation (5.17) and simplifying yields the relativistic velocity addition formula:

$$v = \frac{U + v'}{1 + Uv'/c^2} \quad \text{(special relativity)}. \tag{5.22}$$

Notice how this equation behaves in various limits. If $|Uv'| \ll c^2$, the denominator of equation (5.22) is nearly unity, and the special relativistic formula reduces to the classical case.

On the other hand, if $v' = c$, then equation (5.22) reduces to $v = c$.

In other words, if the object in question is moving at the speed of light in one reference frame, it is moving at the speed of light in all reference frames, i. e., for all possible values of U.

Thus, we have found a velocity addition formula that 1) reduces to the classical formula for low velocities and 2) gives the observed results for very high velocities as well.

Equation (5.22) is valid even if v' is negative, i. e., if the object is moving to the right less rapidly than the primed reference frame, or even if it is moving to the left in the unprimed frame.

Chapter 6

Acceleration and General Relativity

General relativity is Einstein's extension of special relativity to include gravity.

An important aspect of general relativity is that spacetime is no longer necessarily flat, but in fact may be curved under the influence of mass.

Understanding curved spacetime is an advanced topic which is not easily accessible at the level of this text.

However, it turns out that some insight into general relativistic phenomena may be obtained by investigating the effects of acceleration in the flat (but non-Euclidean) space of special relativity.

The central assumption of general relativity is the equivalence principle, which states that gravity is a force which arises from being in an accelerated reference frame.

To understand this we must first investigate the concept of acceleration.

We then see how this leads to phenomena such as the gravitational red shift, event horizons, and black holes.

We also introduce in a preliminary way the notions of force and mass.

6.1 Acceleration

Imagine that you are in a powerful luxury car stopped at a stoplight.

As you sit there, gravity pushes you into the comfortable leather seat.

The light turns green and you “floor it”.

The car accelerates and an additional force pushes you into the seat back.

You round a curve, and yet another force pushes you toward the outside of the curve.

But the well designed seat and seat belt keep you from feeling discomfort!

Let us examine the idea of *acceleration* more closely.

Considering first acceleration in one dimension, figure 6.1 shows the position of an object as a function of time, $x(t)$.

The velocity is simply the time rate of change of the position:

$$v(t) = \frac{dx(t)}{dt}. \tag{6.1}$$

The acceleration is the time rate of change of velocity:

$$a(t) = \frac{dv(t)}{dt} = \frac{d^2x(t)}{dt^2}. \tag{6.2}$$

In figure 6.1, only the segment OA has zero velocity.

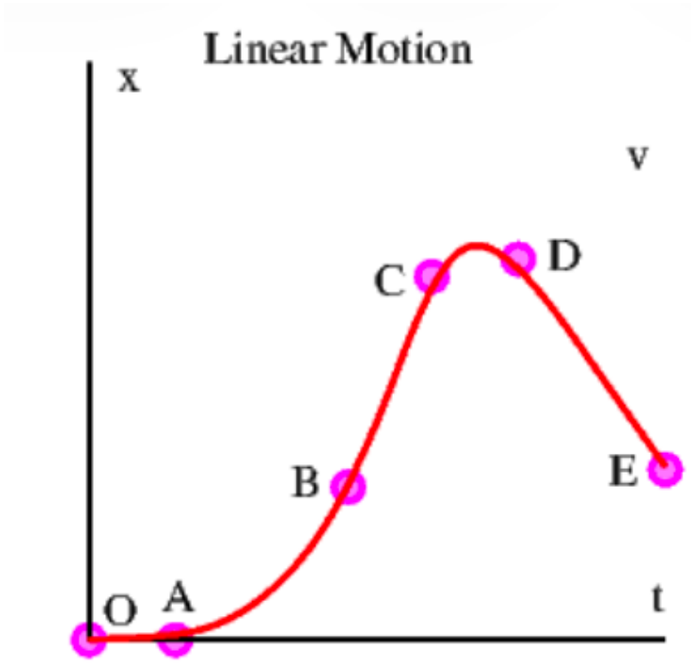


Figure 6.1: Example of linear motion.

Velocity is increasing in AB, and the acceleration is positive there.

Velocity is constant in BC, which means that the acceleration is zero.

Velocity is decreasing in CD, and the acceleration is negative.

Finally, in DE, the velocity is negative and the acceleration is zero.

In two or three dimensions, position **x**, velocity **v**, and acceleration **a** are all vectors, so that the velocity is

$$\mathbf{v}(t) = \frac{d\mathbf{x}(t)}{dt} \tag{6.3}$$

while the acceleration is

$$\mathbf{a}(t) = \frac{d\mathbf{v}(t)}{dt}. \tag{6.4}$$

Thus, over some short time interval Δt, the changes in **x** and **v** can be written

$$\Delta\mathbf{x} = \mathbf{v}\Delta t \qquad \Delta\mathbf{v} = \mathbf{a}\Delta t. \tag{6.5}$$

These are vector equations, so the subtractions implied by the "delta" operations must be done vectorially.

An example where the vector nature of these quantities is important is motion in a circle at constant speed, which is discussed in the next section.

6.2 Circular Motion

Imagine an object constrained by an attached string to move in a circle at constant speed, as shown in the left panel of figure 6.2.

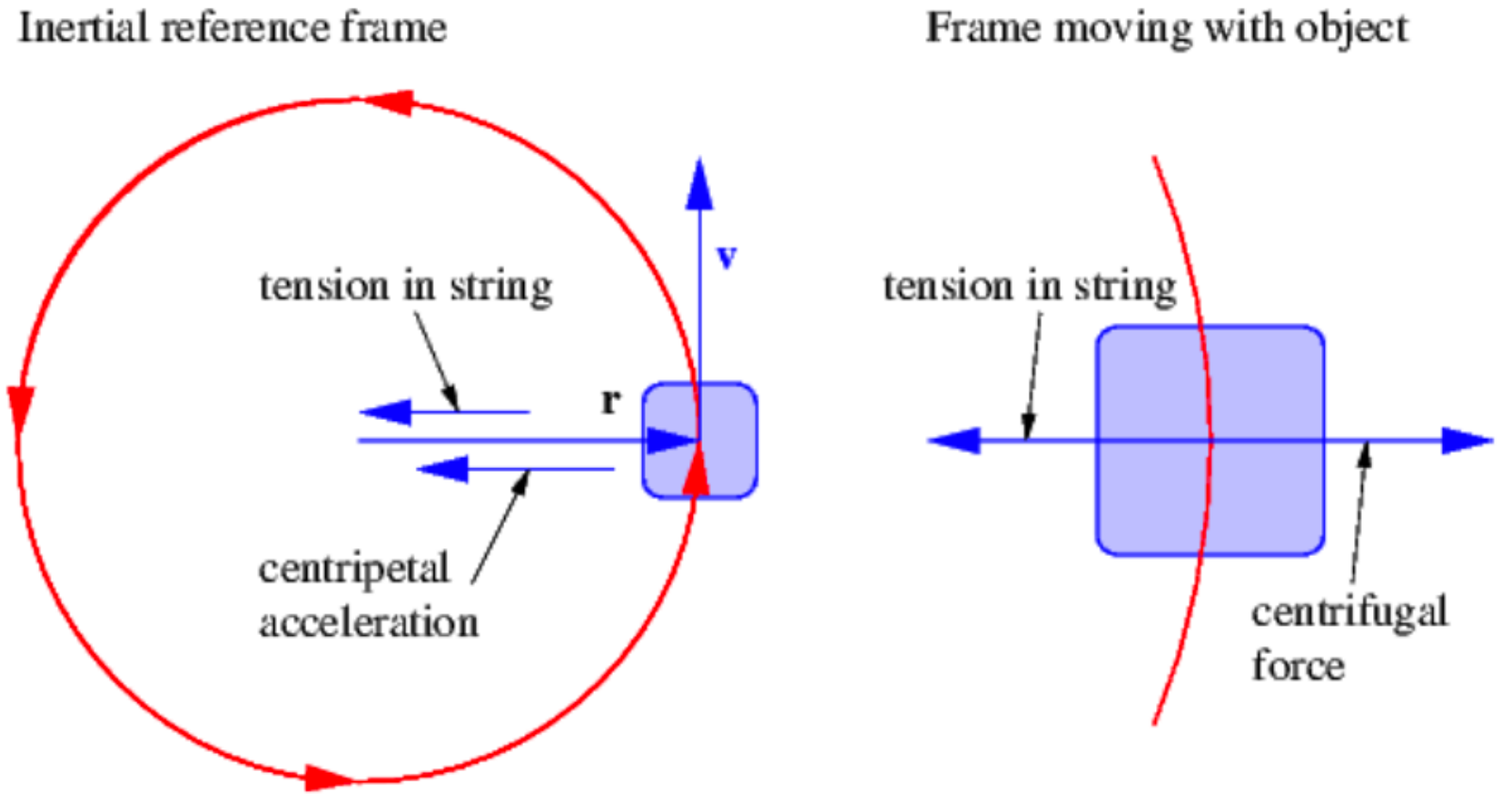


Figure 6.2: Two different views of circular motion of an object. The left panel shows the view from the inertial reference frame at rest with the center of the circle. The tension in the string is the only force and it causes an acceleration toward the center of the circle. The right panel shows the view from an accelerated frame in which the object is at rest. In this frame the tension in the string balances the centrifugal force, which is the inertial force arising from being in an accelerated reference frame, leaving zero net force.

We now demonstrate that the acceleration of the object is toward the center of the circle.

The acceleration in this special case is called the *centripetal acceleration*.

Figure 6.3 shows the position of the object at two times spaced by the time interval Δt.

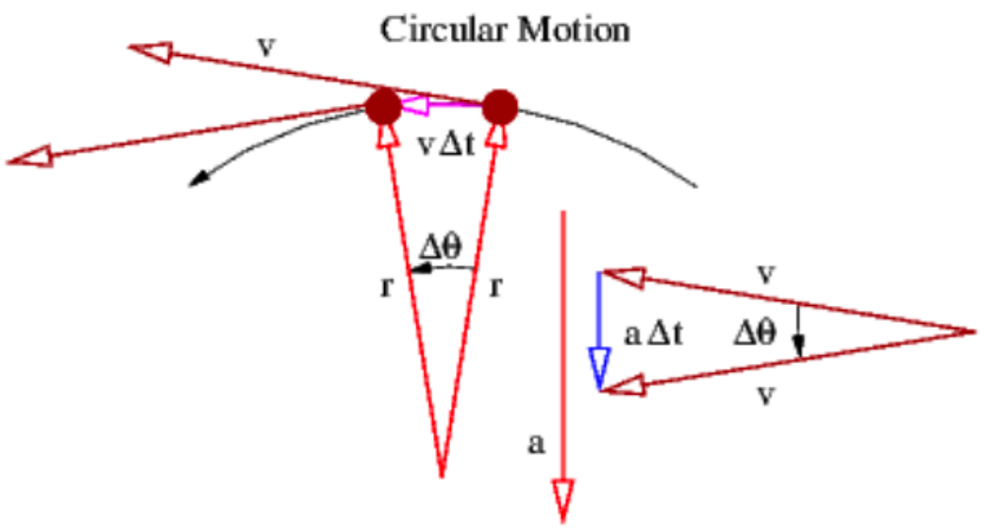


Figure 6.3: Definition sketch for computing centripetal acceleration.

The position vector of the object relative to the center of the circle rotates through an angle $\Delta\theta$ during this interval, so the angular rate of revolution of the object about the center is $\omega = \Delta\theta/\Delta t$.

The magnitude of the velocity of the object is v, so the object moves a distance $v\Delta t$ during the time interval.

To the extent that this distance is small compared to the radius r of the circle, the angle $\Delta\theta = v\Delta t/r$.

Solving for v and using $\omega = \Delta\theta/\Delta t$, we see that

$$v = \omega r \quad (\text{circular motion}). \tag{6.6}$$

The direction of the velocity vector changes over this interval, even though the magnitude v stays the same.

Figure 6.3 shows that this change in direction implies an acceleration a which is directed toward the center of the circle, as noted above.

The magnitude of the vectoral change in velocity in the time interval Δt is $a\Delta t$.

Since the angle between the initial and final velocities is the same as the angle $\Delta\theta$ between the initial and final radius vectors, we see from the geometry of the triangle in figure 6.3 that $a\Delta t/v = \Delta\theta$.

Solving for a results in

$$a = \omega v \quad (\text{circular motion}). \tag{6.7}$$

Combining equations (6.6) and (6.7) yields the equation for centripetal acceleration:

$$a = \omega^2 r = v^2/r \quad \text{(centripetal acceleration)}. \tag{6.8}$$

The second form is obtained by eliminating ω from the first form using equation (6.6).

6.3 Acceleration in Special Relativity

As noted above, acceleration is just the time rate of change of velocity.

We use the above results to determine how acceleration transforms from one reference frame to another.

Figure 6.4 shows the world line of an *accelerated reference frame*, with a time-varying velocity $U(t)$ relative to the unprimed inertial rest frame.

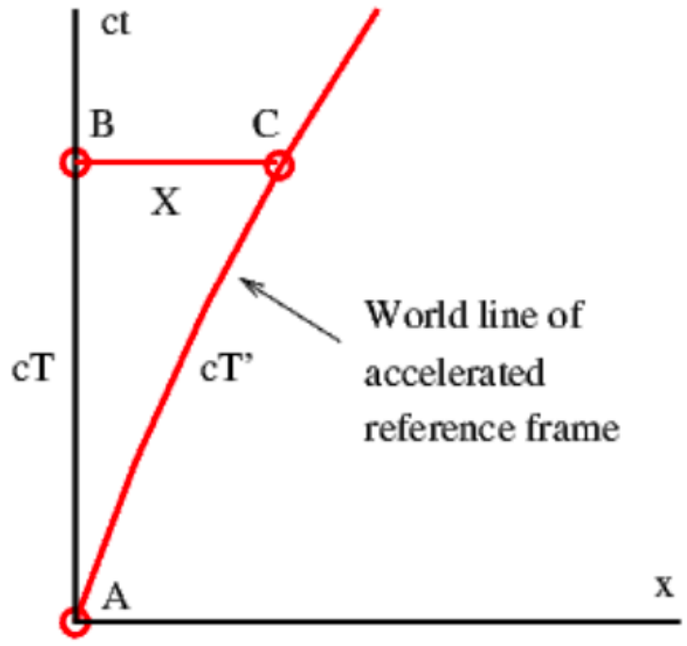


Figure 6.4: World line of the origin of an accelerated reference frame.

Defining $\Delta U = U(T) - U(0)$ as the change in the velocity of the accelerated frame (relative to the unprimed frame) between events A and C, we can relate this to the change of velocity, $\Delta U'$, of the accelerated frame relative to an inertial frame moving with the initial velocity, U(0).

Applying the equation for the relativistic addition of velocities, we find

$$U(T) = U(0) + \Delta U = \frac{U(0) + \Delta U'}{1 + U(0)\Delta U'/c^2}. \tag{6.9}$$

We now note that the mean acceleration of the reference frame between events A and C in the rest frame is just $a = \Delta U/T$, whereas the mean acceleration in the primed frame between the same two events is $a' = \Delta U'/T'$.

From equation (6.9) we find that

$$\Delta U = \frac{\Delta U'[1 - U(0)^2/c^2]}{1 + U(0)\Delta U'/c^2}, \tag{6.10}$$

and the acceleration of the primed reference frame as it appears in the unprimed frame is

$$a = \frac{\Delta U}{T} = \frac{\Delta U'[1 - U(0)^2/c^2]}{T[1 + U(0)\Delta U'/c^2]}. \tag{6.11}$$

Since we are interested in the instantaneous rather than the average acceleration, we let T become small.

This has three consequences.

First, ΔU and $\Delta U'$ become small, which means that the term $U(0)\Delta U'/c^2$ in the denominator of equation (6.11) can be ignored compared to 1.

This means that

$$a \approx \frac{\Delta U'[1 - U(0)^2/c^2]}{T}, \tag{6.12}$$

with the approximation becoming exact as $T \to 0$.

Second, the "triangle" with the curved side in figure 6.4 becomes a true triangle, with the result that $T' = T[1 - U(0)^2/c^2]^{1/2}$.

The acceleration of the primed frame with respect to an inertial frame moving at speed $U(0)$ can therefore be written

$$a' = \frac{\Delta U'}{T'} = \frac{\Delta U'}{T[1 - U(0)^2/c^2]^{1/2}}. \tag{6.13}$$

Third, we can replace $U(0)$ with U, since the velocity of the accelerated frame doesn't change very much over a short time interval.

Dividing equation (6.12) by equation (6.13) results in a relationship between the two accelerations:

$$a = a'(1 - U^2/c^2)^{3/2}, \tag{6.14}$$

which shows that the acceleration of a rapidly moving object, a, as observed from the rest frame, is less than its acceleration relative to an inertial reference frame in which the object is nearly stationary, a', by the factor $(1 - U^2/c^2)^{3/2}$.

We call this latter acceleration the *intrinsic acceleration*.

This difference in observed acceleration between the two inertial reference frames is purely the result of the geometry of spacetime, but it has interesting consequences.

Identifying a with dU/dt, we can integrate the acceleration equation assuming that the intrinsic acceleration a' is constant and that the velocity $U = 0$ at time $t = 0$.

We get the following result (verify this by differentiating with respect to time):

$$a't = \frac{U}{(1 - U^2/c^2)^{1/2}}, \tag{6.15}$$

which may be solved for U/c:

$$\frac{U}{c} = \frac{a't/c}{[1 + (a't/c)^2]^{1/2}}. \tag{6.16}$$

This is plotted in figure 6.5.

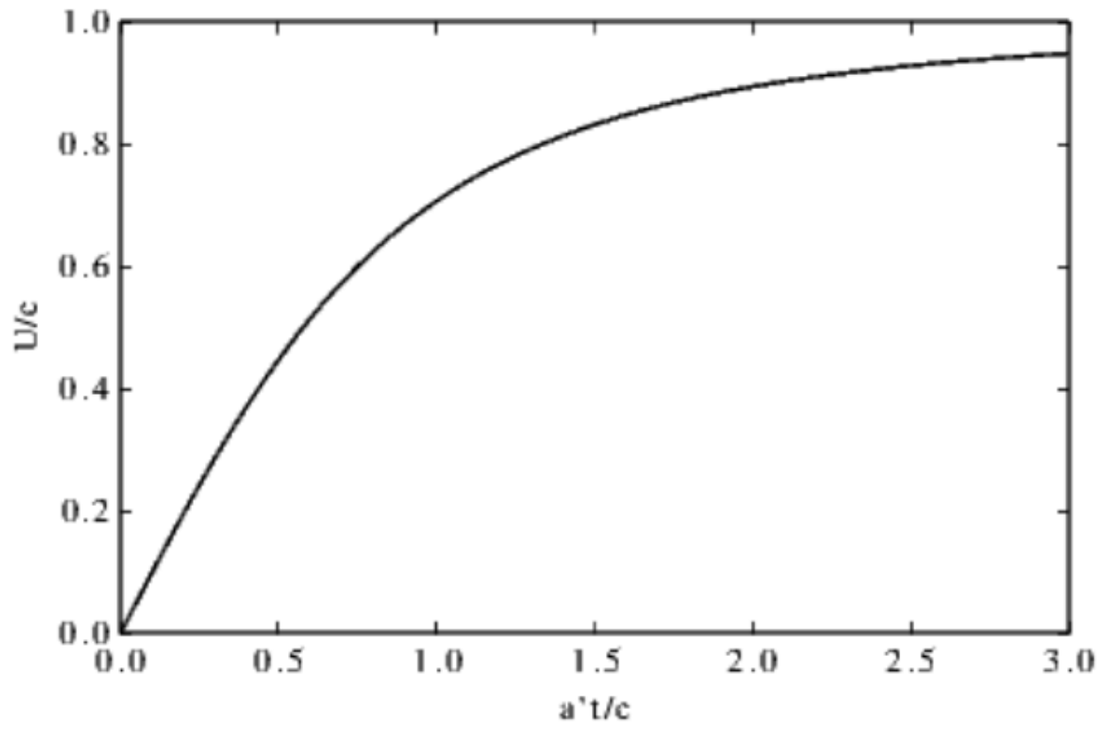


Figure 6.5: Velocity divided by the speed of light as a function of the product of the time and the (constant) acceleration divided by the speed of light.

Classically, the velocity U would reach the speed of light when $a't/c = 1$.

However, as figure 6.5 shows, the rate at which the velocity increases with time slows as the object moves faster, such that U approaches c asymptotically, but never reaches it.

The results for this section are valid only for acceleration components in the direction of motion.

The components perpendicular to this direction behave differently and are treated in more advanced texts.

6.4 Acceleration, Force, and Mass

We have a good intuitive feel for the concepts of force and mass because they are very much a part of our everyday experience.

We think of force as how hard we push on something.

Mass is the resistance of an object to acceleration if it is otherwise free to move.

Thus, pushing on a bicycle on a smooth, level road causes it to accelerate more readily than pushing on a car.

We say that the car has greater mass.

We can summarize this relationship with Newton's second law

$$F = ma \tag{6.17}$$

where F is the total force on an object, m is its mass, and a is the acceleration resulting from the force.

Three provisos apply to equation (6.17).

First, it only makes sense in unmodified form when the velocity of the object is much less than the speed of light.

For relativistic velocities it is best to write this equation in a slightly different form which we introduce later.

Second, the force must be the total force, including all frictional and other incidental forces which might otherwise be neglected by an uncritical observer.

Third, it only works in a reference frame which itself is unaccelerated.

We deal below with accelerated reference frames.

6.5 Accelerated Reference Frames

Referring back to the forces being felt by the occupant of a car, it is clear that the forces associated with accelerations are directed opposite the accelerations and proportional to their magnitudes.

For instance, when accelerating away from a stoplight, the acceleration is forward and the perceived force is backward.

When turning a corner, the acceleration is toward the corner while the perceived force is away from the corner.

Such forces are called *inertial forces*.

The origin of these forces can be understood by determining how acceleration changes when one observes it from a reference frame which is itself accelerated.

Suppose that the primed reference frame is accelerating to the right with acceleration *A* relative to the unprimed frame.

The position x' in the primed frame can be related to the position x in the unprimed frame by

$$x = x' + X, \tag{6.18}$$

where *X* is the position of the origin of the primed frame in the unprimed frame.

Taking the second time derivative, we see that

$$a = a' + A, \tag{6.19}$$

where $a = d^2x/dt^2$ is the acceleration in the unprimed frame and a' to an observer in the primed frame.

We now substitute this into equation (6.17) and move the term involving *A* to the left side:

$$F - mA = ma'. \tag{6.20}$$

This shows that Newton's second law represented by equation (6.17) is not valid in an accelerated reference frame, because the total force *F* and the acceleration a' in this frame don't balance as they do in the unaccelerated frame; the additional term -*mA* messes up this balance.

We can fix this problem by considering -*mA* to be a type of force, in which case we can include it as a part of the total force *F*.

This is the inertial force which we mentioned above.

Thus, to summarize, we can make Newton's second law work when objects are observed from accelerated reference frames if we include as part of the total force an inertial force which is equal to -*mA*, *A* being the acceleration of the reference frame of the observer and *m* the mass of the object being observed.

The right panel of figure 6.2 shows the inertial force observed in the reference frame of an object moving in circular motion at constant speed.

In the case of circular motion the inertial force is called the *centrifugal force*.

It points away from the center of the circle and just balances the tension in the string.

This makes the total force on the object zero in its own reference frame, which is necessary since the object cannot move (or accelerate) in this frame.

General relativity says that gravity is nothing more than an inertial force.

This was called the *equivalence principle* by Einstein.

Since the gravitational force on the Earth points downward, it follows that we must be constantly accelerating upward as we stand on the surface of the Earth!

The obvious problem with this interpretation of gravity is that we don't appear to be moving away from the center of the Earth, which would seem to be a natural consequence of such an acceleration.

However, relativity has a surprise in store for us here.

It follows from the above considerations that something can be learned about general relativity by examining the properties of accelerated reference frames.

In particular, we can gain insight into the above apparent paradox.

Equation (6.16) shows that the velocity of an object undergoing constant intrinsic acceleration a (note that we have dropped the "prime" from a to simplify the notation) is

$$v = \frac{dx}{dt} = \frac{at}{[1 + (at/c)^2]^{1/2}}, \tag{6.21}$$

where t is the time and c is the speed of light.

A function $x(t)$ which satisfies equation (6.21) is

$$x(t) = (c^2/a)[1 + (at/c)^2]^{1/2}. \tag{6.22}$$

You can verify this by differentiating it.

The interval OB in figure 6.6 is of length $x(0) = c^2/A$.

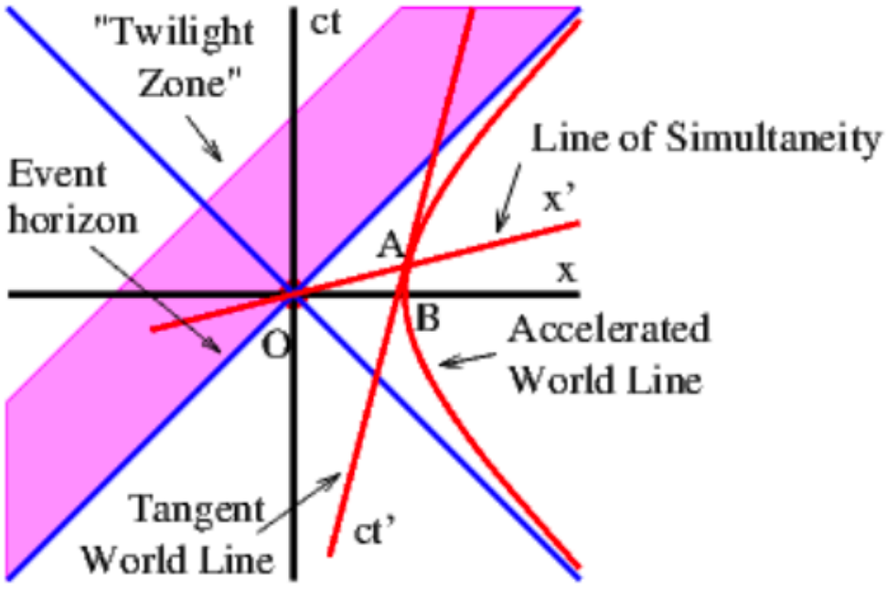


Figure 6.6: Spacetime diagram showing the world line of the origin of a reference frame undergoing constant acceleration.

The slanted line OA is a line of simultaneity associated with the unaccelerated world line tangent to the accelerated world line at point A.

This line of simultaneity goes through the origin, as is shown in figure 6.6.

To demonstrate this, multiply equations (6.21) and (6.22) together and solve for v/c:

$$v/c = ct/x. \tag{6.23}$$

From figure 6.6 we see that ct/x is the slope of the line OA, where (x, ct) are the coordinates of event A.

Equation (6.23) shows that this line is indeed the desired line of simultaneity, since its slope is the inverse of the slope of the world line, c/v.

Since there is nothing special about the event A, we infer that *all* lines of simultaneity associated with the accelerated world line pass through the origin.

We now inquire about the length of the invariant interval OA in figure 6.6.

Recalling that $I^2 = x^2$ and using equation (6.22), we find that the length of OA is

$$I = (x^2 - c^2t^2)^{1/2} = (c^4/a^2)^{1/2} = c^2/a, \tag{6.24}$$

which is the same as the length of the interval OB.

By extension, *all* events on the accelerated world line are the same invariant interval from the origin.

Recalling that the interval along a line of simultaneity is the distance in the associated reference frame, we reach the astonishing conclusion that even though the object associated with the curved world line in figure 6.6 is accelerating away from the origin, it always remains the same distance (in its own frame) from the origin.

The analogy between this problem and the apparent paradox in which one remains a fixed distance from the center of the earth while accelerating away from it is not perfect.

In particular, the center of the earth as it exists is not analogous to point O in figure 6.6 as it is not an event horizon.

The analogy would be closer if the earth's entire mass were compressed into a very small volume.

In this case the earth would be a black hole with an event horizon corresponding to the event horizon in figure 6.6.

6.6 Gravitational Red Shift

Light emitted at a lower level in a gravitational field has its frequency reduced as it travels to a higher level.

This phenomenon is called the *gravitational red shift*.

Figure 6.7 shows why this happens.

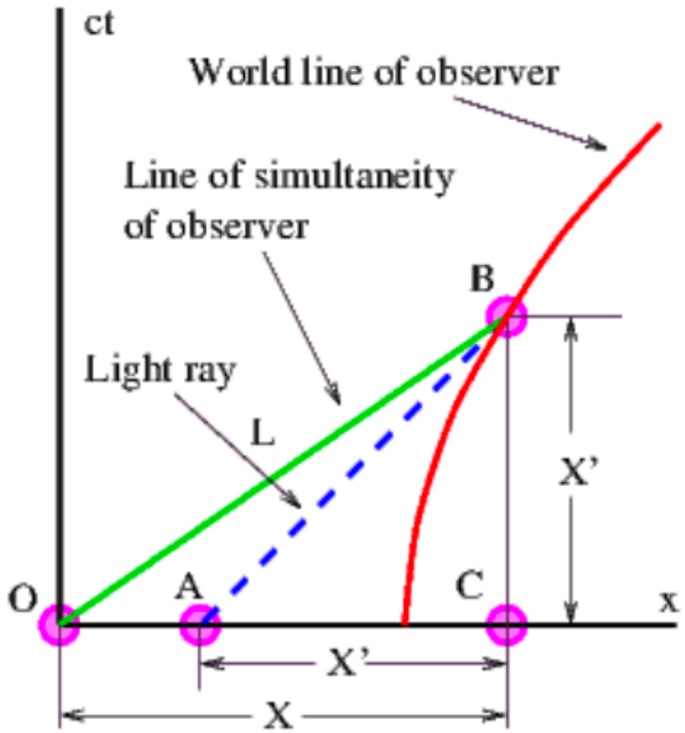


Figure 6.7: Spacetime diagram for explaining the gravitational red shift. Why is the interval AC equal to the interval BC? *L* is the length of the invariant interval OB.

Since experiencing a gravitational force is equivalent to being in an accelerated reference frame, we can use the tools of special relativity to view the process of light emission and absorption from the point of view of the unaccelerated or inertial frame.

In this reference frame the observer of the light is accelerating to the right, as indicated by the curved world line in figure 6.7, which is equivalent to a gravitational force to the left.

The light is emitted at point A with frequency ω by a source which is stationary at this instant.

At this instant the observer is also stationary in this frame.

However, by the time the light gets to the observer, he or she has a velocity to the right which means that the observer measures a Doppler shifted frequency ω' for the light.

Since the observer is moving away from the source, $\omega' < \omega$, as indicated above.

The relativistic Doppler shift is given by

$$\frac{\omega'}{\omega} = \left(\frac{1 - U/c}{1 + U/c}\right)^{1/2}, \tag{6.25}$$

so we need to compute U/c.

The line of simultaneity for the observer at point B goes through the origin, and is thus given by line segment OB in figure 6.7.

The slope of this line is U/c, where U is the velocity of the observer at point B.

From the figure we see that this slope is also given by the ratio X'/X.

Equating these, eliminating X in favor of $L = (X^2 - X'^2)^{1/2}$, which is the actual invariant distance of the observer from the origin, and substituting into equation (6.25) results in our gravitational red shift formula:

$$\frac{\omega'}{\omega} = \left(\frac{X - X'}{X + X'}\right)^{1/2} = \left(\frac{(L^2 + X'^2)^{1/2} - X'}{(L^2 + X'^2)^{1/2} + X'}\right)^{1/2} . \qquad (6.26)$$

If $X' = 0$, then there is no redshift, because the source is collocated with the observer.

On the other hand, if the source is located at the origin, so $X' = X$, the Doppler shifted frequency is zero.

In addition, the light never gets to the observer, since the world line is asymptotic to the light world line passing through the origin.

If the source is at a higher level in the gravitational field than the observer, so that $X' < 0$, then the frequency is shifted to a higher value, i. e., it becomes a "blue shift".

6.7 Event Horizons

The 45° diagonal line passing through the origin in figure 6.6 is called the *event horizon* for the accelerated observer in this figure.

Notice that light from the "twilight zone" above and to the left of the event horizon cannot reach the accelerated observer.

However, the reverse is not true — a light signal emitted to the left by the observer can cross the event horizon into the "twilight zone".

The event horizon thus has a peculiar one-way character — it passes signals from right to left, but not from left to right.

Chapter 7

Matter Waves

We begin our study of quantum mechanics by discussing the diffraction undergone by X-rays and electrons when they interact with a crystal.

X-rays are a form of electromagnetic radiation with wavelengths comparable to the distances between atoms.

Scattering from atoms in a regular crystalline structure results in an interference pattern which is in many ways similar to the pattern from a diffraction grating.

We first develop Bragg's law for diffraction of X-rays from a crystal.

Two practical techniques for doing X-ray diffraction are then described.

It turns out that electrons have wave-like properties and also undergo Bragg diffraction by crystals.

Bragg diffraction thus provides a crucial bridge between the worlds of waves and particles.

With this bridge we introduce the classical ideas of momentum and energy by relating them to the wave vector and frequency of a wave. The properties of waves also give rise to the Heisenberg uncertainty principle.

Table 7.1 shows a table of the Nobel prizes associated with the ideas presented in this chapter.

This gives us a feel for the chronology of these discoveries and indicates how important they were to the development of physics in the early 20th century.

Year	Recipient	Contribution
1901	W. K. Röntgen	Discovery of X-rays
1906	J. J. Thomson	Discovery of electron
1914	M. von Laue	X-ray diffraction in crystals
1915	W. and L. Bragg	X-ray analysis of crystal structure
1918	M. Planck	Energy quantization
1921	A. Einstein	Photoelectric effect
1922	N. Bohr	Structure of atoms
1929	L.-V. de Broglie	Wave nature of electrons
1932	W. Heisenberg	Quantum mechanics
1933	Schrödinger and Dirac	Atomic theory
1937	Davisson and Thomson	Electron diffraction in crystals

Table 7.1: Selected Nobel prize winners, year of award, and contribution.

7.1 Bragg's Law

Figure 7.1 schematically illustrates interference between waves scattering from two adjacent rows of atoms in a crystal.

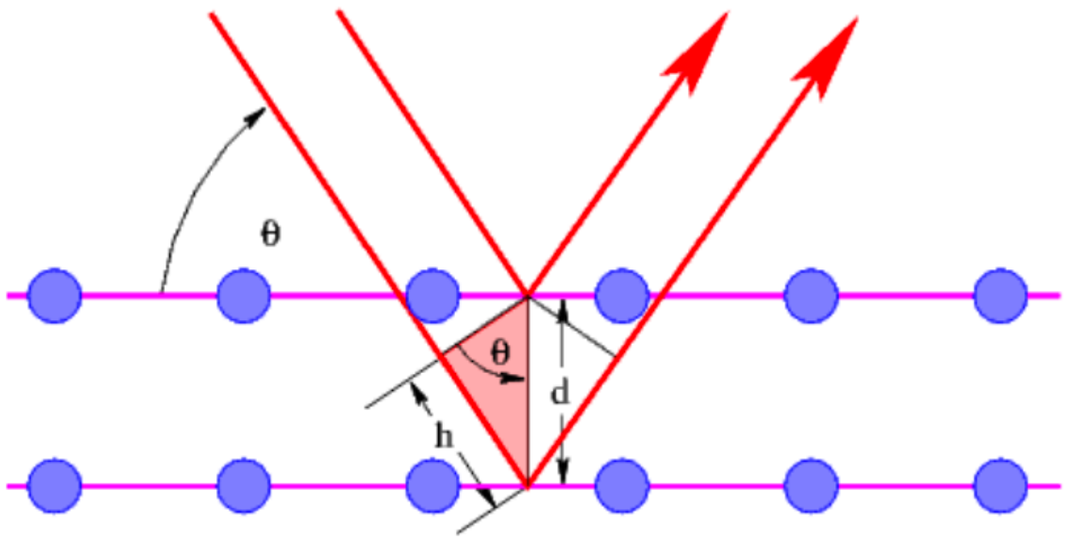


Figure 7.1: Schematic diagram for determining Bragg's law.

The net effect of scattering from a single row is equivalent to partial reflection from a mirror imagined to be aligned with the row.

Thus, the angle of "reflection" equals the angle of incidence for each row.

Interference then occurs between the beams reflecting off different rows of atoms in the crystal.

For the two adjacent rows shown in figure 7.1, the path difference between beams is $2h = 2d \sin\theta$.

For constructive interference this must be an integer number of wavelengths, $m\lambda$, where the integer m is called the *order of interference*.

The result is Bragg's law of diffraction:

$$m\lambda = 2d\sin\theta, \quad m = 1, 2, 3 \ldots \quad \text{(Bragg's law)}. \tag{7.1}$$

If only two rows are involved, the transition from constructive to destructive interference as θ changes is gradual.

However, if interference from many rows occurs, then the constructive interference peaks become very sharp with mostly destructive interference in between.

This sharpening of the peaks as the number of rows increases is very similar to the sharpening of the diffraction peaks from a diffraction grating as the number of slits increases.

7.2 X-Ray Diffraction Techniques

Two types of targets are used in Bragg diffraction experiments: single crystals and powder targets.

7.2.1 Single Crystal

In a single crystal setup, an X-ray detector is mounted as shown in figure 7.2.

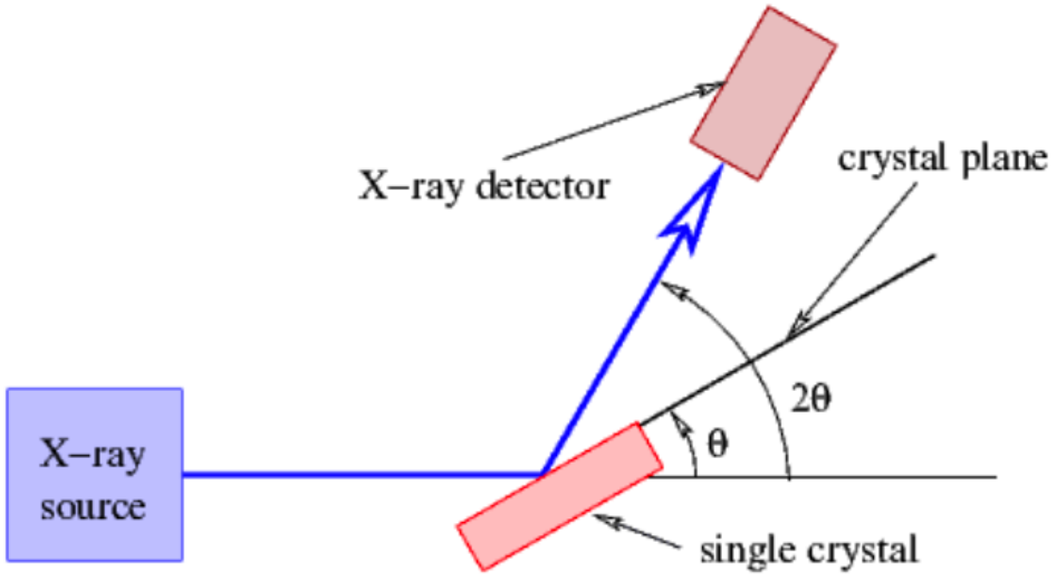


Figure 7.2: Setup for single crystal Bragg diffraction.

A mechanical device keeps the detector oriented so that the angle of incidence equals the angle of reflection for the desired crystal plane.

Peaks in the X-ray detection rate are sought as the angle θ is varied.

The advantage of this type of apparatus is that diffraction peaks from only the selected crystal plane are observed.

7.2.2 Powder Target

The powder in a powder target is really a conglomeration of many tiny crystals randomly oriented.

Thus, for each possible Bragg diffraction angle there are crystals oriented correctly for Bragg diffraction to take place.

The detector is usually a photographic plate or an equivalent electronic device as illustrated in figure 7.3.

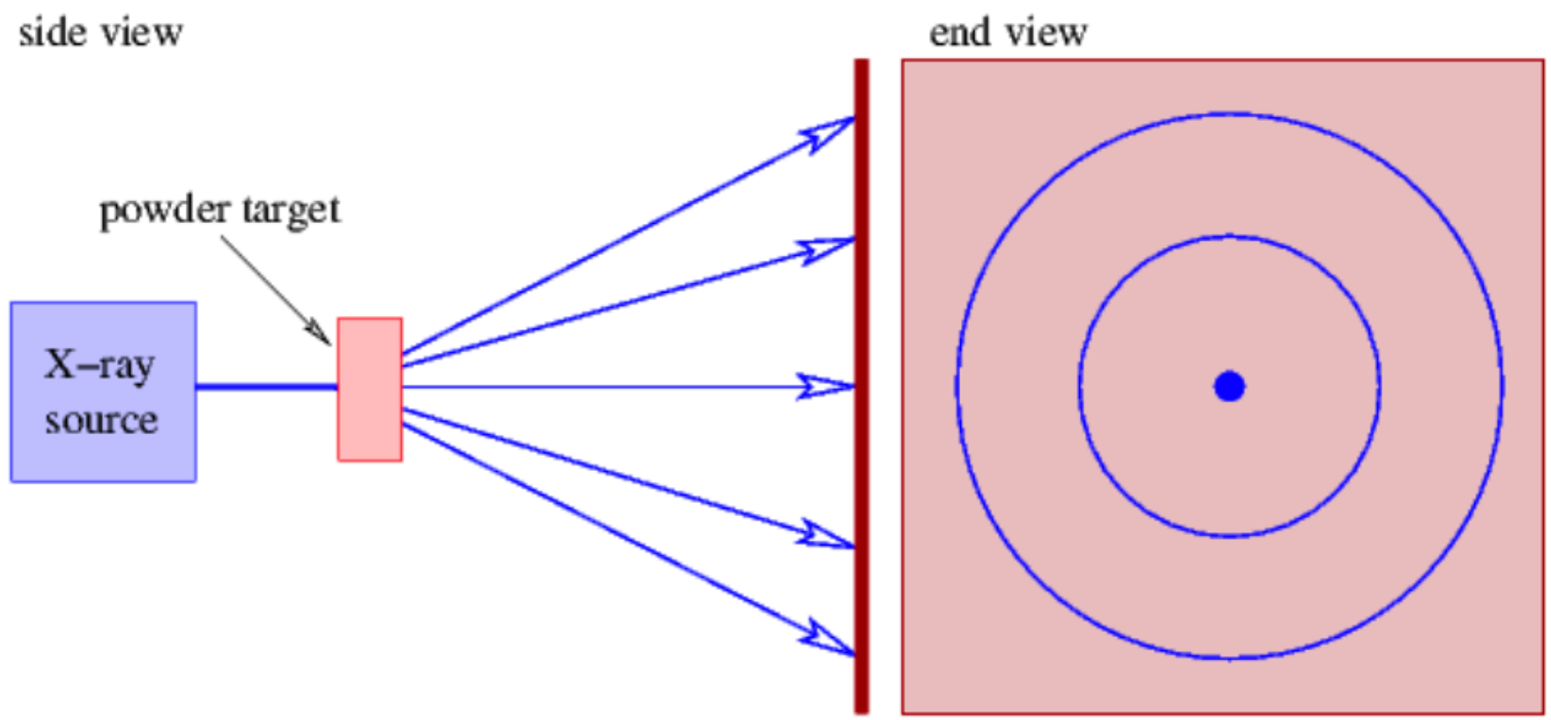


Figure 7.3: Setup for powder target Bragg diffraction.

For each Bragg diffraction angle one sees a ring on the plate concentric with the axis of the incident X-ray beam.

The advantage of this type of system is that no knowledge is needed of the crystal plane orientations.

Furthermore, a single large crystal is not required.

However, all possible Bragg scattering angles are seen at once, which can lead to confusion in the interpretation of the results.

7.3 Meaning of Quantum Wave Function

Bragg diffraction illustrates the most difficult thing to understand about quantum mechanics, namely that particles can have wave-like properties and waves can have particle-like properties.

The variation of X-ray intensity with angle seen in a Bragg diffraction apparatus is very difficult to explain in any terms other than wave interference.

Yet, X-rays are typically detected by a device such as a Geiger counter which produces a pulse of electricity for each X-ray particle, or *photon*, which hits it.

Thus, X-rays sometimes act like particles and sometimes like waves.

Light isn't alone in having both particle and wave properties.

Davisson and Germer and later G. P. Thomson (son of J. J. Thomson, the discoverer of the electron) showed that electrons also can act like waves.

They did this by demonstrating that electrons undergo Bragg diffraction in crystals, much as X-rays do.

Most physicists (including Albert Einstein) have found quantum mechanics to be extremely bizarre, so if you feel the same way, you are in good company!

However, there is a useful interpretation of quantum mechanics which at least allows us to get on with using it to solve problems, even though it may not satisfy our intuitive reservations about the theory.

The displacement of the matter wave associated with a particle is usually called the wave function, ψ.

It is not at all clear what ψ is a displacement *of*, but its use is straightforward.

The absolute square of the wave function, $|\psi(x,t)|^2$, is proportional to the *probability* of finding the associated particle at position x and time t.

The *absolute* square is taken because under many circumstances the wave function is actually *complex*, i. e., it has both real and imaginary parts.

The reasons for this will be discussed later.

Due to the interpretation of the wave function, quantum mechanics is a probabilistic theory.

It does not tell us with certainty what happens to a particular particle.

Instead, it tells us the probability for detecting the particle in any given location.

If many experiments are done, with one particle per experiment, the numbers of experiments with particles being detected in the various possible locations are in proportion to the quantum mechanical probabilities.

7.4 Sense and Nonsense in Quantum Mechanics

The essential mystery of quantum mechanics becomes clearer when discussed in the conceptually simpler context of two slit interference.

If light and electrons can have both particle and wave properties, then one might ask through which of the two slits the particle passed.

However, in physics a question simply doesn't make sense if it cannot be answered by experiment.

One can indeed perform an experiment to determine which slit the X-ray photon or electron passes through in the two slit experiment.

However, by the very act of making this measurement, the form of the associated wave is altered.

In particular, since the absolute square of the wave displacement represents the probability of finding the particle, once the particle has definitely been found passing through one or the other of the slits, the wave function collapses into a very small wave packet located at the observed position of the particle.

Thus, the wave displacement becomes zero at the slit it didn't pass through.

However, the interference pattern results from the superposition of waves emanating from two slits.

If no wave comes from one of the slits (because the wave displacement is zero there), then there can be no interference pattern!

We can now make the inverse argument.

If there is an interference pattern, then we know that the wave displacement is non-zero at both slits.

From the probability interpretation of the wave displacement, we conclude that we cannot say, even in principle, through which slit the particle passed.

It is not just that we don't know the answer to this question; there is simply no experiment which can give us an answer without destroying the interference pattern.

In other words, the question "Which slit did the particle pass through?" is a nonsensical question in the case where an interference pattern is actually produced.

The American physicist Richard Feynman noticed that the above behavior can be interpreted as violating the normal laws of probability.

These laws say that the probability of an event is the sum of the probabilities of alternate independent ways for that event to occur.

For instance, the probability for a particle to reach point A on the detection screen of a two slit setup is just the probability P_1 for the particle to reach point A after going through slit 1, plus the probability P_2 for the particle to reach point A after going through slit 2.

Thus, if $P_1 = P_2 = 0.1$, then the probability for the particle to reach point A irrespective of which slit it went through should be $P_{total} = P_1 + P_2 = 0.2$.

However, if point A happens to be a point of destructive interference, then we know that $P_{total} = 0$.

Feynman proposed that the above rule stating that alternate independent probabilities add, is simply incorrect.

In its place Feynman asserted that *probability amplitudes* add instead, where the probability amplitude in this case is just the wave function associated with the particle.

The total probability for a process is obtained by adding the alternate probability amplitudes together and taking the absolute square of the sum.

Feynman's view is a particularly compact expression of the so-called *Copenhagen interpretation* of quantum mechanics which evolved from the ideas of Niels Bohr, Werner Heisenberg, Max Born, and others in the 1920s.

It dispenses with the wave-particle duality and other philosophical baggage by saying "Particles are real in that we can observe them, but the only theory we have is about probability amplitudes for particles."

This interpretation of quantum mechanics may be weird, but it appears to be self-consistent and in agreement with experiment.

7.5 Mass, Momentum, and Energy

In this section we relate the classical ideas of mass, momentum and energy to what we have done so far.

Historically, these connections were first made by Max Planck and Louis de Broglie with help from Albert Einstein.

Bragg diffraction of electrons is invoked as an experimental test of the Planck and de Broglie relations.

Technically, we don't need the ideas of mass, momentum, and energy to do physics – the notions of wavenumber, frequency, and group velocity are sufficient to describe and explain all observed phenomena.

However, mass, momentum, and energy are so firmly embedded in physics that one couldn't talk to other physicists without an understanding of these quantities!

7.5.1 Planck, Einstein, and de Broglie

Max Planck was the first to develop a theory explaining the energy density of electromagnetic radiation in a box at a fixed temperature.

Albert Einstein extended Planck's ideas by postulating that the energy of electromagnetic radiation is quantized into chunks called *photons*.

The energy E of a photon is related to the frequency of the electromagnetic radiation by the equation

$$E = hf = \hbar\omega \quad \text{(Planck-Einstein relation)}, \tag{7.2}$$

where f is the rotational frequency of the associated electromagnetic wave and ω is its angular frequency.

The constant $h = 6.63 \times 10^{-34}$ kg m^2 s^{-1} is called *Planck's constant*.

The related constant $\hbar = h/2\pi = 1.06 \times 10^{-34}$ kg m^2 s^{-1} is also referred to as Planck's constant, but to avoid confusion with the original constant, we will generally refer to it as "h bar"

Notice that a new physical dimension has appeared, namely mass, with the unit *kilogram*, abbreviated "kg".

The physical meaning of mass is much like our intuitive understanding of the concept, i. e., as a measure of the resistance of an object to its velocity being changed.

The precise scientific meaning will emerge shortly.

Einstein showed that Planck's idea could be used to explain the emission of electrons which occurs when light impinges on the surface of a metal.

This emission, which is called the *photoelectric effect*, can only occur when electrons are supplied with a certain minimum energy E_B required to break them loose from the metal.

Experiment shows that this emission occurs only when the frequency of the light exceeds a certain minimum value.

This value turns out to equal $\omega_{min} = E_B/\hbar$, which suggests that electrons gain energy by absorbing a single photon.

If the photon energy, $\hbar\omega$, exceeds E_B, then electrons are emitted, otherwise they are not.

It is much more difficult to explain the photoelectric effect from the classical theory of light.

The value of E_B, called the binding energy or work function, is different for different metals.

Louis de Broglie proposed that Planck's energy-frequency relationship be extended to all kinds of particles.

In addition he hypothesized that the *momentum* $\mathbf{\Pi}$ of the particle and the wave vector $\mathbf{k}$ of the corresponding wave were similarly related:

$$\boldsymbol{\Pi} = \hbar \mathbf{k} \quad \text{(de Broglie relation)}. \tag{7.3}$$

Note that this can also be written in scalar form in terms of the wavelength as $\Pi = h/\lambda$.

We use $\boldsymbol{\Pi}$ rather than the more common $\mathbf{p}$ for momentum, because as we shall see, there are two different kinds of momentum, one related to the wavenumber, the other related to the velocity of a particle.

In many cases they are equal, but there are certain important situations in which they are not.

De Broglie's hypothesis was inspired by the fact that wave frequency and wavenumber are components of the same four-vector according to the theory of relativity, and are therefore closely related to each other.

Thus, if the energy of a particle is related to the frequency of the corresponding wave, then there ought to be some similar quantity which is correspondingly related to the wavenumber.

It turns out that the momentum is the appropriate quantity.

The physical meaning of momentum will become clear as we proceed.

We will also find that the rest frequency, μ, of a particle is related to its mass, m:

$$E_{rest} \equiv mc^2 = \hbar\mu. \tag{7.4}$$

The quantity E_{rest} is called the *rest energy* of the particle.

From our perspective, energy, momentum, and rest energy are just scaled versions of frequency, wave vector, and rest frequency, with a scaling factor $\hbar$.

We can therefore define a *four-momentum* as a scaled version of the wave four-vector:

$$\underline{\Pi} = \hbar \underline{k}. \tag{7.5}$$

The spacelike component of $\underline{\Pi}$ is just $\mathbf{\Pi}$, while the timelike part is E/c.

Planck, Einstein, and de Broglie had extensive backgrounds in classical mechanics, in which the concepts of energy, momentum, and mass have precise meaning.

In this text we do not presuppose such a background.

Perhaps the best strategy at this point is to think of these quantities as scaled versions of frequency, wavenumber, and rest frequency, where the scale factor is $\hbar$.

The significance of these quantities to classical mechanics will emerge bit by bit.

7.5.2 Wave and Particle Quantities

Let us now recapitulate what we know about relativistic waves, and how this knowledge translates into knowledge about the mass, energy, and momentum of particles.

In the following equations, the form on the left is expressed in wave terms, i. e., in terms of frequency, wavenumber, and rest frequency.

The form on the right below is the identical equation expressed in terms of energy, momentum, and mass.

Since the latter variables are just scaled forms of the former, the two forms of each equation are equivalent.

We begin with the dispersion relation for relativistic waves:

$$\omega^2 = k^2c^2 + \mu^2 \qquad E^2 = \Pi^2c^2 + m^2c^4. \tag{7.6}$$

Calculation of the group velocity, $u_g = d\omega / dk$, from the dispersion relation yields

$$u_g = \frac{c^2k}{\omega} \qquad u_g = \frac{c^2\Pi}{E}. \tag{7.7}$$

These two sets of equations represent what we know about relativistic waves, and what this knowledge tells us about the relationships between the mass, energy, and momentum of relativistic particles.

When in doubt, refer back to these equations, as they work in all cases, including for particles with zero mass!

It is useful to turn equations (7.6) and (7.7) around so as to express the frequency as a function of rest frequency and group velocity,

$$\omega = \frac{\mu}{(1 - u_g^2/c^2)^{1/2}} \qquad E = \frac{mc^2}{(1 - u_g^2/c^2)^{1/2}}, \tag{7.8}$$

and the wavenumber as a similar function of these quantities:

$$k = \frac{\mu u_g/c^2}{(1-u_g^2/c^2)^{1/2}} \qquad \Pi = \frac{m u_g}{(1-u_g^2/c^2)^{1/2}}. \tag{7.9}$$

Note that equations (7.8) and (7.9) work only for particles with non-zero mass!

For zero mass particles both the numerators and denominators of equations (7.8) and (7.9) are zero, making these equations undefined, and you need to use equations (7.6) and (7.7) with $m = 0$ and $\mu = 0$ instead.

The quantity ω - μ indicates how much the frequency exceeds the rest frequency.

Notice that if $\omega = \mu$, then from equation (7.6) $k = 0$.

Thus, positive values of $\omega_k \equiv \omega$ - μ indicate $|k| > 0$, which means that the particle is moving according to equation (7.7).

Let us call ω_k the *kinetic frequency*:

$$\omega_k = \left[\frac{1}{(1-u_g^2/c^2)^{1/2}} - 1\right]\mu \qquad K = \left[\frac{1}{(1-u_g^2/c^2)^{1/2}} - 1\right] mc^2. \tag{7.10}$$

We call K the *kinetic energy* for similar reasons.

Again, equation (7.10) only works for particles with non-zero mass.

For zero mass particles the kinetic energy equals the total energy.

Note that the results of this section are valid only for free particles, i. e., particles to which no force is applied.

Force in classical and quantum mechanics is treated in the next chapter.

7.5.3 Non-Relativistic Limits

When the mass is non-zero and the group velocity is much less than the speed of light, it is useful to compute approximate forms of the above equations valid in this limit.

Using the approximation $(1 + \epsilon)^x \approx 1 + x\epsilon$, we find that the dispersion relation becomes

$$\omega = \mu + \frac{k^2 c^2}{2\mu} \qquad E = mc^2 + \frac{\Pi^2}{2m}, \tag{7.11}$$

and the group velocity equation takes the approximate form

$$v_g = \frac{c^2 k}{\mu} \qquad v_g = \frac{\Pi}{m}. \tag{7.12}$$

The non-relativistic limits for equations (7.8) and (7.9) become

$$\omega = \mu + \frac{\mu u_g^2}{2c^2} \qquad E = mc^2 + \frac{mu_g^2}{2} \tag{7.13}$$

and

$$k = \mu u_g / c^2 \qquad \Pi = m u_g, \tag{7.14}$$

while the approximate kinetic energy equation is

$$\omega_k = \frac{\mu u_g^2}{2c^2} \qquad K = \frac{m u_g^2}{2}. \tag{7.15}$$

Just a reminder — the equations in this section are not valid for massless particles!

7.5.4 An Experimental Test

How can we test the above predictions against experiment?

The key point is to be able to relate the wave aspects to the particle aspects of a quantum mechanical wave-particle.

Equation (7.9), or equation (7.14) in the non-relativistic case, relates a particle's wavenumber k to its velocity u_g.

Both of these quantities can be measured in a Bragg's law experiment with electrons.

In this experiment electrons are fired at a crystal with known atomic dimensions at a known speed, which we identify with the group velocity u_g.

The Bragg angle which yields constructive interference can be used to calculate the wavelength of the corresponding electron wave, and hence the wavenumber and momentum.

If the momentum is plotted against group velocity in the non-relativistic case, a straight line should be found, the slope of which is the particle's mass.

In the fully relativistic case one needs to plot momentum versus $u_g / (1 - u_g^2/c^2)^{1/2}$.

Again, a straight line indicates agreement with the theory and the slope of the line is the particle's mass.

This particular experiment is difficult to do, but the corresponding theories verify in many other experiments.

7.6 Heisenberg Uncertainty Principle

Classically, we consider the location of a particle to be a knowable piece of information.

In quantum mechanics the position of a particle is well known if the wave packet representing it is small in size.

However, quantum mechanics imposes a price on accurately knowing the position of a particle in terms of the future predictability of its position.

This is because a small wave packet, which corresponds to accurate knowledge of the corresponding particle's position, implies the superposition of plane waves corresponding to a broad distribution of wavenumbers.

This translates into a large uncertainty in the wavenumber, and hence the momentum of the particle.

In contrast, a broad wave packet corresponds to a narrower distribution of wavenumbers, and correspondingly less uncertainty in the momentum.

Referring back to chapters 1 and 2, recall that both the longitudinal (along the direction of motion) and transverse (normal to the direction of motion) dimensions of a wave packet, Δx_L and Δx_T , can be related to the spread of longitudinal and transverse wavenumbers, Δk_L and Δk_T :

$$\Delta k_L \Delta x_L \approx 1, \tag{7.16}$$

$$\Delta k_T \Delta x_T \approx 1. \tag{7.17}$$

We have omitted numerical constants which are order unity in these approximate relations so as to show their essential similarity.

The above equations can be interpreted in the following way.

Since the absolute square of the wave function represents the probability of finding a particle, Δx_L and Δx_Trepresent the uncertainty in the particle's position.

Similarly, Δk_L and Δk_T represent the uncertainty in the particle's longitudinal and transverse wave vector components.

This latter uncertainty leads to uncertainty in the particle's future motion — larger or smaller longitudinal k results respectively in larger or smaller particle speed, while uncertainty in the transverse wavenumber results in uncertainty in the particle's direction of motion.

Thus uncertainties in any component of **k** result in uncertainties in the corresponding component of the particle's velocity, and hence in its future position.

The equations (7.16) and (7.17) show that uncertainty in the present and future positions of a particle are complimentary.

If the present position is accurately known due to the small size of the associated wave packet, then the future position is not very predictable, because the wave packet disperses rapidly.

On the other hand, a broad-scale initial wave packet means that the present position is poorly known, but the uncertainty in position, poor as it is, doesn't rapidly increase with time, since the wave packet has a small uncertainty in wave vector and thus disperses slowly.

This is a statement of the *Heisenberg* uncertainty principle.

The uncertainty principle also applies between frequency and time:

$$\Delta\omega\Delta t \approx 1. \tag{7.18}$$

This shows up in the beat frequency equation $1/T_{beat} = \Delta f = \Delta\omega/\ 2\pi$.

The beat period T_{beat} may be thought of as the size of a "wave packet in time".

The beat frequency equation may be rewritten as $\Delta\omega\, T_{beat} = 2\pi$, which is the same as equation (7.18) if the factor of 2π is ignored and T_{beat} is identified with Δt.

The above forms of the uncertainty principle are not relativistically invariant.

A useful invariant form may be obtained by transforming to the coordinate system in which a particle is stationary.

In this reference frame the time t becomes the proper time τ associated with the particle.

Furthermore, the frequency ω becomes the rest frequency μ.

The uncertainty principle thus becomes

$$\Delta\mu\Delta\tau \approx 1 \tag{7.19}$$

in this reference frame.

However, since $\Delta\mu$ and $\Delta\tau$ are relativistic invariants, this expression of the uncertainty principle is valid in *any* reference frame.

It is more common to express the uncertainty principle in terms of the mass, momentum, and energy by multiplying equations (7.16) - (7.19) by $\hbar$.

Lumping the momentum equations, we find

$$\Delta\Pi\Delta x \approx \hbar, \tag{7.20}$$

$$\Delta E \Delta t \approx \hbar, \tag{7.21}$$

and

$$\Delta(mc^2)\Delta\tau \approx \hbar. \tag{7.22}$$

Classical mechanics is the realm of quantum mechanics in which the dimensions of the system of interest are much larger than the wavelengths of the waves corresponding to the particles constituting the system.

In this case the uncertainties induced by the uncertainty principle are unimportant.

This limit is analogous to the geometrical optics limit for light.

Thus, we can say that classical mechanics is the geometrical optics limit of quantum mechanics.

Chapter 8

Geometrical Optics and Newton's Laws

The question that motivates us to study physics is "What makes things go?"

The answers we conceive to this question constitute the subject of *dynamics*.

This is in contrast to the question we have primarily addressed so far, namely "How do things go?" The latter question is about kinematics.

Extensive preparation in the kinematics of waves and particles in relativistic spacetime is needed to intelligently address dynamics.

This preparation is now complete.

In this chapter we outline three different dynamical principles based respectively on pre-Newtonian, Newtonian, and quantum mechanical thinking.

We first discuss the Newtonian mechanics of conservative forces in one dimension.

Certain ancillary concepts in mechanics such as work and power are introduced at this stage.

We then show that Newtonian and quantum mechanics are consistent with each other in the realm in which they overlap, i. e., in the geometrical optics limit of quantum mechanics.

For simplicity, this relationship is first developed in one dimension in the non-relativistic limit.

Higher dimensions require the introduction of partial derivatives, and the relativistic case will be considered later.

8.1 Fundamental Principles of Dynamics

Roughly speaking, there have been three eras of physics, characterized by three different answers to the question of what makes things go.

8.1.1 Pre-Newtonian Dynamics

Aristotle expounded a view of dynamics which agrees closely with our everyday experience of the world.

Objects only move when a force is exerted upon them.

As soon as the force goes away, the object stops moving.

The act of pushing a box across the floor illustrates this principle — the box certainly doesn't move by itself!

8.1.2 Newtonian Dynamics

In contrast to earthly behavior, the motions of celestial objects seem effortless.

No obvious forces act to keep the planets in motion around the sun.

In fact, it appears that celestial objects simply coast along at constant velocity unless something acts on them.

The Newtonian view of dynamics — objects change their *velocity* rather than their position when a force is exerted on them — is expressed by *Newton's* second law:

$$\mathbf{F} = m\mathbf{a} \quad \text{(Newton's second law)}, \tag{8.1}$$

where **F** is the force exerted on a body, *m* is its mass, and **a** is its acceleration.

Newton's first law, which states that an object remains at rest or in uniform motion unless a force acts on it, is actually a special case of Newton's second law which applies when $\mathbf{F} = 0$.

It is no wonder that the first successes of Newtonian mechanics were in the celestial realm, namely in the predictions of planetary orbits.

It took Newton's genius to realize that the same principles which guided the planets also applied to the earthly realm as well.

In the Newtonian view, the tendency of objects to stop when we stop pushing on them is simply a consequence of frictional forces opposing the motion.

Friction, which is so important on the earth, is negligible for planetary motions, which is why Newtonian dynamics is more obviously valid for celestial bodies.

Note that the principle of relativity is closely related to Newtonian physics and is incompatible with pre-Newtonian views.

After all, two reference frames moving relative to each other cannot be equivalent in the pre-Newtonian view, because objects with nothing pushing on them can only come to rest in one of the two reference frames.

Newton's second law obeys the principle of relativity because the acceleration of an object is the same when viewed from two different reference frames moving at a constant velocity with respect to each other.

Einstein's relativity is often viewed as a repudiation of Newton, but this is far from the truth — Newtonian physics makes the theory of relativity possible through its invention of the principle of relativity.

Compared with the differences between pre-Newtonian and Newtonian dynamics, the changes needed to go from Newtonian to Einsteinian physics constitute minor tinkering.

8.1.3 Quantum Dynamics

In quantum mechanics, particles are represented by matter waves, with the absolute square of the wave displacement yielding the probability of finding the particle.

The behavior of particles thus follows from the reflection, refraction, diffraction, and interference of the associated waves.

The connection with Newtonian dynamics comes from tracing the trajectories of matter wave packets.

Changes in the speed and direction of motion of these packets correspond to the accelerations of classical mechanics.

When wavelengths are small compared to the natural length scale of the problem at hand, the wave packets can be made small, thus pinpointing the position of the associated particle, without generating excessive uncertainty in the particle's momentum.

This is the geometrical optics limit of quantum mechanics.

8.2 Potential Energy

We now address Newtonian mechanics in the case where the force on a particle is *conservative*.

A conservative force is one that can be derived from a so-called *potential energy U*.

We assume that the potential energy of the particle depends only on its position.

The force is obtained from the potential energy by the equation

$$F = -\frac{dU}{dx}. \tag{8.2}$$

Using this equation we write Newton's second law as

$$-\frac{dU}{dx} = ma. \tag{8.3}$$

We then notice that the acceleration can be written in terms of the x derivative along the particle's trajectory of $v^2/2$:

$$a = \frac{dv}{dt} = \frac{dv}{dx}\frac{dx}{dt} = \frac{dv}{dx}v = \frac{1}{2}\frac{dv^2}{dx}. \tag{8.4}$$

The last step in the above derivation can be verified by applying the product rule: $dv^2/dt = d(vv)/dt = v(dv/dt) + (dv/dt)v = 2v(dv/dt)$.

Putting equations (8.3) and (8.4) together, we find that $d(mv^2/2 + U)/dt = 0$, which implies that $mv^2/2 + U$ is constant.

We call this constant the *total energy* E and the quantity $K = mv^2/2$ the *kinetic energy*.

We thus have the principle of conservation of energy for conservative forces:

$$E = K + U = \text{constant}. \tag{8.5}$$

Recall that in quantum mechanics the momentum is related to the group velocity u_g by

$$\Pi = mu_g \quad (\text{momentum}) \tag{8.6}$$

in the nonrelativistic case.

Equating the group velocity to v and eliminating it in the kinetic energy results in an alternate expression for this quantity:

$$K \equiv \frac{1}{2}mu_g^2 = \frac{\Pi^2}{2m} \quad (\text{kinetic energy}). \tag{8.7}$$

Since the total energy E is constant or *conserved*, increases in the potential energy coincide with decreases in the kinetic energy and vice versa, as is illustrated in figure 8.1.

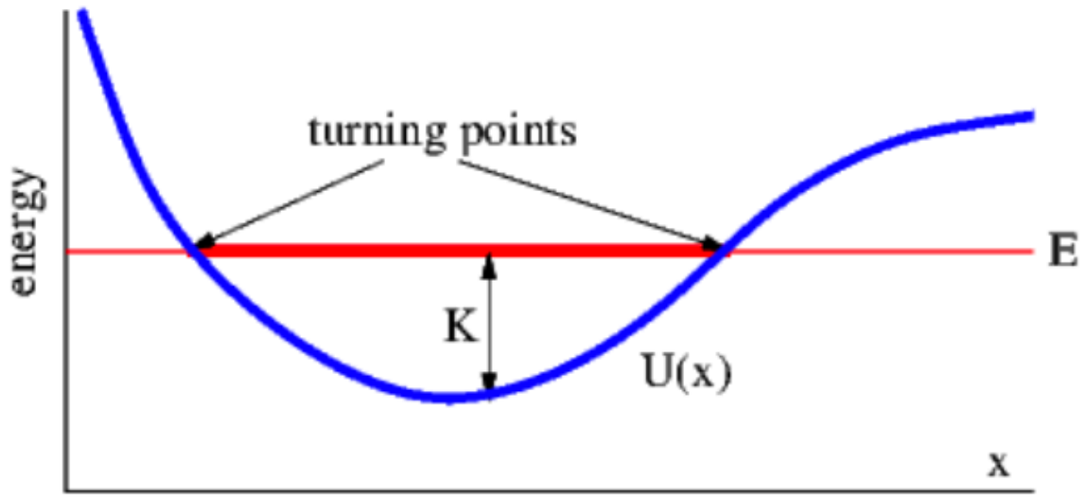


Figure 8.1: Example of spatially variable potential energy $U(x)$ for a particle with fixed total energy E. The kinetic energy $K = E - U$ is zero where the E and U lines cross. These points are called turning points. The thick part of the horizontal line indicates the allowed range for the particle with the given energy.

In classical mechanics the kinetic energy cannot be negative, since it is the product of half the mass and the square of the velocity, both of which are positive.

Thus, a particle with total energy E and potential energy U is forbidden to venture into regions in which the kinetic energy $K = E - U$ is less than zero.

The points at which the kinetic energy is zero are called *turning points*.

This is because a particle decreases in speed as it approaches a turning point, stops there for an instant, and reverses direction.

Note also that a particle with a given total energy always has the same speed at some point x regardless of whether it approaches this point from the left or the right:

$$\text{speed} = |u_g| = | \pm [2(E - U)/m]^{1/2}|. \tag{8.8}$$

8.2.1 Gravity as a Conservative Force

An example of a conservative force is gravity.

An object of mass *m* near the surface of the earth has the gravitational potential energy

$$U = mgz \quad \text{(gravity near earth's surface)} \tag{8.9}$$

where z is the height of the object above some reference point such as the earth's surface and $g = 9.8$ m s^{-2} is the local value of the *gravitational field* near the surface.

Notice that the gravitational potential energy increases upward.

The speed of the object in this case is $|u_g| = [2(E - mgz)/m]^{1/2}$.

If $|u_g|$ is known to equal the constant value u_0 at elevation $z = 0$, then equations (8.8) and (8.9) tell us that $u_0 = (2E/m)^{1/2}$ and $|u_g| = (u_0^2 - 2gz)^{1/2}$.

There are certain types of questions which energy conservation cannot directly answer.

For instance, if an object is released at elevation h with zero velocity at $t = 0$, at what time will it reach $z = 0$ under the influence of gravity?

In such cases it is often easiest to return to Newton's second law.

Since the force on the object is $F = -dU/dz = -mg$ in this case, we find that the acceleration is $a = F/m = -mg/m = -g$.

However, $a = du/dt = d^2z/dt^2$, so

$$u = -gt + C_1 \qquad z = -gt^2/2 + C_1 t + C_2 \quad \text{(constant gravity)}, \tag{8.10}$$

where C_1 and C_2 are constants to be determined by the initial conditions.

These results can be verified by differentiating to see if the original acceleration is recovered.

Since $u = 0$ and $z = h$ at $t = 0$, we have $C_1 = 0$ and $C_2 = h$.

With these results it is easy to show that the object reaches $z = 0$ when $t = (2h/g)^{1/2}$.

8.3 Work and Power

When a force is exerted on an object, energy is transferred to the object.

The amount of energy transferred is called the *work* done on the object.

However, energy is only transferred if the object moves.

The work W done is

$$W = F\Delta x \tag{8.11}$$

where the distance moved by the object is Δx and the force exerted on it is F.

Notice that work can either be positive or negative.

The work is positive if the object being acted upon moves in the same direction as the force, with negative work occurring if the object moves opposite to the force.

Equation (8.11) assumes that the force remains constant over the full displacement Δx.

If it is not, then it is necessary to break up the displacement into a number of smaller displacements, over each of which the force can be assumed to be constant.

The total work is then the sum of the works associated with each small displacement.

If more than one force acts on an object, the works due to the different forces each add or subtract energy, depending on whether they are positive or negative.

The total work is the sum of these individual works.

There are two special cases in which the work done on an object is related to other quantities.

If F is the total force acting on the object, then $W = F\Delta x = ma\Delta x$ by Newton's second law.

However, $a = dv/dt$ where v is the velocity of the object, and $\Delta x = (\Delta x/\Delta t)\Delta t \approx v\Delta t$, where Δt is the time required by the object to move through distance Δx.

The approximation becomes exact when Δx and Δt become very small.

Putting all of this together results in

$$W_{total} = m\frac{dv}{dt}v\Delta t = \frac{d}{dt}\left(\frac{mv^2}{2}\right)\Delta t = \Delta K \quad \text{(total work)}, \tag{8.12}$$

where K is the kinetic energy of the object.

Thus, when F is the only force, $W = W_{total}$ is the total work on the object, and this equals the change in kinetic energy of the object.

This is called the *work-energy theorem*, and it demonstrates that work really is a transfer of energy to an object.

The other special case occurs when the force is conservative, but is not necessarily the total force acting on the object.

In this case

$$W_{cons} = -\frac{dU}{dx}\Delta x = -\Delta U \quad \text{(conservative force)}, \tag{8.13}$$

where ΔU is the change in the potential energy of the object associated with the force of interest.

The *power* associated with a force is simply the amount of work done by the force divided by the time interval Δt over which it is done.

It is therefore the energy per unit time transferred to the object by the force of interest.

From equation (8.11) we see that the power is

$$P = \frac{F\Delta x}{\Delta t} = Fv \quad \text{(power)}, \tag{8.14}$$

where v is the velocity at which the object is moving.

The *total power* is just the sum of the powers associated with each force.

It equals the time rate of change of kinetic energy of the object:

$$P_{total} = \frac{W_{total}}{\Delta t} = \frac{dK}{dt} \quad \text{(total power)}. \tag{8.15}$$

8.4 Mechanics and Geometrical Optics

Louis de Broglie made an analogy between matter waves and light waves, pointing out that wave packets of light change their velocity as the result of spatial variations in the index of refraction of the medium in which they are traveling.

This behavior comes about because the dispersion relation for light traveling through a medium with index of refraction n is $\omega = kc/n$, so that the group velocity, $u_g = \mathrm{d}\omega/\mathrm{dk} = c/n$.

Thus, when n increases, u_g decreases, and vice versa.

In this section we pursue de Broglie's analogy to see if we can come up with a theory of matter waves which gives the same results as classical mechanics in the geometrical optics limit of these waves.

The dispersion relation for free matter waves is $\omega = (k^2c^2 + \mu^2)^{1/2}$.

In the non-relativistic limit $k^2c^2 \ll \mu^2$.

As done previously, we use $(1 + \epsilon)_n \approx 1 + n\epsilon$ for small ϵ.

In the non-relativistic limit, the dispersion relation for free waves thus becomes

$$\omega = \mu(1 + k^2c^2/\mu^2)^{1/2} \approx \mu + k^2c^2/(2\mu). \tag{8.16}$$

The above equation can be transformed into the total energy equation for a free, non-relativistic particle, $E = mc^2 + K$, where mc^2 is the rest energy and K is the kinetic energy, by multiplying by $\hbar$.

We convert the free particle energy equation into the equation for a particle subject to a conservative force by adding the potential energy U the right side.

The analogous change to equation (8.16) is to add $S = U/\hbar$ to the right side, resulting in a modified dispersion relation:

$$\omega = S(x) + k^2c^2/(2\mu). \tag{8.17}$$

Since the rest energy is just a constant, we have absorbed it into S.

This gives us the dispersion relation for one-dimensional matter waves subject to a spatially varying potential energy.

The quantity S, which we see is just a scaled potential energy, plays a role for matter waves which is analogous to the role played by a spatially variable index of refraction for light waves.

Let us now imagine that all parts of the wave governed by this dispersion relation oscillate in phase.

The only way this can happen is if ω is constant, i. e., it takes on the same value in all parts of the wave.

If ω is constant, the only way S can vary with x in equation (8.17) is if the wavenumber varies in a compensating way.

Thus, constant frequency and spatially varying S together imply that $k = k(x)$.

Solving equation (8.17) for k yields

$$k(x) = \pm \left[\frac{2\mu[\omega - S(x)]}{c^2} \right]^{1/2} . \tag{8.18}$$

Since ω is constant, the wavenumber becomes smaller and the wavelength larger as the wave moves into a region of increased S.

In the geometrical optics limit, we assume that S doesn't change much over one wavelength so that the wave remains reasonably sinusoidal in shape with approximately constant wavenumber over a few wavelengths.

However, over distances of many wavelengths the wavenumber and amplitude of the wave are allowed to vary considerably.

The group velocity calculated from the dispersion relation given by equation (8.17) is

$$u_g = \frac{d\omega}{dk} = \frac{kc^2}{\mu} = \pm\left(\frac{2c^2(\omega - S)}{\mu}\right)^{1/2} \tag{8.19}$$

where k is eliminated in the last step with the help of equation (8.18).

The resulting equation tells us how the group velocity varies as a matter wave traverses a region of slowly varying S.

Thus, as S increases, u_g decreases and vice versa.

We can now calculate the acceleration of a wave packet resulting from the spatial variation in S.

We assume that $x(t)$ represents the position of the wave packet, so that $u_g = dx/dt$.

Using the chain rule $du_g/dt = (du_g/dx)(dx/dt) = (du_g/dx)u_g$, we find

$$a = \frac{du_g}{dt} = \frac{du_g}{dx}u_g = \frac{du_g^2/2}{dx} = -\frac{c^2}{\mu}\frac{dS}{dx} = -\frac{\hbar}{m}\frac{dS}{dx}. \tag{8.20}$$

The group velocity is eliminated in favor of S by squaring equation (8.19) and substituting the result into equation (8.20).

Recalling that $U = \hbar S$, equation (8.20) becomes

$$a = -\frac{1}{m}\frac{dU}{dx} = \frac{F}{m}, \tag{8.21}$$

which is just Newton's second law!

Thus, the geometrical optics approach to particle motion is completely equivalent to the classical mechanics of a particle moving under the influence of a conservative force, at least in one dimension.

We therefore have two ways of solving for the motion of a particle subject to a potential energy $U(x)$.

We can apply the principles of classical mechanics to get the force and the acceleration of the particle, from which we can derive the motion.

Alternatively, we can apply the principles of geometrical optics to compute the spatially variable velocity of the wave packet using equation (8.19).

The results are completely equivalent, though the methods are conceptually very different.

8.5 Math Tutorial – Partial Derivatives

In order to understand the generalization of Newtonian mechanics to two and three dimensions, we first need to understand a new type of derivative called the *partial derivative*.

The partial derivative is used in functions of more than one variable.

It is just like an ordinary derivative, except that when taking the derivative of the function with respect to one of the variables, the other variables are held constant.

As an example, let us consider the function

$$f(x,y) = Ax^4 + Bx^2y^2 + Cy^4 \tag{8.22}$$

where A, B, and C are constants.

The partial derivative of f with respect to x is

$$\frac{\partial f}{\partial x} = 4Ax^3 + 2Bxy^2 \tag{8.23}$$

and the partial derivative with respect to y is

$$\frac{\partial f}{\partial y} = 2Bx^2y + 4Cy^3. \tag{8.24}$$

That's it!

Note that a special symbol "∂" is used in place of the normal "d" for the partial derivative.

This is sometimes called a "curly d".

8.6 Motion in Two and Three Dimensions

When a matter wave moves through a region of variable potential energy in one dimension, only the wavenumber changes.

In two or three dimensions, the wave vector can change in both direction and magnitude.

This complicates the calculation of particle movement.

However, we already have an example of how to handle this situation, namely, the refraction of light.

In that case Snell's law tells us how the direction of the wave vector changes, while the dispersion relation combined with the constancy of the frequency gives us information about the change in the magnitude of the wave vector.

For matter waves a similar procedure works, though the details are different, because we seek the consequences of a change in potential energy rather than a change in the index of refraction.

Figure 8.2 illustrates the refraction of matter waves at a discontinuity in the potential energy.

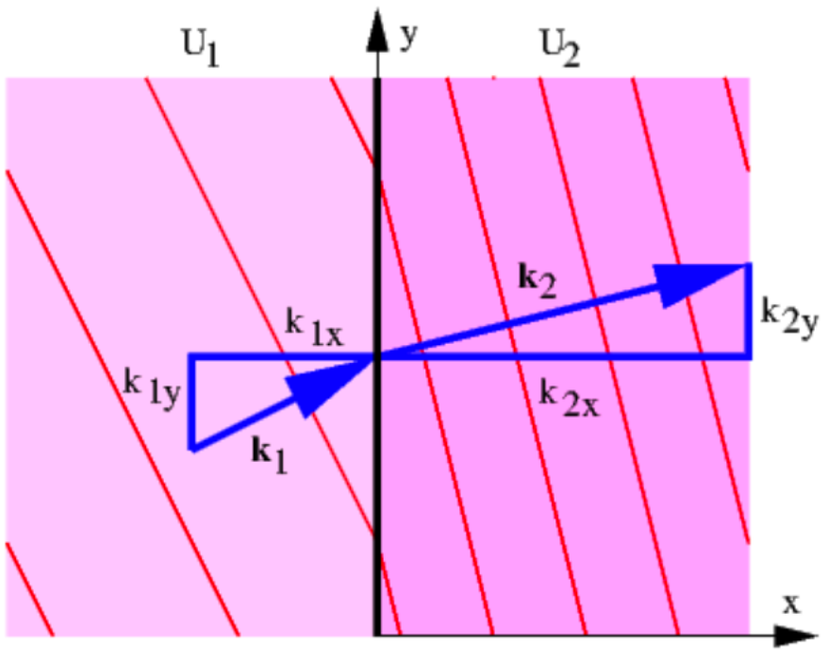


Figure 8.2: Refraction of a matter wave by a discontinuity in potential energy. The component of the wave vector parallel to the discontinuity, k_y, doesn't change, so $k_{1y} = k_{2y}$.

Let us suppose that the discontinuity occurs at $x = 0$.

If the matter wave to the left of the discontinuity is $\psi_1 = \sin(k_{1x}x + k_{1y}y - \omega_1 t)$ and to the right is $\psi_2 = \sin(k_{2x}x + k_{2y}y - \omega_2 t)$, then the wavefronts of the waves will match across the discontinuity for all time only if $\omega_1 = \omega_2 \equiv \omega$ and $k_{1y} = k_{2y} \equiv k_y$.

We are already familiar with the first condition from the one-dimensional problem, so the only new ingredient is the constancy of the y component of the wave vector.

In two dimensions the momentum is a vector: $\mathbf{\Pi} = m\mathbf{u}$ when $|\mathbf{u}| \ll c$, where $\mathbf{u}$ is the particle velocity.

Furthermore, the kinetic energy is $K = m|\mathbf{u}|^2/2 = |\mathbf{\Pi}|^2/(2m) = (\Pi_x^2 + \Pi_y^2)/(2m)$.

The relationship between kinetic, potential, and total energy is unchanged from the one-dimensional case, so we have

$$E = U + (\Pi_x^2 + \Pi_y^2)/(2m) = \text{constant}. \tag{8.25}$$

The de Broglie relationship tells us that $\mathbf{\Pi} = \hbar\mathbf{k}$, so the constancy of k_y across the discontinuity in U tells us that

$$\Pi_y = \text{constant} \tag{8.26}$$

there.

Let us now approximate a continuously variable $U(x)$ by a series of steps of constant U oriented normal to the x axis.

The above analysis can be applied at the jumps or discontinuities in U between steps, as illustrated in figure 8.3, with the result that equations (8.25) and (8.26) are valid across all discontinuities.

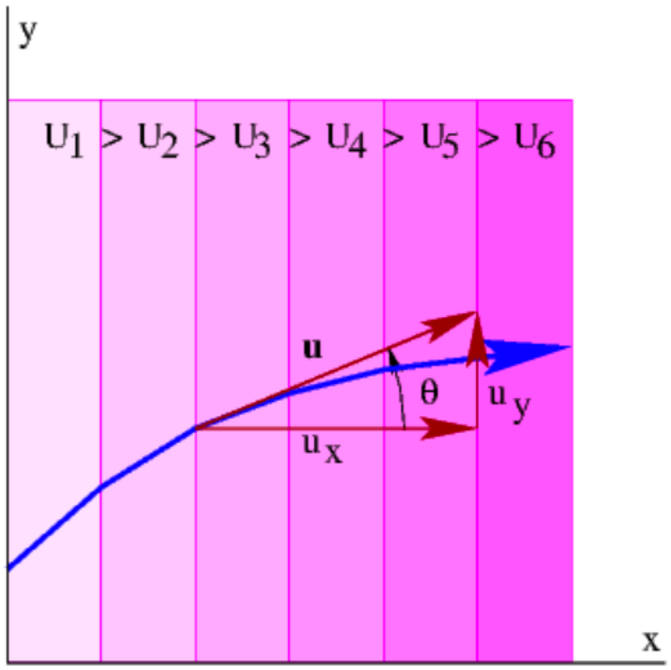


Figure 8.3: Trajectory of a wave packet through a variable potential energy, $U(x)$, which decreases to the right.

If we now let the step width go to zero, these equations then become valid for U continuously variable in x.

An example from classical mechanics of a problem of this type is a ball rolling down an inclined ramp with an initial velocity component across the ramp, as illustrated in figure 8.4.

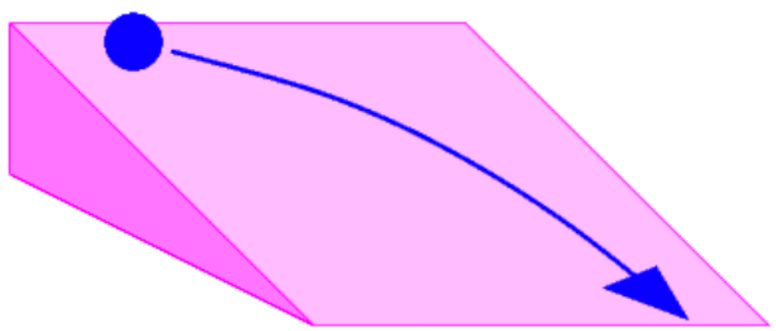


Figure 8.4: Classical mechanics example of the problem illustrated in figure 8.3.

The potential energy decreases in the down ramp direction, resulting in a force down the ramp.

This accelerates the ball in that direction, but leaves the component of momentum across the ramp unchanged.

Using the procedure which we invoked before, we find the force components associated with $U(x)$ in the x and y directions to be $F_x = -dU/dx$ and $F_y = 0$.

This generalizes to

$$\mathbf{F} = -\left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z}\right) \quad \text{(3-D conservative force)} \tag{8.27}$$

in the three-dimensional case where the orientation of constant U surfaces is arbitrary.

It is also valid when $U(x,y,z)$ is not limited to a simple ramp form, but takes on a completely arbitrary structure.

The definitions of work and power are slightly different in two and three dimensions.

In particular, work is defined as

$$W = \mathbf{F} \cdot \Delta\mathbf{x} \tag{8.28}$$

where $\Delta\mathbf{x}$ is now a vector displacement.

The vector character of this expression yields an additional possibility over the one dimensional case, where the work is either positive or negative depending on the direction of $\Delta\mathbf{x}$ relative to $\mathbf{F}$.

If the force and the displacement of the object on which the force is acting are perpendicular to each other, the work done by the force is actually zero, even though the force and the displacement both have non-zero magnitudes.

The power exhibits a similar change:

$$P = \mathbf{F} \cdot \mathbf{u}. \tag{8.29}$$

Thus the power is zero if an object's velocity is normal to the force being exerted on it.

As in the one-dimensional case, the total work done on a particle equals the change in the particle's kinetic energy.

In addition, the work done by a conservative force equals minus the change in the associated potential energy.

Energy conservation by itself is somewhat less useful for solving problems in two and three dimensions than it is in one dimension.

This is because knowing the kinetic energy at some point tells us only the magnitude of the velocity, not its direction.

If conservation of energy fails to give us the information we need, then we must revert to Newton's second law, as we did in the one-dimensional case.

For instance, if an object of mass m has initial velocity $\mathbf{u}_0 = (u_0, 0)$ at location $(x, z) = (0, h)$ and has the gravitational potential energy $U = mgz$, then the force on the object is $\mathbf{F} = (0,-mg)$.

The acceleration is therefore $\mathbf{a} = \mathbf{F}/m = (0,-g)$. Since $\mathbf{a} = d\mathbf{u}/dt = d^2\mathbf{x}/dt^2$ where $\mathbf{x} = (x, z)$ is the object's position, we find that

$$\mathbf{u} = (C_1, -gt + C_2) \qquad \mathbf{x} = (C_1 t + C_3, -gt^2/2 + C_2 t + C_4), \tag{8.30}$$

where C_1, C_2, C_3, and C_4 are constants to be evaluated so that the solution reduces to the initial conditions at $t = 0$.

The specified initial conditions tell us that $C_1 = u_0$, $C_2 = 0$, $C_3 = 0$, and $C_4 = h$ in this case.

From these results we can infer the position and velocity of the object at any time.

8.7 Kinetic and Total Momentum

If you have previously taken a physics course then you have probably noticed that a rather odd symbol is used for momentum, namely $\mathbf{\Pi}$, rather than the more commonly employed $\mathbf{p}$.

The reason for this peculiar usage is that there are actually two kinds of momentum, *kinetic momentum* and *total momentum*, just as there are two kinds of energy, kinetic and total.

The symbol $\mathbf{\Pi}$ represents total momentum while $\mathbf{p}$ represents kinetic momentum.

Normally we don't need to distinguish between the two quantities, as they are generally equal to each other.

However, we will find later in the course that it is crucial to make this distinction in the case of charged particles in a magnetic field.

As a general rule, the total momentum is related to a particle's wave vector via the de Broglie relation, $\mathbf{\Pi} = \hbar\mathbf{k}$, while the kinetic momentum is related to a particle's velocity, $\mathbf{p} = m\mathbf{u}/(1 - u^2/c^2)^{1/2}$.

Chapter 9

Symmetry and Bound States

When quantum mechanics was first invented, the dynamical principles used were the same as those underlying classical mechanics.

The initial development of the field thus proceeded largely by imposing quantum laws on classical variables such as position, momentum, and energy.

However, as quantum mechanics advanced, it became clear that there were many situations in which no classical analogs existed for new types of quantum mechanical systems, especially those which arose in the study of elementary particles.

To understand these systems it was necessary to seek guidance from novel sources.

One of the most important of these sources was the idea of symmetry, and in particular the relationship between symmetry and conserved variables.

This type of relationship was first developed in the early 20th century by the German mathematician Emmy Nöther in the context of classical mechanics.

However, her idea is easier to express and use in quantum mechanics than it is in classical mechanics.

Emmy Nöther showed that there is a relationship between the symmetries of a system and conserved dynamical variables.

This idea is naturally called *Nöther's theorem*.

In classical mechanics a particle is bound or confined to a particular region if its total energy exceeds the potential energy only in this region.

In quantum mechanics, matter waves can also be confined for the same reason, though the confinement is often less perfect than in the classical case.

In this chapter we examine the consequences of wave confinement.

We first look at the so-called “particle in a box” in one spatial dimension.

We find that confined particles can take on only discrete energy values.

When confinement isn’t perfect we see how a quantum mechanical particle can leak through a potential energy barrier which is classically impenetrable.

Movement of a particle on a circular ring leads us to another form of confinement and the introduction of angular momentum.

This brings us finally to a discussion of the intrinsic or spin angular momentum of elementary particles.

9.1 Math Tutorial — Complex Waves

Until now we have represented quantum mechanical plane waves by sine and cosine functions, just as with other types of waves.

However, plane matter waves cannot be truly represented by sines and cosines.

We need instead mathematical functions in which the wave displacement is complex rather than real.

This requires the introduction of a bit of new mathematics, which we tackle first.

Using our new mathematical tool, we are then able to explore two crucially important ideas in quantum mechanics; (1) the relationship between symmetry and conservation laws, and (2) the dynamics of spatially confined waves.

A *complex number* z is the sum of a real number and an imaginary number.

An imaginary number is just a real number multiplied by $i \equiv (-1)^{1/2}$.

Thus, we can write $z = a + ib$ for any complex z, where a and b are real.

The quantities a and b are the *real* and *imaginary* parts of z, sometimes written Re(z) and Im(z).

Quantum mechanics requires wave functions to be complex, i. e., to possess real and imaginary parts.

Plane waves in quantum mechanics actually take the form $\psi = \exp[i(kx - \omega t)]$ rather than, say, $\cos(kx - \omega t)$.

The reason for this is the need to distinguish between waves with positive and negative frequencies.

If we replace k and ω with $-k$ and $-\omega$ in the cosine form, we get $\cos(-kx + \omega t) = \cos[-(kx - \omega t)] = \cos(kx - \omega t)$.

In other words, changing the sign of k and ω results in no change in a wave expressed as a cosine function.

The two quantum mechanical states, one with wavenumber and frequency k and ω and the other with $-k$ and $-\omega$, yield indistinguishable wave functions and therefore would represent physically indistinguishable states.

The cosine form is thus insufficiently flexible to represent quantum mechanical waves.

On the other hand, if we replace k and ω with their negatives in the complex exponential form of a plane wave we get $\psi = \exp[-i(kx - \omega t)]$, which is different from $\exp[i(kx - \omega t)]$.

These two wave functions are distinguishable and thus correspond to distinct physical states.

It is not immediately obvious that a complex exponential function provides the oscillatory behavior needed to represent a plane wave.

However, the complex exponential can be expressed in terms of sines and cosines using Euler's equation:

$$\exp(i\phi) = \cos(\phi) + i\sin(\phi) \quad \text{(Euler's equation)}. \tag{9.1}$$

If we define $r = (a_2 + b_2)^{1/2}$ and $\varphi = \tan^{-1}(b/a)$, then an alternate way of expressing a complex number is $z = r\exp(i\varphi)$, which by Euler's equation equals $r\cos(\varphi) + ir\sin(\varphi)$.

Comparison shows that $a = r\cos(\varphi)$ and $b = r\sin(\varphi)$.

Thus, a complex number can be thought of as a point in the a-b plane with Cartesian coordinates a and b and polar coordinates r and φ.

The a-b plane is called the *complex plane*.

We now see how the complex wave function represents an oscillation.

If $\psi = \exp[i(kx-\omega t)]$, the complex function $\psi(x,t)$ moves round and round the unit circle in the complex plane as x and t change, as illustrated in figure 9.1.

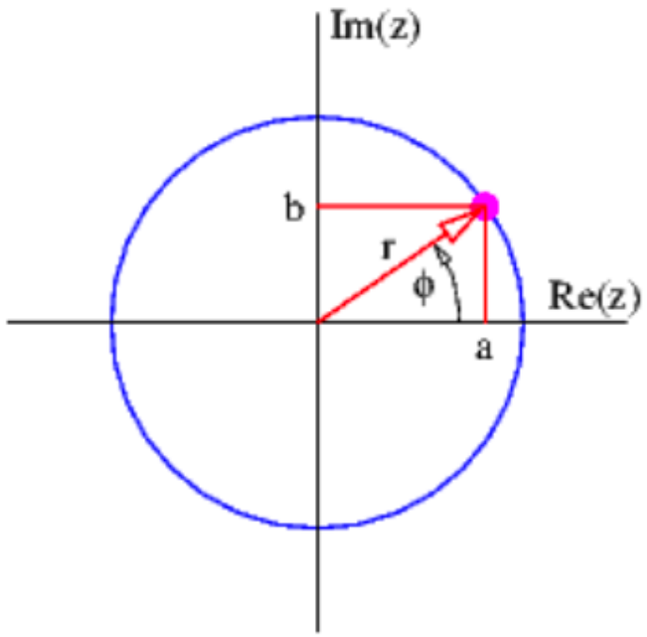


Figure 9.1: Graphical representation of a complex number z as a point in the complex plane. The horizontal and vertical Cartesian components give the real and imaginary parts of z respectively.

This contrasts with the back and forth oscillation along the horizontal axis of the complex plane represented by $\cos(kx - \omega t)$.

We will not present a formal proof of Euler's equation — you will eventually see it in a calculus course.

However, it may be helpful to note that the φ derivatives of $\exp(i\varphi)$ and $\cos(\varphi) + i\sin(\varphi)$ have the same behavior:

$$\frac{d}{d\phi}\exp(i\phi) = i\exp(i\phi); \tag{9.2}$$

$$\begin{aligned} \frac{d}{d\phi}[\cos(\phi) + i\sin(\phi)] &= -\sin(\phi) + i\cos(\phi) \\ &= i[\cos(\phi) + i\sin(\phi)]. \end{aligned} \tag{9.3}$$

In the second of these equations we have replaced the minus sign in front of the sine function by i^2 and then extracted a common factor of i.

The φ derivative of both of these functions thus yields the function back again times i.

This is a strong hint that $\exp(i\varphi)$ and $\cos(\varphi) + i \sin(\varphi)$ are different ways of representing the same function.

We indicate the *complex conjugate* of a complex number z by a superscripted asterisk, i. e., z^*. It is obtained by replacing i by $-i$. Thus, $(a + ib)^* = a - ib$.

The *absolute square* of a complex number is the number times its complex conjugate:

$$|z|^2 = |a + ib|^2 \equiv (a + ib)(a - ib) = a^2 + b^2 = r^2. \tag{9.4}$$

Notice that the absolute square of a complex exponential function is one:

$$|\exp(i\phi)|^2 = \exp(i\phi)\exp(-i\phi) = \exp(i\phi - i\phi) = \exp(0) = 1. \tag{9.5}$$

In quantum mechanics the absolute square of the wave function at any point expresses the relative probability of finding the associated particle at that point.

Thus, the probability of finding a particle represented by a plane wave is uniform in space.

Contrast this with the relative probability associated with a sine wave: $|\sin(kx - \omega t)|^2 = \sin^2(kx - \omega t)$.

This varies from zero to one, depending on the phase of the wave.

The "waviness" in a complex exponential plane wave resides in the phase rather than in the magnitude of the wave function.

One more piece of mathematics is needed.

The complex conjugate of Euler's equation is

$$\exp(-i\phi) = \cos(\phi) - i\sin(\phi). \tag{9.6}$$

Taking the sum and the difference of this with the original Euler's equation results in the expression of the sine and cosine in terms of complex exponentials:

$$\cos(\phi) = \frac{\exp(i\phi) + \exp(-i\phi)}{2} \qquad \sin(\phi) = \frac{\exp(i\phi) - \exp(-i\phi)}{2i}. \tag{9.7}$$

We aren't used to having complex numbers show up in physical theories and it is hard to imagine how we would measure such a number.

However, everything observable comes from taking the absolute square of a wave function, so we deal only with real numbers in experiments.

9.2 Symmetry and Quantum Mechanics

The idea of symmetry plays a huge role in physics.

We have already used symmetry arguments in the theory of relativity — applying the principle of relativity to obtain the dispersion relation for relativistic matter waves is just such an argument.

In this section we begin to explore how symmetry can be used to increase our understanding of quantum mechanics.

9.2.1 Free Particle

For our first example we take the case of a free particle in quantum mechanics, i. e., a particle subject to no force.

The wave function for a free particle of definite momentum Π and energy E is given by

$$\psi = \exp[i(kx - \omega t)] = \exp[i(\Pi x - Et)/\hbar] \quad \text{(free particle)}. \tag{9.8}$$

For this wave function $|\psi|^2 = 1$ everywhere, so the probability of finding the particle anywhere in space and time is uniform.

This contrasts with the probability distribution which arises if we assume a free particle to have the wave function $\psi = \cos[(\Pi x - Et)/\hbar]$.

In this case $|\psi|^2 = \cos^2[(\Pi x - Et)/\hbar]$, which varies with position and time, and is inconsistent with a uniform probability distribution.

9.2.2 Symmetry and Definiteness

Quantum mechanics is a probabilistic theory, in the sense that the predictions it makes tell us, for instance, the probability of finding a particle somewhere in space.

If we know nothing about a particle's previous history, and if there are no physical constraints that would make it more likely for a particle to be at one point along the x axis than any another, then the probability distribution must be $P(x) = \text{constant}$.

This is an example of a *symmetry* argument.

Expressed more formally, it states that if the above conditions apply, then the probability distribution ought to be subject to the condition $P(x + D) = P(x)$ for any constant value of D.

The only possible $P(x)$ in this case is $P = constant$.

In the language of physics, if there is nothing that gives the particle a higher probability of being at one point rather than another, then the probability is independent of position and the system is *invariant under displacement* in the x direction.

The above argument doesn't suffice for quantum mechanics, since as we have learned, the fundamental quantity describing a particle is not the probability distribution, but the wave function $\psi(x)$.

Thus, the wave function rather than the probability distribution ought to be the quantity which is invariant under displacement, i. e., $\psi(x + D) = \psi(x)$.

This condition turns out to be too restrictive, because it implies that $\psi(x) = constant$, whereas we know that a one-dimensional plane wave, which describes a particle with a uniform probability of being found anywhere along the x axis, has the form $\psi(x) = \exp(ikx)$.

For simplicity we temporarily ignore the time dependence.

If we make the substitution $x \rightarrow x + D$ in a plane wave, we get $\exp[ik(x + D)] = \exp(ikx) \exp(ikD)$.

The wave function is thus technically not invariant under displacement, in that the displaced wave function is multiplied by the factor $\exp(ikD)$.

However, the probability distribution of the displaced wave function still equals one everywhere, so there is no change in what we observe.

Thus, in determining invariance under displacement, we are allowed to ignore changes in the wave function which consist only of multiplying it by a complex constant with an absolute value of one.

Such a multiplicative constant is called a *phase factor*.

It is easy to convince oneself by trial and error or by more sophisticated means that the only form of wave function $\psi(x)$ which satisfies the condition $\psi(x + D) = \psi(x) \times$ (phase factor) is $\psi(x) = A \exp(ikx)$ where A is a (possibly complex) constant.

This is just in the form of a complex exponential plane wave with wavenumber k.

Thus, not only is the complex exponential wave function invariant under displacements in the manner defined above, it is the *only* wave function which is invariant to displacements.

Furthermore, the phase factor which appears for a displacement D of such a plane wave takes the form $\exp(iC) = \exp(ikD)$, where k is the wavenumber of the plane wave.

As an experiment, let us see if a wave packet is invariant under displacement.

Let's define a wave packet consisting of two plane waves:

$$\psi(x) = \exp(ik_1 x) + \exp(ik_2 x). \tag{9.9}$$

Making the substitution $x \rightarrow x + D$ in this case results in

$$\begin{aligned} \psi(x+D) &= \exp[ik_1(x+D)] + \exp[ik_2(x+D)] \\ &= \exp(ik_1 x)\exp(ik_1 D) + \exp(ik_2 x)\exp(ik_2 D) \\ &\neq [\exp(ik_1 x) + \exp(ik_2 x)] \times (\text{phase factor}). \end{aligned} \tag{9.10}$$

The impossibility of writing $\psi(x + D) = \psi(x) \times$ (phase factor) lends plausibility to the assertion that a single complex exponential is the only possible form of the wave function that is invariant under displacement.

Notice that the wave packet does not have definite wavenumber, and hence, momentum.

In particular, the wave packet is a sum of complex exponentials with wavenumbers k_1 and k_2, which means that the associated particle can have either momentum $\Pi_1 = \hbar k_1$ or $\Pi_2 = \hbar k_2$.

This makes sense from the point of view of the uncertainty principle – for a single plane wave the uncertainty in position is complete and the uncertainty in momentum is zero.

For a wave packet the uncertainty in position is reduced and the uncertainty in the momentum is non-zero.

As we have seen, this idea can be carried further: A definite value of momentum must be associated with a completely indefinite probability distribution in position, i. e., with $P = constant$.

This corresponds to a wave function which has the form of a complex exponential plane wave.

However, such a plane wave is invariant under displacement D, except for the multiplicative phase factor $\exp(ikD)$, which has no physical consequences since it disappears when the probability distribution is obtained.

Thus, we see that invariance under displacement of the wave function and a definite value of the momentum are linked, in that each implies the other:

$$\text{invariance under displacement} \Longleftrightarrow \text{definite momentum} \tag{9.11}$$

The idea of potential energy was introduced in the previous chapter.

In particular, we found that if the total energy is constant, then the momentum cannot be constant in the presence of spatially varying potential energy.

This means that the wavenumber, and hence the wavelength of the oscillations in the wave function also vary with position.

The spatial inhomogeneity of the potential energy gives rise to spatial inhomogeneity in the wave function, and hence an indefinite momentum.

The above argument can be extended to other variables besides momentum.

In particular since the time dependence of a complex exponential plane wave is $\exp(-i\omega t) = \exp(-iEt/\hbar)$, where E is the total energy, we have by analogy with the above argument that

$$\text{invariance under time shift} \iff \text{definite energy}. \tag{9.12}$$

Thus, invariance of the wave function under a displacement in time implies a definite value of the energy of the associated particle.

In the previous chapter we assumed that the frequency (and hence the energy) was definite and constant for a particle passing through a region of variable potential energy.

We now see that this assumption is justified only if the potential energy doesn't change with time.

This is because a time-varying potential energy eliminates the possibility of invariance under time shift.

9.2.3 Invariance

We have seen a few examples of invariance in quantum mechanics.

It is now time to define this concept more precisely.

A quantum mechanical wave function is said to be invariant under some transformation if the transformed wave function is observationally indistinguishable from the original.

In the above examples, the transformation is accomplished by replacing x by $x + D$ in the case of displacement in space and similarly by replacing t by $t + T$ for displacement in time.

However, the idea of a transformation is much more general; other examples will be discussed as they arise.

The idea of "observationally indistinguishable" can be tricky.

For example, if some transformation results in a new wave function which is the old wave function times a constant phase factor, then the new wave function is observationally indistinguishable from the old one.

This is because physical measurements capture phase differences between different parts of wave functions (think of how interferometers work), but not absolute phases.

The constant phase factor disappears in this difference calculation.

However, if the multiplicative phase factor created by some transformation is a function of position, then the phase difference between different parts of a wave function changes as a result of the transformation.

The wave function is not invariant under this transformation.

9.2.4 Compatible Variables

We already know that definite values of certain pairs of variables cannot be obtained simultaneously in quantum mechanics.

For instance, the indefiniteness of position and momentum are related by the uncertainty principle — a definite value of position implies an indefinite value of the momentum and vice versa.

If definite values of two variables can be simultaneously obtained, then we call these variables *compatible*.

If not, the variables are *incompatible*.

If the wave function of a particle is invariant under the displacements associated with both variables, then the variables are compatible.

For instance, the complex exponential plane wave associated with a free particle is invariant under displacements in both space and time.

Since momentum is associated with space displacements and energy with time displacements, the momentum and energy are compatible variables for a free particle.

9.2.5 Compatibility and Conservation

Variables which are compatible with the energy have a special status.

The wave function which corresponds to a definite value of such a variable is invariant to displacements in time.

In other words, the wave function doesn't change under this displacement except for a trivial phase factor.

Thus, if the wave function is also invariant to some other transformation at a particular time, it is invariant to that transformation for all time.

The variable associated with that transformation therefore retains its definite value for all time — i. e., it is *conserved*.

For example, the plane wave implies a definite value of energy, and is thus invariant under time displacements.

At time $t = 0$, it is also invariant under x displacements, which corresponds to the fact that it represents a particle with a known value of momentum.

However, since momentum and energy are compatible for a free particle, the wave function will represent the *same* value of momentum at all other times.

In other words, if the momentum is definite at $t = 0$, it will be definite at all later times, and furthermore will have the same value.

This is how the conservation of momentum (and by extension, the conservation of any other variable compatible with energy) is expressed in quantum mechanics.

9.2.6 New Symmetries and Variables

In modern quantum physics, the discovery of new symmetries leads to new dynamical variables.

In the problems we show how that comes about for the symmetries of *parity* ($x \rightarrow -x$), *time reversal* $t \rightarrow -t$), and *charge conjugation* (the interchange of particles with antiparticles).

One of the key examples of this was the development of the quark theory of matter, which came from the observation that the interchange of certain groups of elementary particles left the universe approximately unchanged, meaning that the universe was (approximately) symmetric under these interchanges.

9.3 Confined Matter Waves

Confinement of a wave to a limited spatial region results in rather peculiar behavior — the wave can only fit comfortably into the confined region if the wave frequency, and hence the associated particle energy, takes on a limited set of possible values.

This is the origin of the famous quantization of energy, from which the "quantum" in quantum mechanics comes.

We will explore two types of confinement, position confinement due to a potential energy well, and rotational confinement due to the fact that rotation of an object through 2π radians returns the object to its original orientation.

9.3.1 Particle in a Box

We now imagine how a particle confined to a region $0 \leq x \leq a$ on the x axis must behave.

As with the displacement of a guitar string, the wave function must be zero at $x = 0$ and a, i. e., at the ends of the guitar string.

A single complex exponential plane wave cannot satisfy this condition, since $|\exp[i(kx - \omega t)]|^2 = 1$ everywhere.

However, a superposition (with a minus sign) of leftward and rightward traveling waves creates a *standing wave*, in which the the wave function separates into a function of space alone times a function of time alone.

$$\psi = \exp[i(kx - \omega t)] - \exp[i(-kx - \omega t)] = 2i \exp(-i\omega t) \sin(kx). \tag{9.13}$$

Notice that the time dependence is still a complex exponential, which means that $|\psi|^2$ is independent of time.

This insures that the probability of finding the particle *somewhere* in the box remains constant with time.

It also means that the wave packet corresponds to a definite energy $E = \hbar\omega$.

Because we took a difference rather than a sum of plane waves, the condition $\psi = 0$ is already satisfied at $x = 0$.

To satisfy it at $x = a$, we must have $ka = n\pi$, where $n = 1, 2, 3, \ldots$.

Thus, the absolute value of the wavenumber must take on the discrete values

$$k_n = \frac{n\pi}{a}, \qquad n = 1, 2, 3, \ldots. \tag{9.14}$$

The wavenumbers of the two plane waves equal plus or minus this absolute value respectively.

This implies that the absolute value of the particle momentum is $\Pi_n = \hbar k_n = n\pi\hbar/a$, which in turn means that the energy of the particle must be

$$E_n = (\Pi_n^2 c^2 + m^2 c^4)^{1/2} = (n^2\pi^2\hbar^2 c^2/a^2 + m^2 c^4)^{1/2}, \tag{9.15}$$

where m is the particle mass.

In the non-relativistic limit this becomes

$$E_n = \frac{\Pi_n^2}{2m} = \frac{n^2\pi^2\hbar^2}{2ma^2} \quad \text{(non-relativistic)} \tag{9.16}$$

where we have dropped the rest energy mc^2 since it is a constant offset.

In the ultra-relativistic case where we can ignore the particle mass, we find

$$E_n = |\Pi_n|c = \frac{n\pi\hbar c}{a} \quad \text{(zero mass)}. \tag{9.17}$$

The shapes of the wave functions for the first three values of n for the particle in the box are illustrated in figure 9.2.

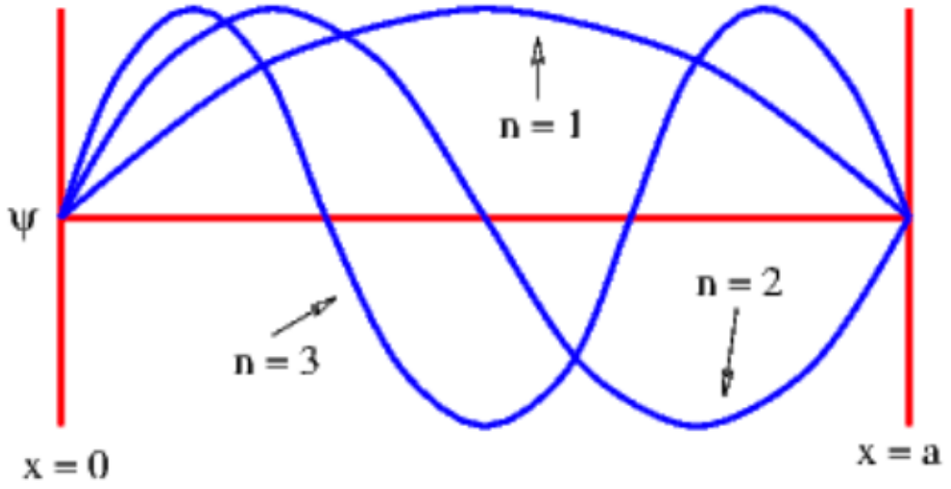


Figure 9.2: First three modes for wave function of a particle in a box.

In both limits the energy takes on only a certain set of possible values.

This is called *energy quantization* and the integer n is called the *energy quantum number*.

In the non-relativistic limit the energy is proportional to n^2 , while in the ultra-relativistic case the energy is proportional to n.

We can graphically represent the allowed energy levels for the particle in a box by an *energy level diagram*.

Such a diagram is shown in figure 9.3 for the non-relativistic case.

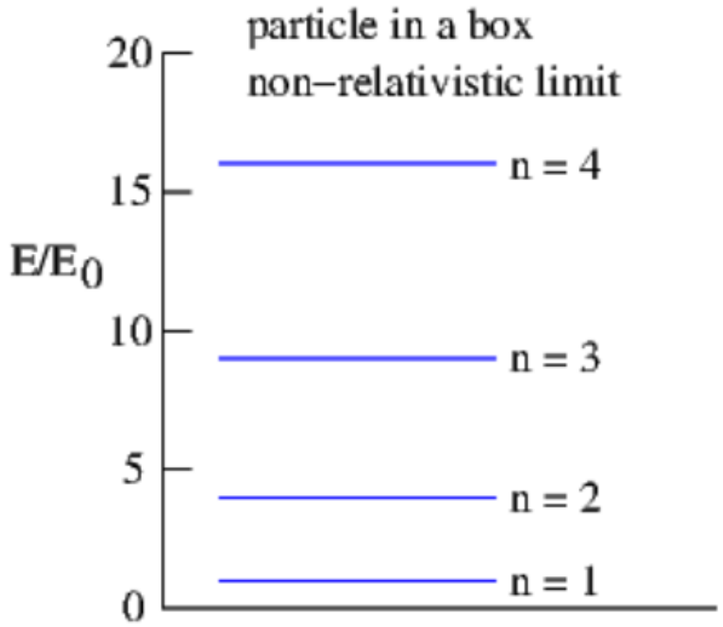


Figure 9.3: Allowed energy levels for the non-relativistic particle in a box. The constant $E_0 = \pi^2\hbar^2 / (2ma^2)$. See text for the meanings of symbols.

One aspect of this problem deserves a closer look.

Equation (9.13) shows that the wave function for this problem is a superposition of two plane waves corresponding to momenta $\Pi_1 = +\hbar k$ and $\Pi_2 = -\hbar k$ and is therefore a kind of wave packet.

Thus, the wave function is not invariant under displacement and does not correspond to a definite value of the momentum — the momentum's absolute value is definite, but its sign is not.

Following Feynman's prescription, equation (9.13) tells us that the amplitude for the particle in the box to have momentum $+\hbar k$ is $\exp[i(kx - \omega t)]$, while the amplitude for it to have momentum $-\hbar k$ is

$- \exp[i(-kx - \omega t)]$.

The absolute square of the sum of these amplitudes gives us the relative probability of finding the particle at position x:

$$P(x) = |2i \exp(-i\omega t) \sin(kx)|^2 = 4 \sin^2(kx). \tag{9.18}$$

Which of the two possible values of the momentum the particle takes on is unknowable, just as it is impossible in principle to know which slit a particle passes through in two slit interference.

If an experiment is done to measure the momentum, then the wave function is irreversibly changed, just as the interference pattern in the two slit problem is destroyed if the slit through which the particle passes is unambiguously determined.

9.3.2 Barrier Penetration

Unlike the situation in classical mechanics, quantum mechanics allows the kinetic energy K to be negative.

This makes the momentum Π (equal to $(2mK)^{1/2}$ in the nonrelativistic case) imaginary, which in turn gives rise to an imaginary wavenumber.

Let us investigate the nature of a wave with an imaginary wavenumber.

Let us assume that $k = i\varkappa$ in a complex exponential plane wave, where $\varkappa$ is real:

$$\psi = \exp[i(kx - \omega t)] = \exp(-\kappa x - i\omega t) = \exp(-\kappa x)\exp(-i\omega t). \qquad (9.19)$$

The wave function doesn't oscillate in space when $K = E - U < 0$, but grows or decays exponentially with x, depending on the sign of $\varkappa$.

For a particle moving to the right, with positive k in the allowed region, $\varkappa$ turns out to be positive, and the solution decays to the right.

Thus, a particle impingent on a potential energy barrier from the left (i. e., while moving to the right) will have its wave amplitude decay in the classically forbidden region, as illustrated in figure 9.4.

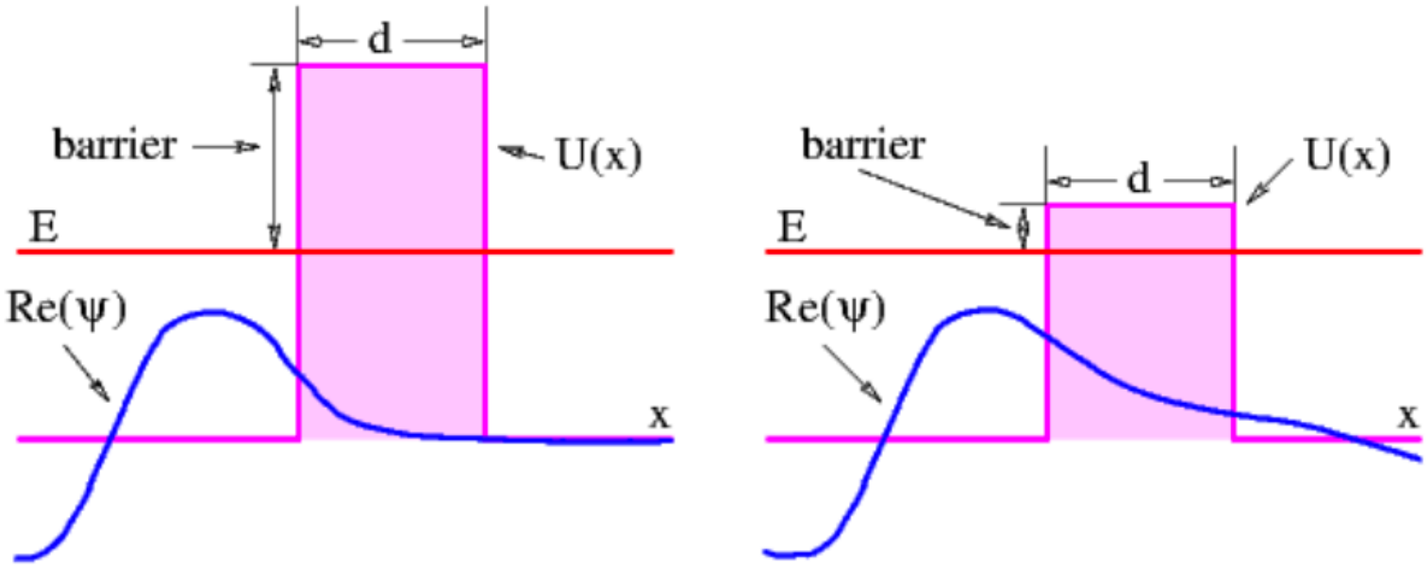


Figure 9.4: Real part of wave function $\mathrm{Re}[\psi(x)]$ for barrier penetration. The left panel shows weak penetration occurring for a large potential energy barrier, while the right panel shows stronger penetration which occurs when the barrier is small.

If this decay is very rapid, then the result is almost indistinguishable from the classical result — the particle cannot penetrate into the forbidden region to any great extent.

However, if the decay is slow, then there is a reasonable chance of finding the particle in the forbidden region.

If the forbidden region is finite in extent, then the wave amplitude will be small, but non-zero at its right boundary, implying that the particle has a finite chance of completely passing through the classical forbidden region.

This process is called *barrier penetration*.

The probability for a particle to penetrate a barrier is the absolute square of the amplitude after the barrier divided by the square of the amplitude before the barrier.

Thus, in the case of the wave function illustrated in equation (9.19), the probability of penetration is

$$P = |\psi(d)|^2 / |\psi(0)|^2 = \exp(-2\kappa d) \tag{9.20}$$

where d is the thickness of the barrier.

The rate of exponential decay with x in the forbidden region is related to how negative K is in this region.

Since

$$-K = U - E = -\frac{\Pi^2}{2m} = -\frac{\hbar^2 k^2}{2m} = \frac{\hbar^2 \kappa^2}{2m}, \tag{9.21}$$

we find that

$$\kappa = \left(\frac{2mB}{\hbar^2}\right)^{1/2} \tag{9.22}$$

where the potential energy barrier is B ≡ -K = U - E.

The smaller B is, the smaller is $\varkappa$, resulting in less rapid decay of the wave function with x.

This corresponds to stronger barrier penetration.

Note that the way B is defined, it is positive in forbidden regions.

If the energy barrier is very high, then the exponential decay of the wave function is very rapid.

In this case the wave function goes nearly to zero at the boundary between the allowed and forbidden regions.

This is why we specify the wave function to be zero at the walls for the particle in a box.

These walls act in effect as infinitely high potential barriers.

Barrier penetration is important in a number of natural phenomena.

Certain types of radioactive decay and the fissioning of heavy nuclei are governed by this process.

9.3.3 Orbital Angular Momentum

Another type of bound state motion occurs when a particle is constrained to move in a circle.

Imagine a bead sliding on a circular loop of wire, as illustrated in figure 9.5.

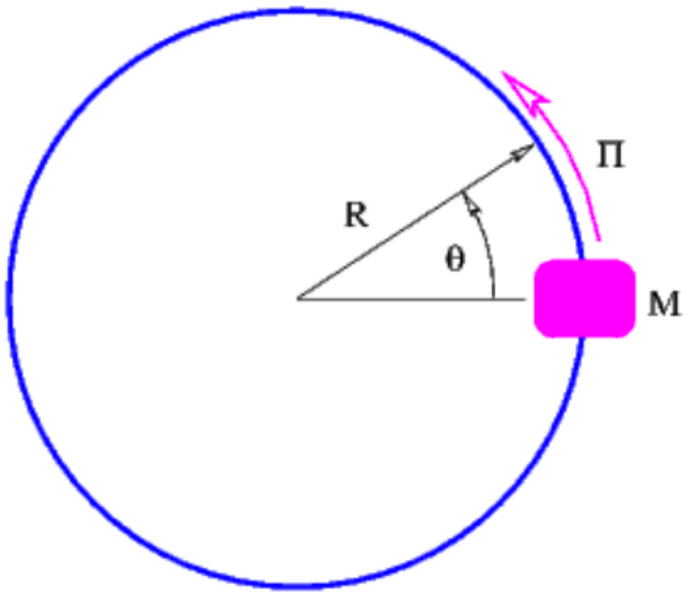


Figure 9.5: Illustration of a bead of mass M sliding (without friction) on a circular loop of wire of radius R with momentum Π.

We can define x in this case as the path length around the wire and relate it to the angle θ: $x = R\theta$.

For a plane wave we have

$$\psi = \exp[i(kx - \omega t)] = \exp[i(kR\theta - \omega t)]. \tag{9.23}$$

This plane wave differs from the normal plane wave for motion along a Cartesian axis in that we must have $\psi(\theta) = \psi(\theta + 2\pi)$.

This can only happen if the circumference of the loop, $2\pi R$, is an integral number of wavelengths, i. e., if $2\pi R/\lambda = m$ where m is an integer.

However, since $2\pi/\lambda = k$, this condition becomes $kR = m$.

Since $\Pi = \hbar k$, the above condition can be written $\Pi_m R = m\hbar$.

The quantity

$$L_m \equiv \Pi_m R \tag{9.24}$$

is called the *angular momentum*, leading to our final result,

$$L_m = m\hbar, \qquad m = 0, \pm 1, \pm 2, \ldots. \tag{9.25}$$

We see that the angular momentum can only take on values which are integer multiples of $\hbar$.

This represents the quantization of angular momentum, and m in this case is called the angular momentum *quantum number*.

Note that this quantum number differs from the energy quantum number for the particle in the box in that zero and negative values are allowed.

The energy of our bead on a loop of wire can be expressed in terms of the angular momentum:

$$E_m = \frac{\Pi_m^2}{2M} = \frac{L_m^2}{2M\,R^2}. \tag{9.26}$$

This means that angular momentum and energy are compatible variables in this case, which further means that angular momentum is a conserved variable.

Just as definite values of linear momentum are related to invariance under translations, definite values of angular momentum are related to invariance under rotations.

Thus, we have

$$\text{invariance under rotation} \Longleftrightarrow \text{definite angular momentum} \tag{9.27}$$

for angular momentum.

We need to briefly address the issue of angular momentum in three dimensions.

Angular momentum is actually a vector oriented perpendicular to the wire loop in the example we are discussing.

The direction of the vector is defined using a variation on the right-hand rule: Curl your fingers in the direction of motion of the bead around the loop (using your right hand!).

The orientation of the angular momentum vector is defined by the direction in which your thumb points.

This tells you, for instance, that the angular momentum in figure 9.5 points out of the page.

In quantum mechanics it is only possible to measure simultaneously the square of the length of the angular momentum vector and one component of this vector.

Two different components of angular momentum cannot be simultaneously measured because of the uncertainty principle.

However, the *length* of the angular momentum vector may be measured simultaneously with one component.

Thus, in quantum mechanics, the angular momentum is completely specified if the length and one component of the angular momentum vector are known.

Figure 9.6 illustrates the angular momentum vector associated with a bead moving on a wire loop which is tilted from the horizontal.

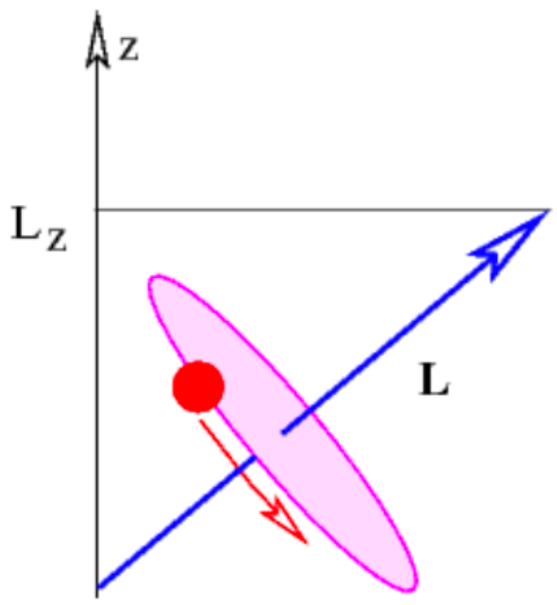


Figure 9.6: Illustration of the angular momentum vector **L** for a tilted loop and its z component L_z.

One component (taken to be the z component) is shown as well.

For reasons we cannot explore here, the square of the length of the angular momentum vector L^2 is quantized with the following values:

$$L_l^2 = \hbar^2 l(l+1), \qquad l = 0, 1, 2, \ldots. \tag{9.28}$$

One component (say, the z component) of angular momentum is quantized just like angular momentum in the two-dimensional case, except that l acts as an upper bound on the possible values of $|m|$.

In other words, if the square of the length of the angular momentum vector is $\hbar^2 l(l+1)$, then the z component can take on the values

$$L_{zm} = \hbar m, \qquad m = -l, -l+1, \ldots, l-1, l. \tag{9.29}$$

The quantity l is called the angular momentum quantum number, while m is called the orientation or *magnetic quantum number*, the latter for historical reasons.

9.3.4 Spin Angular Momentum

The type of angular momentum discussed above is associated with the movement of particles in orbits.

However, it turns out that even stationary particles can possess angular momentum.

This is called *spin* angular momentum.

The *spin quantum number s* plays a role analogous to l for spin angular momentum, i. e., the square of the spin angular momentum vector of a particle is

$$L_s^2 = \hbar^2 s(s+1). \tag{9.30}$$

The spin orientation quantum number m_s is similarly related to *s*:

$$L_{zs} = \hbar m_s, \qquad m_s = -s, -s+1, \ldots, s-1, s. \tag{9.31}$$

The spin angular momentum for an elementary particle is absolutely conserved, i. e., it can never change.

Thus, the value of *s* is an intrinsic property of a particle.

The major difference between spin and orbital angular momentum is that the spin quantum number can take on more values, i. e., $s = 0, 1/2, 1, 3/2, 2, 5/2, \ldots$.

Particles with integer spin values $s = 0, 1, 2, \ldots$ are called *bosons* after the Indian physicist Satyendra Nath Bose.

Particles with half-integer spin values $s = 1/2, 3/2, 5/2, \ldots$ are called *fermions* after the Italian physicist Enrico Fermi.

As we shall see later in the course, bosons and fermions play very different roles in the universe.

Chapter 10

Dynamics of Multiple Particles

So far we have considered only the dynamics of a single particle subject to an externally imposed potential energy.

The particle has no way of influencing this external agent.

In the real world particles interact with each other.

In this chapter we learn how this happens.

We first rewrite Newton's second law in terms of momentum.

This is useful in the subsequent consideration of Newton's third law, which leads to the principle of the conservation of momentum.

Collisions between particles and the behavior of rockets and conveyor belts are then studied as applications of the conservation laws to more than one particle.

10.1 Momentum and Newton's Second Law

Up to this point we have stated Newton's second law in its conventional form, $\mathbf{F} = m\mathbf{a}$.

However, in the non-relativistic case $m\mathbf{a} = md\mathbf{u}/dt = d(m\mathbf{u})/dt$, so we can also write Newton's second law as

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} \quad \text{(Newton's second law)} \tag{10.1}$$

where $\mathbf{p} = m\mathbf{u}$ is the non-relativistic kinetic momentum.

This form of Newton's second law is actually closer to Newton's original statement of the law.

It also has the advantage that it is correct even in the relativistic case when the relativistic definition of kinetic momentum, $\mathbf{p} = m\mathbf{u}/(1 - u^2/c^2)^{1/2}$ (as defined earlier), is substituted.

10.2 Newton's Third Law

Newton's third law states that if particle A exerts a force **F** on particle B, then particle B exerts a force -**F** on particle A.

Newton's third law makes it possible to apply Newton's second law to systems of particles without considering the detailed interactions between particles within the system.

For instance, if we (arbitrarily) define the system in figure 10.1 to be the particles A and B inside the dashed lines,

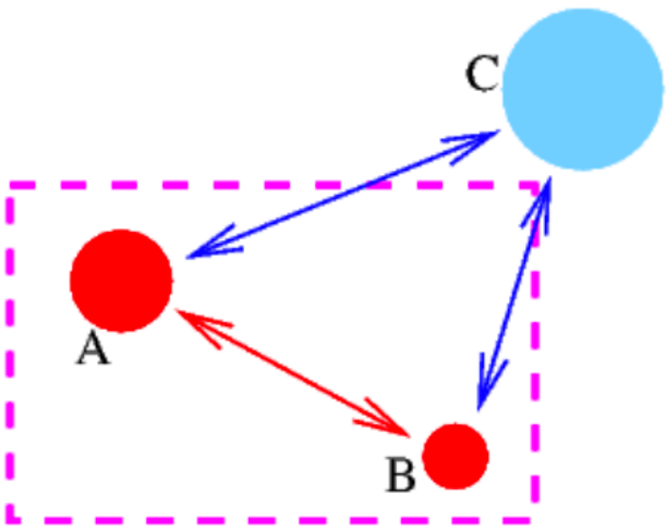


Figure 10.1: Interactions between three particles, A, B, and C. A and B are considered to be part of the system defined by the dashed line.

then we can divide the forces acting on these particles into internal and external parts,

$$\mathbf{F}_A = \mathbf{F}_{A-internal} + \mathbf{F}_{A-external} = \frac{d\mathbf{p}_A}{dt}, \tag{10.2}$$

$$\mathbf{F}_B = \mathbf{F}_{B-internal} + \mathbf{F}_{B-external} = \frac{d\mathbf{p}_B}{dt}. \tag{10.3}$$

Adding these equations together results in the net force $\mathbf{F}_{net}$ being equal to

$$\mathbf{F}_{A-internal} + \mathbf{F}_{A-external} + \mathbf{F}_{B-internal} + \mathbf{F}_{B-external} = \frac{d}{dt}(\mathbf{p}_A + \mathbf{p}_B). \tag{10.4}$$

However, the internal interactions in this case are A acting on B and B acting on A.

These forces are equal in magnitude but opposite in direction, so they cancel out, leaving us with the net force equal to the sum of the external parts, $\mathbf{F}_{\text{net}} = \mathbf{F}_{A\text{-}external} + \mathbf{F}_{B\text{-}external}$.

The external forces in figure 10.1 are the force of C on A and the force of C on B.

Defining the total kinetic momentum of the system as the sum of the A and B momenta, $\mathbf{p}_{tot} = \mathbf{p}_A + \mathbf{p}_B$, the above equation becomes

$$\mathbf{F}_{net} = \frac{d\mathbf{p}_{tot}}{dt}, \tag{10.5}$$

which looks just like Newton's second law for a single particle, except that it now applies to the system of particles (A and B in the present case) as a whole.

This argument easily generalizes to any number of particles inside and outside the system.

Thus, for instance, even though a soccer ball consists of billions of atoms, we are sure that the forces between atoms within the soccer ball cancel out, and the trajectory of the ball as a whole is determined solely by external forces such as gravity, wind drag, friction with the ground, and the kicks of soccer players.

Remember that for two forces to be a third law pair, they have to be acting on different particles.

Furthermore, if one member of the pair is the force of particle A acting on particle B, then the other must be the force of particle B acting on particle A.

A counterexample would be gravity and the upward normal force acting on a mass sitting on a table; these forces are equal and opposite in the stationary case, but act on the same object, and therefore *do not* constitute a third law pair.

However, the upward normal force of the table on the mass and the downward normal force of the mass on the table *would be* a third law pair.

10.3 Conservation of Momentum

If all external forces on a system are zero, then equation (10.5) reduces to

$$\mathbf{p}_{tot} = const \quad \text{(isolated system)}. \tag{10.6}$$

A system of particles with no external forces acting on it is called *isolated*.

Newton's third law thus tells us that the kinetic momentum of an isolated system doesn't change with time.

This law is called the *conservation of momentum*.

10.4 Collisions

Let us now consider the situation in which two particles collide with each other.

There can be several outcomes to this collision, of which we will study two:

- The two particles collide *elastically*, in essence bouncing off of each other.
- The two particles stick together, resulting in the production of a single particle, or a single particle breaks apart into two particles. These are *inelastic processes*.

In both of the above cases energy and momentum are conserved.

We assume that the forces acting between the particles are short range, so that except in the instant of collision, we need not worry about potential energy or potential momentum — all energy is in the form of rest plus kinetic energy except in this short interval, and all momenta are kinetic momenta.

Because of the principle of relativity, we are free to consider collisions in any convenient reference frame.

We can then transform the results to any reference frame we please.

Generally speaking, the most convenient reference frame to consider is the one in which the total momentum of the two particles is zero.

For the sake of simplicity we only consider collisions in one dimension.

10.4.1 Elastic Collisions

Suppose a particle with mass m_1 and initial velocity u_1 in the *center of momentum* frame, i. e., the reference frame in which the total momentum is zero, collides elastically with another particle of mass m_2 with initial velocity u_2.

The momenta of the two particles are

$$p_1 = \frac{m_1 u_1}{(1 - u_1^2/c^2)^{1/2}} \quad p_2 = \frac{m_2 u_2}{(1 - u_2^2/c^2)^{1/2}}. \tag{10.7}$$

In the center of momentum frame we must have

$$p_1 = -p_2. \tag{10.8}$$

Figure 10.2 shows what happens when these two particles collide.

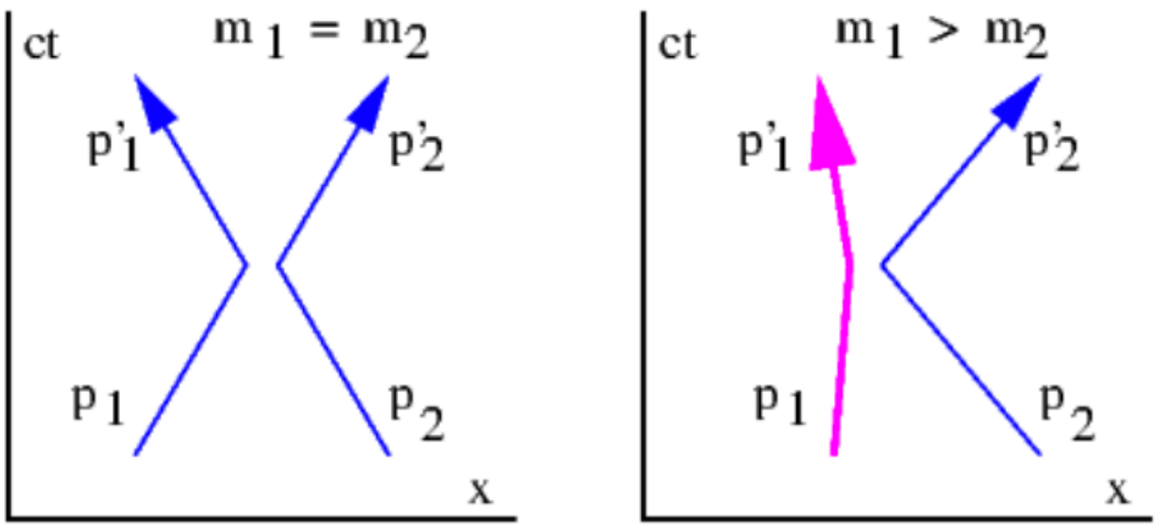


Figure 10.2: One-dimensional elastic collisions of two particles in the center of momentum frame as seen in spacetime diagrams

The first particle acquires momentum p_1' while the second acquires momentum p_2'.

The conservation of momentum tells us that the total momentum after the collision is the same as before the collision, namely zero, so

$$p_1' = -p_2'. \tag{10.9}$$

In the center of momentum frame we know that $|p_1| = |p_2|$ and we know that the two momentum vectors point in opposite directions.

Similarly, $|p_1'| = |p_2'|$.

However, we as yet don't know how p_1' is related to p_2'.

Conservation of energy,

$$E_1 + E_2 = E_1' + E_2', \tag{10.10}$$

gives us this information.

Notice that if $p_1' = -p_1$, then $E_1'^2 = p_1'^2c^2 + m_1^2c^4 = p_1^2c^2 + m_1^2c^4 = E_1^2$.

Assuming positive energies, we therefore have $E_1' = E_1$.

If $p_2' = -p_2$, then we can similarly infer that E_2' - E_2.

If these conditions are satisfied, then so is equation (10.10).

Therefore, a complete solution to the problem is

$$p_1 = -p_1' = -p_2 = p_2' \equiv p \tag{10.11}$$

and

$$E_1 = E_1' = (p^2c^2 + m_1^2c^4)^{1/2} \quad E_2 = E_2' = (p^2c^2 + m_2^2c^4)^{1/2}. \tag{10.12}$$

In other words, the particles just exchange momenta.

The left panel of figure 10.2 shows what happens in a collision when the masses of the two colliding particles are equal.

If $m_1 = m_2$, then the incoming and outgoing velocities of the two particles are the same, as indicated by the inverse slopes of the world lines.

On the other hand, if $m_1 > m_2$, then the velocity of particle 2 is greater than the velocity of particle 1, as is illustrated in the right panel of figure 10.2.

Suppose we wish to view the results of an elastic collision in a reference frame in which particle 2 is initially stationary.

All we have to do is to transform the velocities into a reference frame moving with the initial velocity of particle 2, as illustrated in figure 10.3.

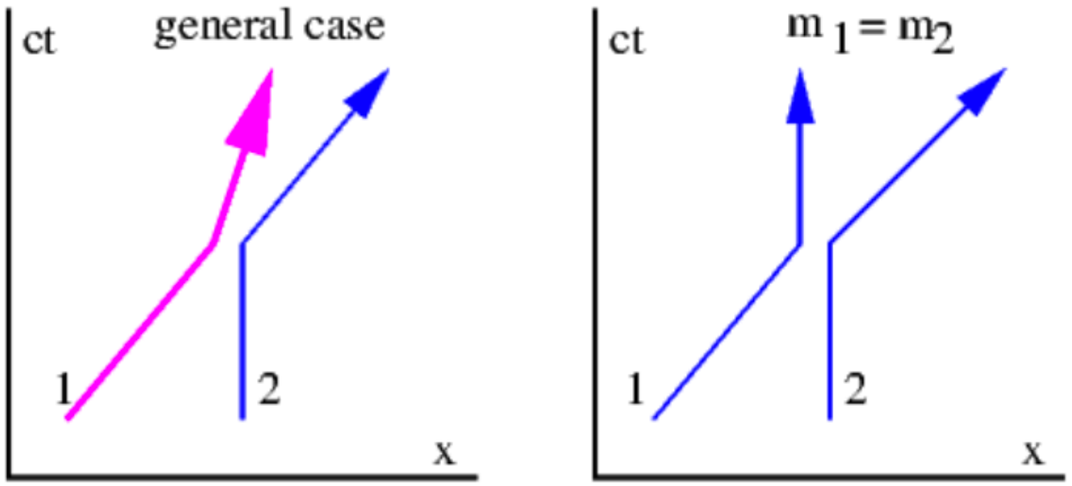


Figure 10.3: Elastic collisions viewed from a reference frame in which one particle is initially stationary.

We do this by relativistically adding $U = -u_2$ to each velocity.

Note that the velocity U of the moving frame is positive since u_2 is negative.

Using the relativistic velocity translation formula, we find that

$$v_1 = \frac{u_1 + U}{1 + u_1 U/c^2} \quad v_1' = \frac{u_1' + U}{1 + u_1' U/c^2} \quad v_2' = \frac{u_2' + U}{1 + u_2' U/c^2} \tag{10.13}$$

where u_1, u_1', u_2, and u_2' indicate velocities in the original, center of momentum reference frame and v_1,v_1' , etc., indicate velocities in the transformed frame.

In the special case where the masses of the two particles are equal to each other, we have $v_1 = 2U/(1+U^2/c^2)$, $v_1' = 0$, and $v_2' = 2U/(1 + U^2/c^2) = v_1$.

Thus, when the masses are equal, the particles simply exchange velocities.

If the velocities are nonrelativistic, then the simpler Galilean transformation law $v = u + U$ can be used in place of the relativistic equations invoked above.

10.4.2 Inelastic Collisions

An inelastic collision is one in which the particles coming out of the collision are not the same as the particles going into it.

Inelastic collisions conserve both total momentum and energy just as elastic collisions do.

However, unlike elastic collisions, inelastic collisions generally do not conserve the total *kinetic* energy of the particles, as some rest energy is generally created or destroyed.

Figure 10.4 shows the fundamental building blocks of inelastic collisions.

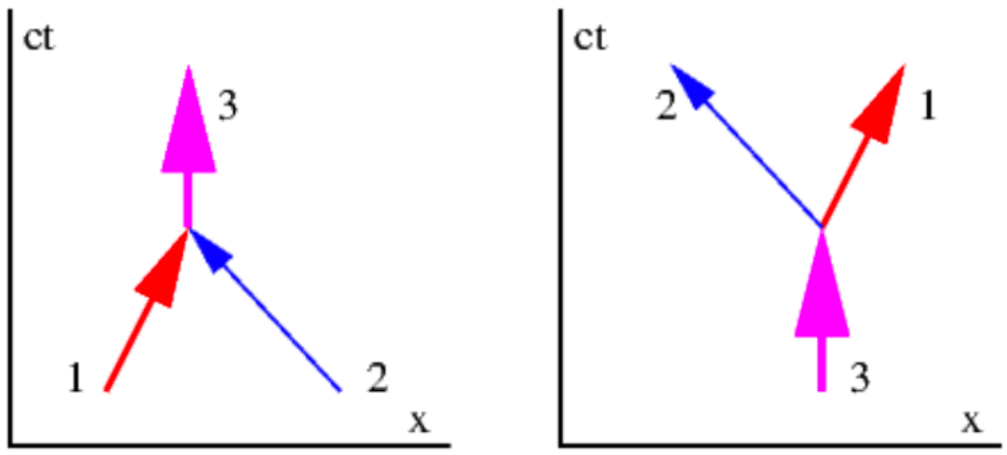


Figure 10.4: Building blocks of inelastic collisions. In the left panel two particles collide to form a third particle. In the right panel a particle breaks up, forming two particles.

We can consider even the most complex inelastic collisions to be made up of composites of only two processes, the creation of one particle from two, and the disintegration of one particle into two.

Let us consider each of these in the center of momentum frame.

In both cases the single particle must be stationary in this frame since it carries the total momentum of the system, which has to be zero.

By conservation of momentum, if particle 1 in the left panel of figure 10.4 has momentum p, then the momentum of particle 2 is -p.

If the two particles have masses m_1 and m_2, then their energies are $E_1 = (p^2c^2 + m_1^2c^4)^{1/2}$ and $E_2 = (p^2c^2 + m_2^2c^4)^{1/2}$.

The energy of particle 3 is therefore $E_3 = E_1 + E_2$, and since it is at rest, all of its energy is in the form of "mc^2" or rest energy, and so the mass of this particle is

$$\begin{aligned} m_3 &= (E_1 + E_2)/c^2 \quad \text{(center of momentum frame)} \\ &= (p^2/c^2 + m_1^2)^{1/2} + (p^2/c^2 + m_2^2)^{1/2} \\ &= m_1[1 + p^2/(m_1^2c^2)]^{1/2} + m_2[1 + p^2/(m_2^2c^2)]^{1/2}. \end{aligned} \tag{10.14}$$

The last line in the above equation shows that $m_3 > m_1 + m_2$ because it is in the form $m_1A + m_2B$ where both A and B are greater than one.

Thus, rest energy is created in the amount $\Delta E_{\text{rest}} = (m_3 - m_1 - m_2)c^2$.

Actually, it is easy to calculate the mass of particle 3 in the above case from any reference frame as long as the momenta and energies of particles 1 and 2 are known in this frame.

By conservation of energy and momentum, $E_3 = E_1 + E_2$ and $\mathbf{p}_3 = \mathbf{p}_1 + \mathbf{p}_2$.

Furthermore, $E_3^2 = p_3^2c^2 + m_3^2c^4$, so we can solve for m_3:

$$m_3 = [(E_1 + E_2)^2/c^4 - (\mathbf{p}_1 + \mathbf{p}_2) \cdot (\mathbf{p}_1 + \mathbf{p}_2)/c^2]^{1/2} \quad \text{(any frame)}. \tag{10.15}$$

The right panel of figure 10.4 shows the process of particle decay.

This is just the inverse of the particle creation process, and all of the analysis we have done for creation is valid for particle decay except that rest energy is converted to kinetic energy rather than vice versa.

10.5 Rockets and Conveyor Belts

Normally when we define a system to which Newton's second law is to be applied, the system is *closed* in the sense that mass cannot enter or exit the system.

However, sometimes it is convenient to work with *open* systems for which this is not true.

The classic example is the rocket, where exhaust gases leave the system, thus decreasing the mass of the rocket with time.

Open systems can be analyzed if momentum is considered to be a quantity which is accounted for much as money is accounted for in a bank account.

The bank account can change in three ways: money can be deposited in the account, it can be withdrawn from the account, and the amount can grow or shrink as a consequence of interest payments or fees.

Similarly, the amount of momentum in a system can change as the result of mass entering the system, mass leaving the system, and forces acting on the system.

The time rate of change of momentum in a system is therefore

$$\frac{dp}{dt} = F + \left(\frac{dp}{dt}\right)_{in} - \left(\frac{dp}{dt}\right)_{out}, \tag{10.16}$$

where F is the net force on the system, $(dp/dt)_{in}$ is the momentum per unit time added by mass entering the system, and $(dp/dt)_{out}$ is the amount lost per unit time by mass exiting the system.

In the non- relativistic case, $(dp/dt)_{in} = u_{in}(dm/dt)_{in}$ and $(dp/dt)_{out} = u_{out}(dm/dt)_{out}$ where $(dm/dt)_{in}$ is the mass entering the system per unit time with velocity u_{in} and $(dm/dt)_{out}$ is the mass per unit time exiting the system with velocity *uout*.

For non-relativistic velocities, the momentum of the system can be written as $p = mu$ so that

$$\frac{dp}{dt} = \frac{dm}{dt}u + m\frac{du}{dt}. \tag{10.17}$$

To complete the analysis, we need an accounting of the mass entering and leaving the system:

$$\frac{dm}{dt} = \left(\frac{dm}{dt}\right)_{in} - \left(\frac{dm}{dt}\right)_{out} . \qquad (10.18)$$

Let us see how to apply this to a rocket for which all velocities are non-relativistic.

As figure 10.5 indicates, a rocket spews out a stream of exhaust gas.

Figure 10.5: Rocket moving with velocity V while expelling gas at a rate R with velocity V - u_x.

The system is defined by the dashed box and includes the rocket and the part of the exhaust gas inside the box.

The reaction to the momentum carried off in this stream of gas is what causes the rocket to accelerate.

We note that $(dm/dt)_{in} = 0$ since no mass is entering the system, and $(dm/dt)_{out} = R$, the rate at which mass is ejected by the rocket in the form of exhaust gas.

The rocket is assumed to be moving to the right at speed V and the gas is ejected at a speed u_x relative to the rocket, which means that its actual velocity after ejection is V - u_x.

We call u_x the *exhaust velocity*. Notice that V - u_x may be either positive or negative, depending on how big V is.

Equating the mass of the rocket to the system mass, we find that $R = -dm/dt$.

The momentum balance equation (10.16) becomes $dp/dt = -(V - u_x)R$.

The force on the rocket is actually zero, so the force term does not enter the momentum balance equation.

This is non-intuitive, because we are used to acceleration being the result of a force.

However, nothing, including the ejected gas, is actually *pushing* on the system, so we must indeed conclude that there is no force — all of the change in the system's momentum arises from the ejection of gas with the opposite momentum.

Finally, we see that $dp/dt = (dm/dt)V + m(dV/dt) = -RV + m(dV/dt)$.

Equating this to the results of the momentum balance calculation gives us $-RV + m(dV/dt) = -(V - u_x)R$.

Solving for the acceleration dV/dt results in

$$\frac{dV}{dt} = \frac{u_x R}{m} \quad \text{(rocket acceleration)}. \tag{10.19}$$

Thus, the acceleration of the rocket depends on the exhaust velocity of the ejected gas, the rate at which the gas is being ejected, and the mass of the rocket.

Figure 10.6 illustrates another type of open system problem.

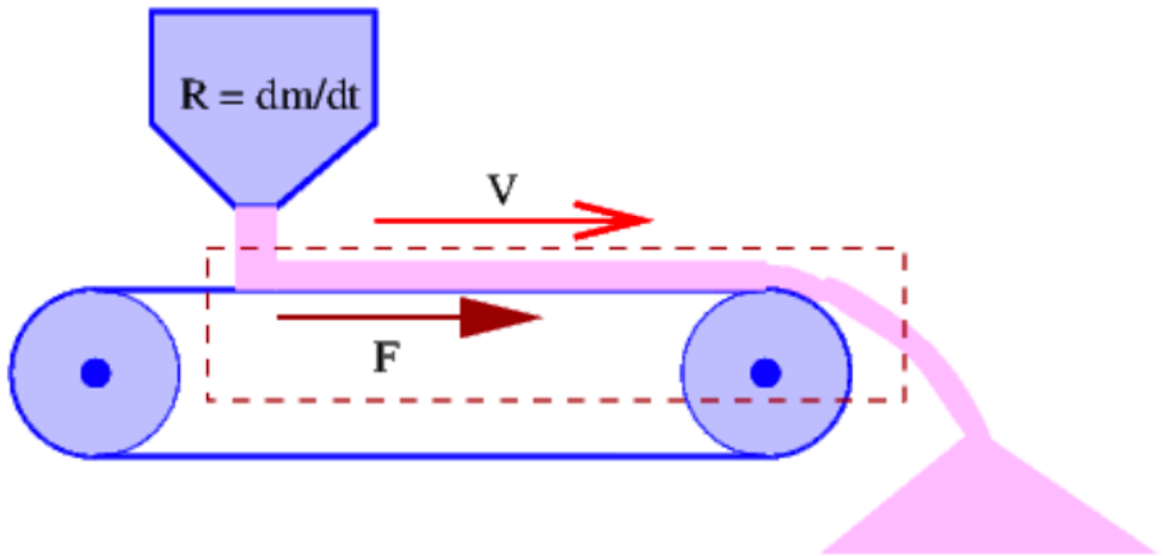


Figure 10.6: Sand is dumped on a conveyor belt and in turn is dumped off the end of the belt.

A hopper dumps sand on a conveyor belt at a rate of R kilograms per second.

The conveyor belt is moving to the right at (non-relativistic) speed V and the sand is dumped off at the end.

What force F is needed to keep the conveyor belt moving at a constant speed, assuming that the conveyor belt mechanism itself is frictionless?

In this case $(dm/dt)_{in} = (\mathrm{dm}/\mathrm{dt})_{\mathrm{out}} = R$.

Furthermore, since the system outlined by the dashed line is in a steady state, $dp/dt = 0$.

The key to understanding this problem is that the sand enters the system with zero horizontal velocity, but exits the system with the horizontal velocity of the conveyor belt, V .

The momentum balance equation is thus

$$0 = F - VR \tag{10.20}$$

and the force is

$$F = VR \quad \text{(force on conveyor belt)}. \tag{10.21}$$

This force serves to accelerate the sand up to the velocity of the conveyor belt.

Chapter 11

Rotational Dynamics

We have already seen the quantum mechanical treatment of angular momentum and rotational dynamics.

In this section we study these subjects in a classical, non-relativistic context.

We first define the concepts of torque and angular momentum in order to understand the orbital motion of a single particle.

Next we examine two particles in arbitrary motion and learn how kinetic energy and angular momentum are partitioned between orbital and internal components.

Two particles fixed to the ends of a light rod constitute a dumbbell, which serves as a prototype for the rotation of rigid bodies.

We then see how what we learned for two particles extends to an arbitrary number of particles.

Finally, we explore the physics of structures in static equilibrium.

Before we begin, we need to extend our knowledge of vectors to the cross product.

11.1 Math Tutorial — Cross Product

There are two ways to multiply two vectors together, the dot product and the cross product.

We have already studied the dot product of two vectors, which results in a scalar or single number.

The cross product of two vectors results in a third vector, and is written symbolically as follows:

$$\mathbf{C} = \mathbf{A} \times \mathbf{B}. \tag{11.1}$$

As illustrated in figure 11.1, the cross product of two vectors is perpendicular to the plane defined by these vectors.

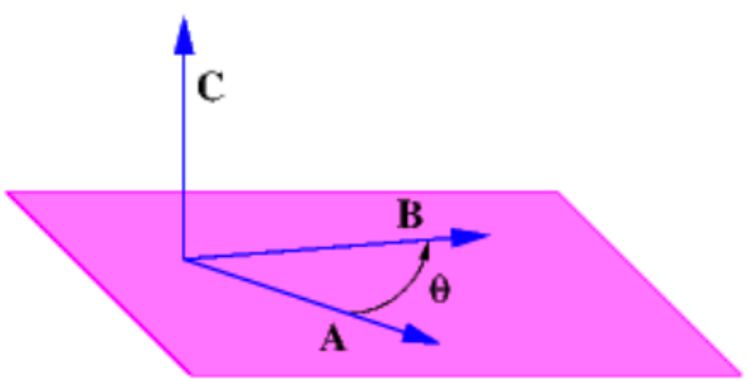


Figure 11.1: Illustration of the cross product of two vectors **A** and **B**. The resulting vector **C** is perpendicular to the plane defined by **A** and **B**.

However, this doesn't tell us whether the resulting vector in figure 11.1 points upward out of the plane or downward.

This ambiguity is resolved using the *right-hand rule*:

1. Point the uncurled fingers of your right hand along the direction of the first vector **A**.

2. Rotate your arm until you can curl your fingers in the direction of the second vector **B**.

3. Your stretched out thumb now points in the direction of the cross product vector **C**.

The magnitude of the cross product is given by

$$|\mathbf{C}| = |\mathbf{A}||\mathbf{B}|\sin(\theta), \tag{11.2}$$

where |**A**| and |**B**| are the magnitudes of **A** and **B**, and θ is the angle between these two vectors.

Note that the magnitude of the cross product is zero when the vectors are parallel or anti-parallel, and maximum when they are perpendicular.

This contrasts with the dot product, which is maximum for parallel vectors and zero for perpendicular vectors.

Notice that the cross product does not commute, i. e., the order of the vectors is important.

In particular, it is easy to show using the right-hand rule that

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}. \tag{11.3}$$

An alternate way to compute the cross product is most useful when the two vectors are expressed in terms of components, i. e., $\mathbf{A} = (A_x, A_y, A_z)$ and $\mathbf{B} = (B_x, B_y, B_z)$:

$$\begin{aligned} C_x &= A_y B_z - A_z B_y \\ C_y &= A_z B_x - A_x B_z \\ C_z &= A_x B_y - A_y B_x. \end{aligned} \tag{11.4}$$

Notice that once you have the first of these equations, the other two can be obtained by cyclically permuting the indices, i. e., $x \rightarrow y$, $y \rightarrow z$, and $z \rightarrow x$.

This is useful as a memory aid.

11.2 Torque and Angular Momentum

Torque is the action of a force **F** on a mass *M* which induces it to revolve about some point, called the *origin*.

It is defined

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}, \tag{11.5}$$

where **r** is the position of the mass relative to the origin, as illustrated in figure 11.2.

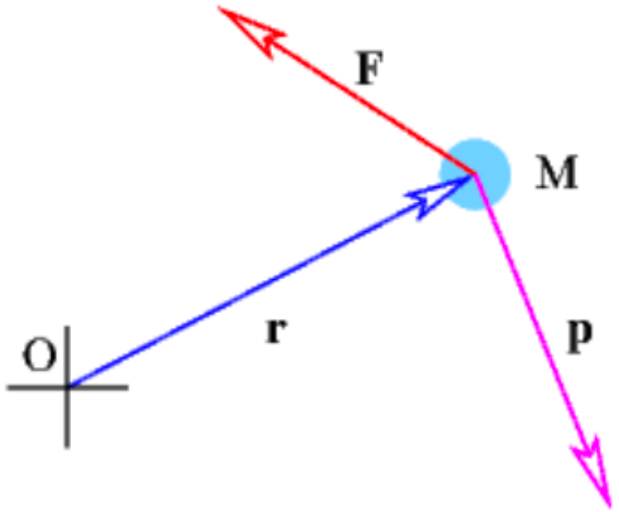


Figure 11.2: A mass *M* located at **r** relative to the origin O has momentum **p** and has a force **F** applied to it. By the right-hand rule the torque $\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}$ points out of the page, while the angular momentum $\mathbf{L} = \mathbf{r} \times \mathbf{p}$ points into the page.

Notice that the torque is zero in a number of circumstances.

If the force points directly toward or away from the origin, the cross product is zero, resulting in zero torque, even though the force is non-zero.

Likewise, if $\mathbf{r} = 0$, the torque is zero.

Thus, a force acting at the origin produces no torque.

Both of these limits make sense intuitively, since neither induces the mass to revolve around the origin.

The *angular momentum* of a mass *M* relative to a point O is

$$\mathbf{L} = \mathbf{r} \times \mathbf{p}, \tag{11.6}$$

where **p** is the ordinary kinetic momentum of the mass.

The angular momentum is zero if the motion of the object is directly towards or away from the origin, or if it is located at the origin.

If we take the cross product of the position vector and Newton's second law, we obtain an equation that relates torque and angular momentum:

$$\mathbf{r} \times \mathbf{F} = \mathbf{r} \times \frac{d\mathbf{p}}{dt} = \frac{d}{dt}(\mathbf{r} \times \mathbf{p}) - \frac{d\mathbf{r}}{dt} \times \mathbf{p}. \tag{11.7}$$

The second term on the right side of the above equation is zero because $d\mathbf{r}/dt$ equals the velocity of the mass, which is parallel to its momentum and the cross product of two parallel vectors is zero.

This equation can therefore be written

$$\boldsymbol{\tau} = \frac{d\mathbf{L}}{dt} \quad \text{(Newton's second law for rotation)}. \tag{11.8}$$

It is the rotational version of Newton's second law.

For both torque and angular momentum the location of the origin is arbitrary, and is generally chosen for maximum convenience.

However, it is necessary to choose the same origin for both the torque and the angular momentum.

For the case of a *central force*, i. e., one which acts along the line of centers between two objects (such as gravity), there often exists a particularly convenient choice of origin.

Imagine a planet revolving around the sun, as illustrated in figure 11.3.

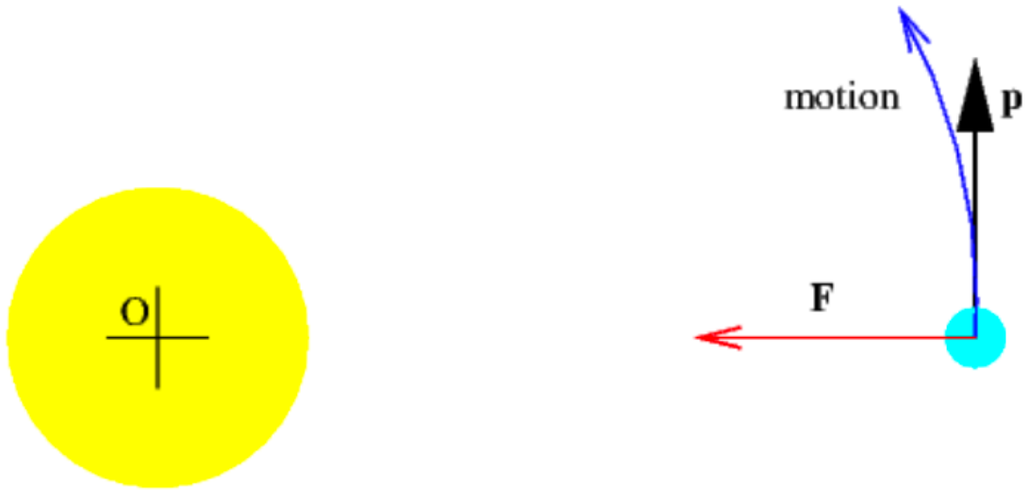


Figure 11.3: A convenient choice of origin for a planet (right-hand sphere) revolving around the sun is simply the center of the sun. In this case the torque of the sun's gravitational force on the planet is zero.

If the origin is placed at the center of the sun (which is assumed not to move under the influence of the planet's gravity), then the torque exerted on the planet by the sun's gravity is zero, which means that the angular momentum of the planet about the center of the sun is constant in time.

No other choice of origin would yield this convenient result.

We already know about two fundamental conservation laws — those of energy and linear momentum.

We believe that angular momentum is similarly conserved in isolated systems.

In other words, particles can exchange angular momentum between themselves, but the vector sum of the angular momentum of all the particles in a system isolated from outside influences must remain constant.

Conservation of angular momentum is not an automatic consequence of the conservation of linear momentum, even though the governing equation (11.8) for angular momentum is derived from Newton's second law.

As an example, figure 11.4 shows a hypothetical situation in which the force $\mathbf{F}_{21}$ of M_1 on M_2 is equal in magnitude but opposite in sign to the force $\mathbf{F}_{12}$ of M_2 on M_1, i. e., Newton's third law holds, and the sum of the momenta of the two masses is conserved

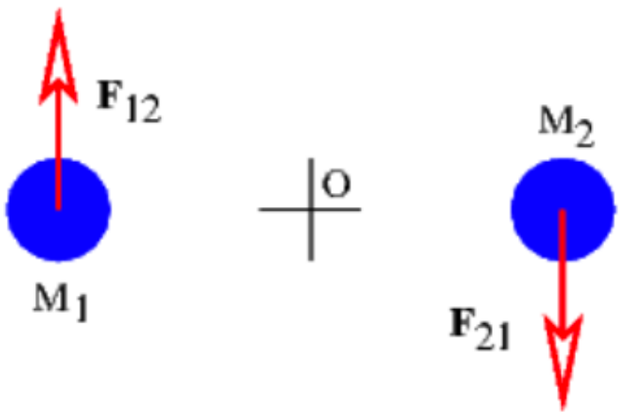


Figure 11.4: Scenario for the non-conservation of angular momentum. $\mathbf{F}_{12}$ is the force of mass M_2 on mass M_1 and vice versa.

However, because the forces are non-central, the angular momentum of the masses is not conserved.

This scenario is impossible if the forces are central.

11.3 Two Particles

Suppose we wish to apply Newton's second law to two particles considered together as a single system.

As we showed previously, only external forces act on the total momentum, $\mathbf{p}_{total} = \mathbf{p}_1 + \mathbf{p}_2$, of the two particles:

$$\mathbf{F}_{external} = \frac{d\mathbf{p}_{total}}{dt}. \tag{11.9}$$

Let's write the total non-relativistic momentum of the two particles in a special way:

$$\mathbf{p}_{total} = M_1\mathbf{v}_1 + M_2\mathbf{v}_2 = M_{total}\left(\frac{M_1\mathbf{v}_1 + M_2\mathbf{v}_2}{M_{total}}\right) \equiv M_{total}\mathbf{V}_{cm}, \tag{11.10}$$

where $M_{total} = M_1 + M_2$.

The quantity $\mathbf{V}_{cm}$ is the velocity of the center of mass and can be expressed as the time derivative of the position of the center of mass, $\mathbf{R}_{cm}$,

$$\mathbf{V}_{cm} = \frac{d\mathbf{R}_{cm}}{dt}, \tag{11.11}$$

where

$$\mathbf{R}_{cm} = \frac{M_1\mathbf{r}_1 + M_2\mathbf{r}_2}{M_{total}}. \tag{11.12}$$

We now see how the kinetic energy and the angular momentum of the two particles may be split into two parts, one having to do with the motion of the center of mass of the two particles, the other having to do with the motion of the two particles relative their center of mass.

Figure 11.5 shows graphically how the vectors $\mathbf{r}'_1 = \mathbf{r}_1 - \mathbf{R}_{cm}$ and $\mathbf{r}'_2 = \mathbf{r}_2 - \mathbf{R}_{cm}$ are defined.

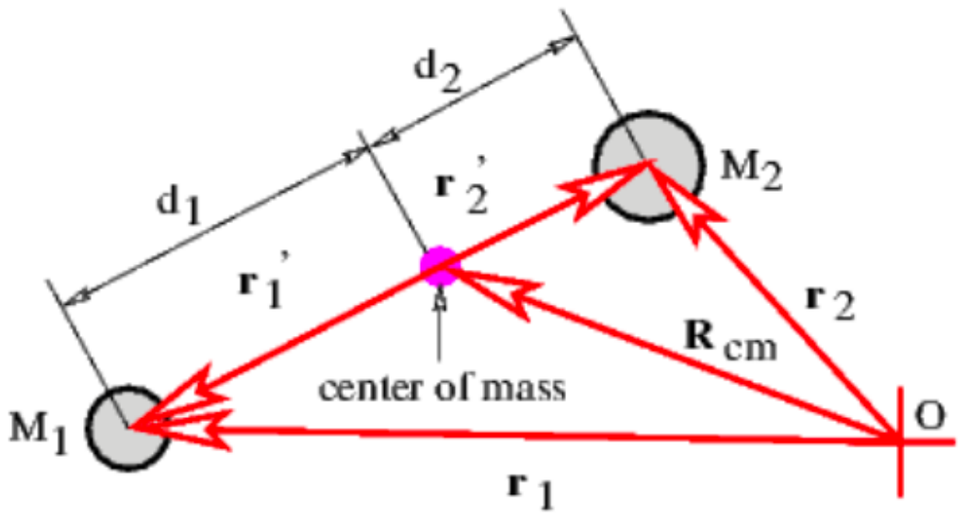


Figure 11.5: Two particles of mass M_1 and M_2 with $M_2 > M_1$.

These vectors represent the positions of the two particles relative to the center of mass.

Substitution into equation (11.12) shows that $M_1\mathbf{r}'_1 + M_2\mathbf{r}'_2 = 0$.

This leads to the conclusion that $M_1 d_1 = M_2 d_2$ in figure 11.5.

We also define the velocity of each mass relative to the center of mass as $\mathbf{v}'_1 = d\mathbf{r}'_1/dt$ and $\mathbf{v}'_2 = d\mathbf{r}'_2/dt$, and we therefore have $M_1\mathbf{v}'_1 + M_2\mathbf{v}'_2 = 0$.

The total kinetic energy is just the sum of the kinetic energies of the two particles, $K = M_1 v_1^2/2 + M_2 v_2^2/2$, where v_1 and v_2 are the magnitudes of the corresponding velocity vectors.

Substitution of $\mathbf{v}_1 = \mathbf{V}_{cm} + \mathbf{v}'_1$ etc., into the kinetic energy formula and rearranging yields

$$K_{total} = K_{trans} + K_{intern} = [M_{total} V_{cm}^2/2] + [M_1 v_1'^2/2 + M_2 v_2'^2/2]. \qquad (11.13)$$

Terms like $\mathbf{V}_{cm} \cdot \mathbf{v}'_1$ cancel out because $M_1\mathbf{v}'_1 + M_2\mathbf{v}'_2 = 0$.

The first term on the right side of equation (11.13) in square brackets is the kinetic energy the two particles would have if all of the mass were concentrated at the center of mass.

The second term is the kinetic energy computed relative to the motion of the center of mass.

The first is called the *translational kinetic energy* of the system while the second is called the *internal kinetic energy*.

The angular momentum of the system is just the sum of the angular momenta of the two particles: $\mathbf{L}_{total} = M_1\mathbf{r}_1 \times \mathbf{v}_1 + M_2\mathbf{r}_2 \times \mathbf{v}_2$.

By reasoning similar to the case of kinetic energy, we can rewrite this as

$$\mathbf{L}_{total} = \mathbf{L}_{orb} + \mathbf{L}_{spin} = [M_{total}\mathbf{R}_{cm} \times \mathbf{V}_{cm}] + [M_1\mathbf{r}'_1 \times \mathbf{v}'_1 + M_2\mathbf{r}'_2 \times \mathbf{v}'_2]. \qquad (11.14)$$

The first term in square brackets on the right is called the *orbital angular momentum* while the second term is called the *spin angular momentum*.

The former is the angular momentum the system would have if all the mass were concentrated at the center of mass, while the latter is the angular momentum of motion about the center of mass.

Interestingly, the idea of center of mass and the corresponding split of kinetic energy and angular momentum into orbital and spin parts has no useful relativistic generalization.

This is due to the factor of $\gamma \equiv (1 - v^2/c^2)^{-1/2}$ in the relativistic definition of momentum, $\mathbf{p} = m\mathbf{v}\gamma$, which means that

$$\frac{dm\mathbf{r}}{dt} \neq \mathbf{p} \quad \text{(relativistic case)}. \qquad (11.15)$$

11.4 The Uneven Dumbbell

So far we have put no restrictions on the movements of the two particles.

An interesting special case occurs when the particles are connected by a lightweight, rigid rod, giving us a dumbbell.

In order to further simplify things, we assume that the rod is connected rigidly to a fixed axle at the center of mass of the two particles, as shown in figure 11.6.

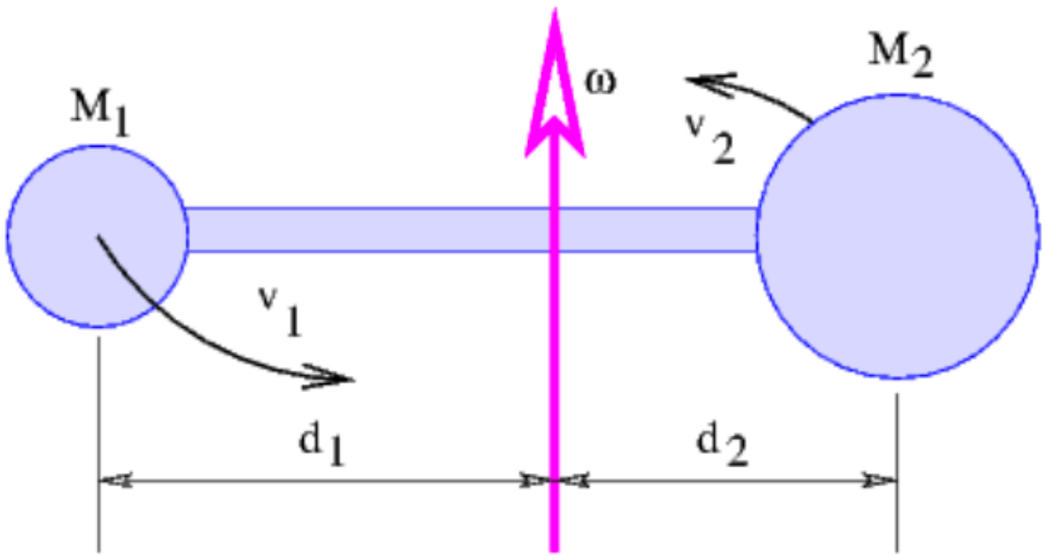


Figure 11.6: Perspective view of the rotating dumbbell attached to an axle labeled ω. The axle attaches to the crossbar at the center of mass. The velocity $\mathbf{v}_1$ is out of the page and $\mathbf{v}_2$ is into the page.

The masses constituting the ends of the dumbbell are therefore free to revolve in circles about the axle, but they are prevented from executing any other motion.

The key effect of this constraint is that both masses rotate about the axle with the same angular frequency ω.

If the particles are respectively distances d_1 and d_2 from the axle, then their speeds are $v_1 = d_1\omega$ and $v_2 = d_2\omega$.

Thus the kinetic energy of the rotating dumbbell is

$$K_{intern} = \frac{1}{2}M_1v_1^2 + \frac{1}{2}M_2v_2^2 = \frac{1}{2}I\omega^2 \quad \text{(fixed axle)}, \tag{11.16}$$

where $I = M_1 d_1^2 + M_2 d_2^2$ is called the *moment of inertia*.

Similarly, the magnitude of the spin angular momentum, which is a vector parallel to the axle, is

$$L_{spin} = M_1 d_1 v_1 + M_2 d_2 v_2 = I\omega \quad \text{(fixed axle)}. \tag{11.17}$$

Finally, Newton's second law for rotation becomes

$$\tau = \frac{dL_{spin}}{dt} = \frac{dI\omega}{dt} = I\frac{d\omega}{dt} \quad \text{(fixed axle)}, \tag{11.18}$$

where τ is the component of torque along the rotation axis.

Note that the rightmost expression in equation (11.18) assumes that I is constant, which only is true if d_1 and d_2 are constant – i. e., the dumbbell must truly be rigid.

11.5 Many Particles

The generalization from two particles to many particles is quite easy in principle.

If a subscripted i indicates the value of a quantity for the ith particle, then the center of mass is given by

$$\mathbf{R}_{cm} = \frac{1}{M_{total}} \sum_i M_i \mathbf{r}_i \tag{11.19}$$

where

$$M_{total} = \sum_i M_i. \tag{11.20}$$

TheFurthermore, if we define $\mathbf{r}'_i = \mathbf{r}_i - \mathbf{R}_{cm}$, etc., then the kinetic energy is just

$$K_{total} = M_{total} V_{cm}^2/2 + \sum_i M_i v_i'^2/2 \tag{11.21}$$

and the angular momentum is

$$\mathbf{L}_{total} = M_{total}\mathbf{R}_{cm} \times \mathbf{V}_{cm} + \sum_i M_i \mathbf{r}'_i \times \mathbf{v}'_i. \tag{11.22}$$

In other words, both the kinetic energy and the angular momentum can be separated into two parts: one part is related to the overall motion of the system and the other is due to motions of system components relative to the center of mass, just as for the case of the dumbbell.

11.6 Rigid Bodies

For a rigid body rotating about a fixed axle, the moment of inertia is

$$I = \sum_i M_i d_i^2, \tag{11.23}$$

where d_i is the perpendicular distance of the ith particle from the axle.

Equations (11.16)-(11.18) are valid for a rigid body consisting of many particles.

Furthermore, the moment of inertia is constant in this case, so it can be taken out of the time derivative:

$$\tau = \frac{dI\omega}{dt} = I\frac{d\omega}{dt} = I\alpha \quad \text{(fixed axle, constant } I\text{)}. \tag{11.24}$$

The quantity $\alpha = d\omega/dt$ is called the angular acceleration.

The sum in the equation for the moment of inertia can be converted to an integral for a continuous distribution of mass.

We shall not pursue this here, but simply quote the results for a number of solid objects of uniform density:

- For rotation of a sphere of mass M and radius R about an axis piercing its center: $I = 2MR^2/5$.
- For rotation of a cylinder of mass M and radius R about its axis of symmetry: $I = MR^2/2$.
- For rotation of a thin rod of mass M and length L about an axis perpendicular to the rod passing through its center: $I = ML^2/12$.
- For rotation of an annulus of mass M, inner radius R_a, and outer radius R_b about its axis of symmetry: $I = M(R_a^2 + R_b^2)/2$.

11.7 Statics

If a rigid body is initially at rest, it will remain at rest if and only if the sum of all the forces and the sum of all the torques acting on the body are zero.

As an example, a mass balance with arms of differing length is shown in figure 11.7.

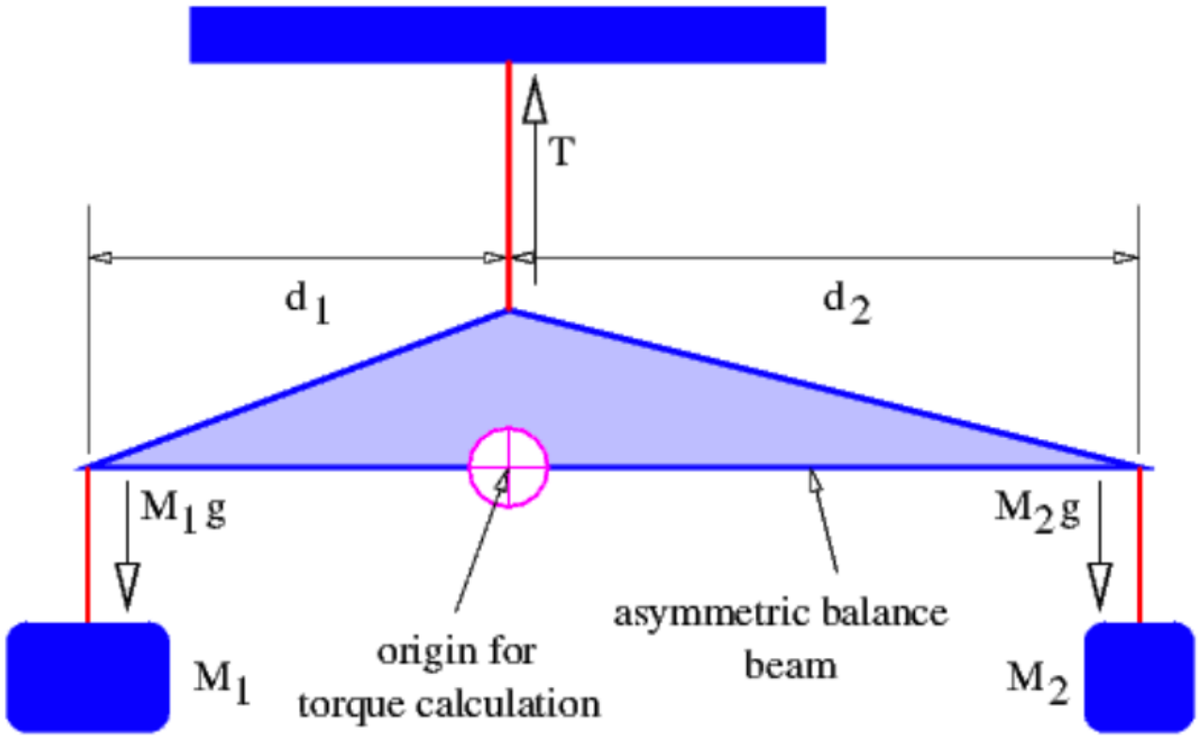


Figure 11.7: Asymmetric mass balance. We assume that the balance beam is massless.

The balance beam is subject to three forces pointing upward or downward, the tension T in the string from which the beam is suspended and the weights M_1g and M_2g exerted on the beam by the two suspended masses.

The parameter g is the local gravitational field and the balance beam itself is assumed to have negligible mass.

Taking upward as positive, the force condition for static equilibrium is

$$T - M_1g - M_2g = 0 \quad \text{(zero net force)}. \tag{11.25}$$

Defining a counterclockwise torque to be positive, the torque balance computed about the pivot point in figure 11.7 is

$$\tau = M_1gd_1 - M_2gd_2 = 0 \quad \text{(zero torque)}, \tag{11.26}$$

where d_1 and d_2 are the lengths of the beam arms.

The first of the above equations shows that the tension in the string must be

$$T = (M_1 + M_2)g, \tag{11.27}$$

while the second shows that

$$\frac{M_1}{M_2} = \frac{d_2}{d_1}. \tag{11.28}$$

Thus, the tension in the string is just equal to the weight of the masses attached to the balance beam, while the ratio of the two masses equals the inverse ratio of the associated beam arm lengths.

Chapter 12

Harmonic Oscillator

Figure 12.1 illustrates the prototypical harmonic oscillator, the mass-spring system.

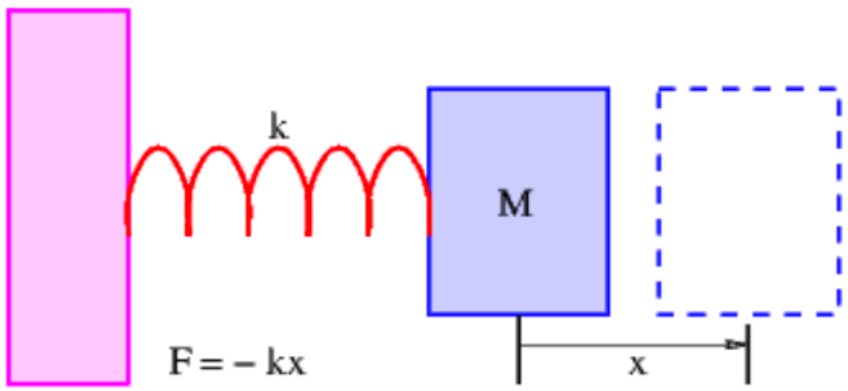


Figure 12.1: Illustration of a mass-spring system.

A mass *M* is attached to one end of a spring.

The other end of the spring is attached to something rigid such as a wall.

The spring exerts a restoring force $F = -kx$ on the mass when it is stretched by an amount x, i. e., it acts to return the mass to its initial position.

This is called *Hooke's law* and k is called the *spring constant*.

12.1 Energy Analysis

The potential energy of the mass-spring system is

$$U(x) = kx^2/2 \tag{12.1}$$

which may be verified by noting that the Hooke's law force is derived from this potential energy: $F = -d(kx^2/2)/dx = -kx$.

This is shown in figure 12.2.

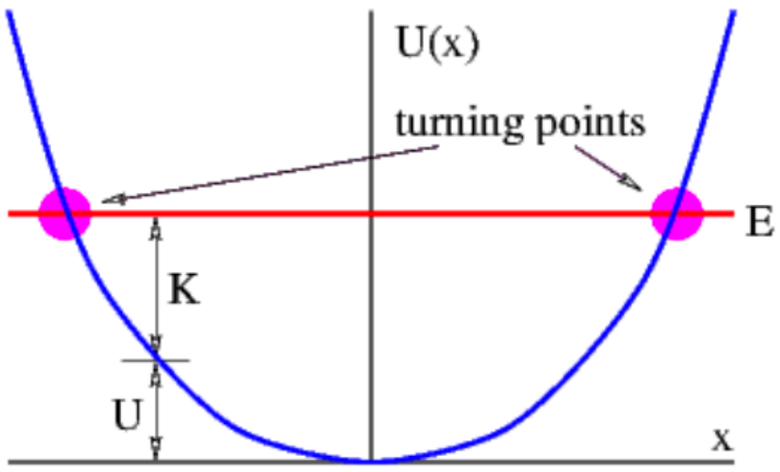


Figure 12.2: Potential, kinetic, and total energy of a harmonic oscillator plotted as a function of spring displacement x.

Since a potential energy exists, the total energy $E = K + U$ is conserved, i. e., is constant in time.

If the total energy is known, this provides a useful tool for determining how the kinetic energy varies with the position x of the mass M: $K(x) = E - U(x)$.

Since the kinetic energy is expressed (non-relativistically) in terms of the velocity u as $K = Mu^2/2$, the velocity at any point on the graph in figure 12.2 is

$$u = \pm \left(\frac{2(E-U)}{M} \right)^{1/2} . \tag{12.2}$$

Given all this, it is fairly evident how the mass moves.

From Hooke's law, the mass is always *accelerating* toward the *equilibrium position*, $x = 0$.

However, at any point the *velocity* can be either to the left or the right.

At the points where $U(x) = E$, the kinetic energy is zero.

This occurs at the turning points

$$x_{TP} = \pm \left(\frac{2E}{k} \right)^{1/2} . \tag{12.3}$$

If the mass is moving to the left, it slows down as it approaches the left turning point.

It stops when it reaches this point and begins to move to the right.

It accelerates until it passes the equilibrium position and then begins to decelerate, stopping at the right turning point, accelerating toward the left, etc.

The mass thus oscillates between the left and right turning points.

Note that equations (12.2) and (12.3) are only true for the harmonic oscillator.)

How does the period of the oscillation depend on the total energy of the system?

Notice that from equation (12.2) the maximum speed of the mass (i. e., the speed at $x = 0$) is equal to $u_{max} = (2E/M)^{1/2}$.

The average speed must be some fraction of this maximum value.

Let us guess here that it is half the maximum speed:

$$u_{average} \approx \frac{u_{max}}{2} = \left(\frac{E}{2M}\right)^{1/2} \quad \text{(approximate)}. \tag{12.4}$$

However, the distance d the mass has to travel for one full oscillation is twice the distance between turning points, or $d = 4(2E/k)^{1/2}$.

Therefore, the period of oscillation must be approximately

$$T = \frac{d}{u_{average}} \approx 4\left(\frac{2E}{k}\right)^{1/2}\left(\frac{2M}{E}\right)^{1/2} = 8\left(\frac{M}{k}\right)^{1/2} \quad \text{(approximate)}. \tag{12.5}$$

12.2 Analysis Using Newton's Laws

The acceleration of the mass at any time is given by Newton's second law:

$$a = \frac{d^2x}{dt^2} = \frac{F}{M} = -\frac{kx}{M}. \tag{12.6}$$

An equation of this type is known as a *differential equation* since it involves a derivative of the dependent variable x.

Equations of this type are generally more difficult to solve than algebraic equations, as there are no universal techniques for solving all forms of such equations.

In fact, it is fair to say that the solutions of most differential equations were originally obtained by *guessing*!

We already have the basis on which to make an intelligent guess for the solution to equation (12.6) since we know that the mass oscillates back and forth with a period that is independent of the amplitude of the oscillation.

A function which might fill the bill is the sine function.

Let us try substituting $x = \sin(\omega t)$, where ω is a constant, into this equation.

The second derivative of x with respect to t is $-\omega^2\sin(\omega t)$, so performing this substitution results in

$$-\omega^2 \sin(\omega t) = -\frac{k}{M} \sin(\omega t). \tag{12.7}$$

Notice that the sine function cancels out, leaving us with $-\omega^2 = -k/M$.

The guess thus works if we set

$$\omega = \left(\frac{k}{M}\right)^{1/2}. \tag{12.8}$$

The constant ω is the angular oscillation frequency for the oscillator, from which we infer the period of oscillation to be $T = 2\pi(M/k)^{1/2}$.

This agrees with the earlier approximate result of equation (12.5), except that the approximation has a numerical factor of 8 rather than $2\pi \approx 6$.

Thus, the earlier guess is only off by about 30%!

It is easy to show that $x = A \sin(\omega t)$ is also a solution of equation (12.6), where A is any constant and $\omega = (k/M)^{1/2}$.

This confirms that the oscillation frequency and period are independent of amplitude.

Furthermore, the cosine function is equally valid as a solution: $x = B \cos(\omega t)$, where B is another constant.

In fact, the most general possible solution is just a combination of these two, i. e.,

$$x = A\sin(\omega t) + B\cos(\omega t). \tag{12.9}$$

The values of A and B depend on the position and velocity of the mass at time $t = 0$.

12.3 Forced Oscillator

If we wiggle the left end of the spring by the amount $d = d_0\sin(\omega_F t)$, as in figure 12.3, rather than rigidly fixing it as in figure 12.1, we have a *forced* harmonic oscillator.

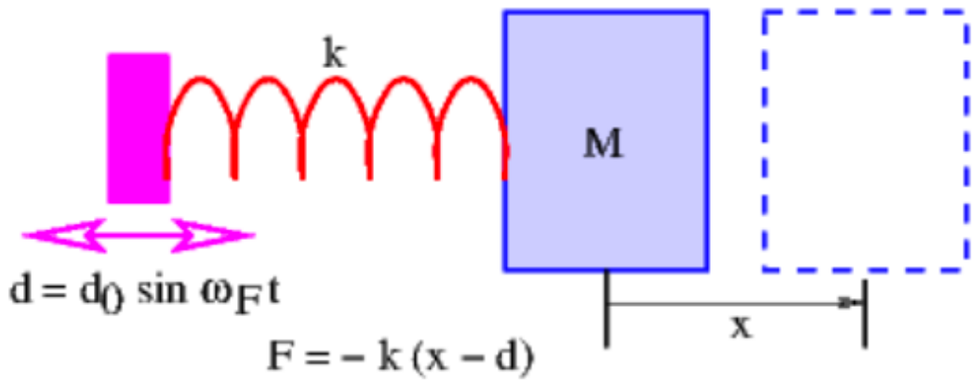


Figure 12.3: Illustration of a forced mass-spring oscillator. The left end of the spring is wiggled back and forth with an angular frequency ω_F and a maximum amplitude d_0.

The constant d_0 is the amplitude of the imposed wiggling motion.

The *forcing frequency* ω_F is not necessarily equal to the natural or *resonant* frequency $\omega = (k/M)^{1/2}$ of the mass-spring system.

Very different behavior occurs depending on whether ω_F is less than, equal to, or greater than ω.

Given the above wiggling, the force of the spring on the mass becomes $F = -k(x - d) = -k[x - d_0\sin(\omega_F t)]$ since the length of the spring is the difference between the positions of the left and right ends.

Proceeding as for the unforced mass-spring system, we arrive at the differential equation

$$\frac{d^2x}{dt^2} + \frac{kx}{M} = \frac{kd_0}{M}\sin(\omega_F t). \tag{12.10}$$

The solution to this equation turns out to be the sum of a forced part in which x is proportional to $\sin(\omega_F t)$ and a free part which is the same as the solution to the unforced equation (12.9).

We are primarily interested in the forced part of the solution, so let us set $x = x_0\sin(\omega_F t)$ and substitute this into equation (12.10):

$$-\omega_F^2 x_0 \sin(\omega_F t) + \frac{kx_0}{M}\sin(\omega_F t) = \frac{kd_0}{M}\sin(\omega_F t). \tag{12.11}$$

Again the sine factor cancels and we are left with an algebraic equation for x_0, the amplitude of the oscillatory motion of the mass.

Solving for the ratio of the oscillation amplitude of the mass to the amplitude of the wiggling motion, x_0/d_0, we find

$$\frac{x_0}{d_0} = \frac{1}{1-\omega_F^2/\omega^2}, \tag{12.12}$$

where we have recognized that $k/M = \omega^2$, the square of the frequency of the *free* oscillation.

This function is plotted in figure 12.4.

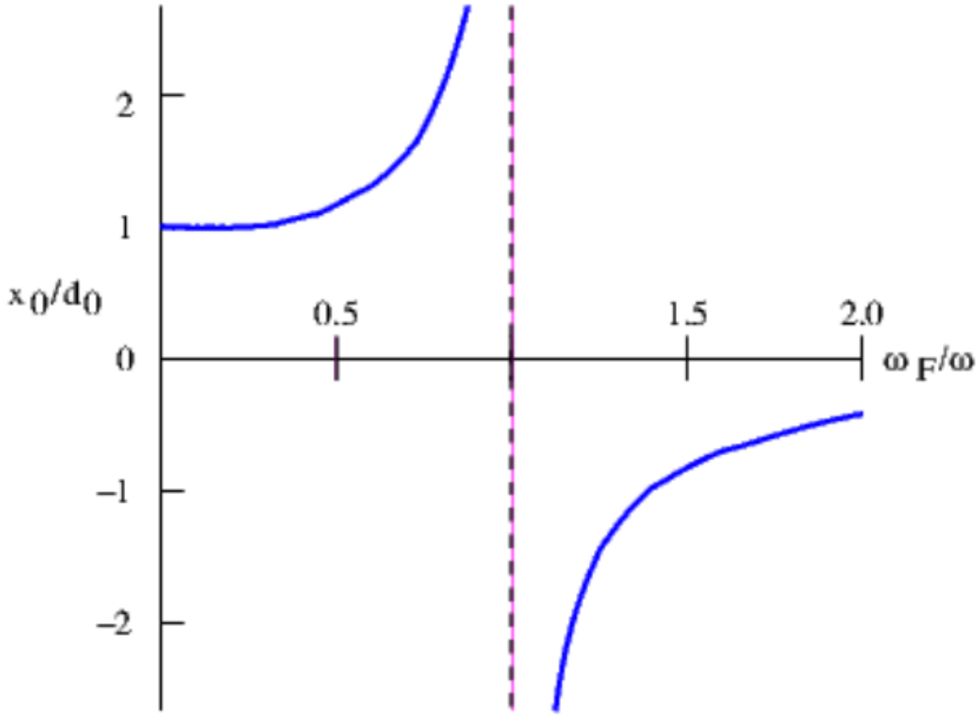


Figure 12.4: Plot of the ratio of response to forcing vs. the ratio of forced to free oscillator frequency for the mass-spring system.

Notice that if $\omega_F < \omega$, the motion of the mass is in phase with the wiggling motion and the amplitude of the mass oscillation is greater than the amplitude of the wiggling.

As the forcing frequency approaches the natural frequency of the oscillator, the response of the mass grows in amplitude.

When the forcing is at the resonant frequency, the response is technically infinite, though practical limits on the amplitude of the oscillation will intervene in this case — for instance, the spring cannot stretch or shrink an infinite amount.

In many cases friction will act to limit the response of the mass to forcing near the resonant frequency.

When the forcing frequency is greater than the natural frequency, the mass actually moves in the opposite direction of the wiggling motion — i. e., the response is out of phase with the forcing.

The amplitude of the response decreases as the forcing frequency increases above the resonant frequency.

Forced and free harmonic oscillators form an important part of many physical systems.

For instance, any elastic material body such as a bridge or an airplane wing has harmonic oscillatory modes.

A common engineering problem is to ensure that such modes are damped by friction or some other physical mechanism when there is a possibility of exictation of these modes by naturally occurring processes.

A number of disasters can be traced to a failure to properly account for oscillatory forcing in engineered structures.

12.4 Quantum Mechanical Harmonic Oscillator

The quantum mechanical harmonic oscillator shares the characteristic of other quantum mechanical bound state problems in that the total energy can take on only discrete values.

Calculation of these values is too difficult for these notes, but the problem is sufficiently important to warrant reporting the results here.

The energies accessible to a quantum mechanical mass-spring system are given by the formula

$$E_n = (n + 1/2)\hbar(k/M)^{1/2}, \qquad n = 0, 1, 2, \ldots . \tag{12.13}$$

In other words, the energy *difference* between successive quantum mechanical energy levels in this case is constant and equals the classical resonant frequency for the oscillator, $\omega = (k/M)^{1/2}$, times $\hbar$.

Chapter 13

Newton's Law of Gravitation

In this chapter we study the law that governs gravitational forces between massive bodies.

We first introduce the law and then explore its consequences.

The notion of a test mass and the gravitational field is developed, followed by the idea of gravitational flux.

We then learn how to compute the gravitational field from more than one mass, and in particular from extended bodies with spherical symmetry.

We finally examine Kepler's laws and learn how these laws and the conservation laws for energy and angular momentum may be used to solve problems in orbital dynamics.

13.1 The Law of Gravitation

Of Newton's accomplishments, the discovery of the universal law of gravitation ranks as one of the greatest.

Imagine two masses, M_1 and M_2, separated by a distance r.

The force has the magnitude

$$F = \frac{M_1 M_2 G}{r^2}, \tag{13.1}$$

where $G = 6.67 \times 10^{-11}$ m^3 kg^{-1} s^{-2} is the *universal gravitational constant*.

The gravitational force is always attractive and it acts along the line of centers between the two masses.

13.2 Gravitational Field

The gravitational field at any point is equal to the gravitational force on some test mass placed at that point divided by the mass of the test mass.

The dimensions of the gravitational field are length over time squared, which is the same as acceleration.

For a single point mass M (other than the test mass), Newton's law of gravitation tells us that

$$\mathbf{g} = -\frac{GM\mathbf{r}}{r^3} \quad \text{(point mass)}, \tag{13.2}$$

where $\mathbf{r}$ is the position of the test point relative to the mass M.

Note that we have written this equation in vector form, reflecting the fact that the gravitational field is a vector.

Thus, $\mathbf{r} = \mathbf{x}_{test} - \mathbf{x}_{mass}$, where $\mathbf{x}_{test}$ and $\mathbf{x}_{mass}$ are the position vectors of the test point and the mass M.

The vector **r** points *from the mass to the test point*.

The quantity $r = |\mathbf{r}|$ is the distance from the mass to the test point.

If there is more than one mass, then the total gravitational field at a test point is obtained by computing the individual fields produced by each mass at the test point and vectorially adding these fields.

This process is schematically illustrated in figure 13.1.

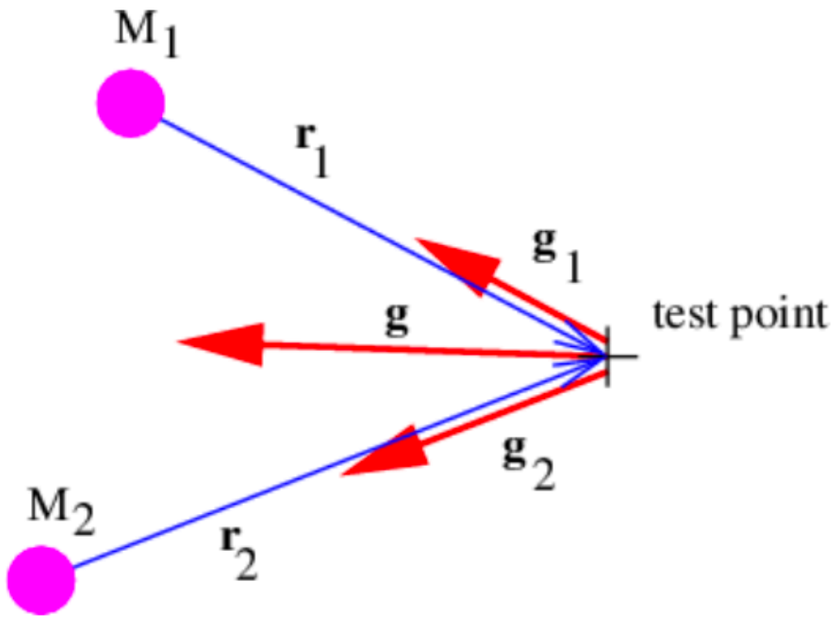


Figure 13.1: Sketch showing the addition of gravitational fields at a test point resulting from two masses.

13.3 Gravitational Flux

The next concept we need to discuss is the *gravitational flux*.

Figure 13.2 shows a rectangular area S with a vector **S** perpendicular to the rectangle.

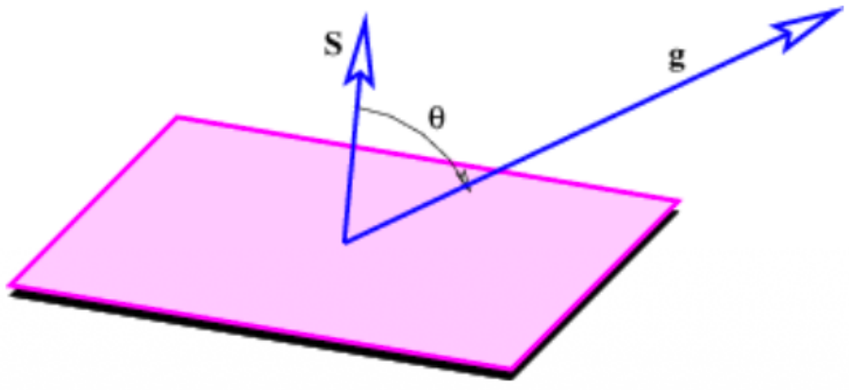


Figure 13.2: Definition sketch for the gravitational flux through the directed area **S**.

The vector **S** is defined to have length S, so it is a compact way of representing the size and orientation of a rectangle in three dimensional space.

The vector **S** could point either upward or downward, and the choice of directions turns out to be important.

This is why we say that **S** represents a *directed area*.

Figure 13.2 also shows a vector **g**, representing the gravitational field on the surface of the rectangle.

Its value is assumed here not to vary with position on the rectangle.

The angle θ is the angle between the vectors **S** and **g**.

The *gravitational flux* through the rectangle is defined as

$$\Phi_g = \mathbf{S} \cdot \mathbf{g} = Sg\cos\theta = Sg_n, \tag{13.3}$$

where $g_n = g \cos \theta$ is the component of **g** normal to the rectangle.

The flux is thus larger for larger areas and for larger gravitational fields.

However, only the component of the gravitational field normal to the rectangle (i. e., parallel to **S**) counts in this calculation.

A consequence is that the gravitational flux through area 1, $\mathbf{S}_1 \cdot \mathbf{g}$, in figure 13.3 is the same as the flux through area 2, $\mathbf{S}_2 \cdot \mathbf{g}$.

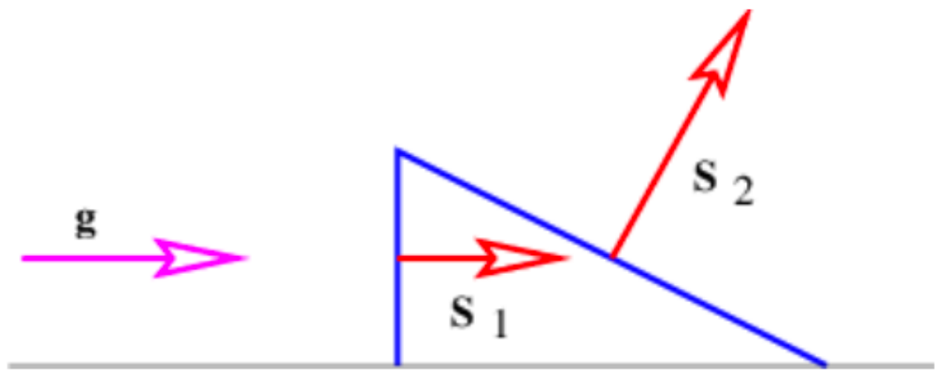


Figure 13.3: Two areas with the same projected area normal to **g**. Is the flux through area 2 greater than, less than, or equal to the flux through area 1? (The two areas are being viewed edge-on and are assumed to have some dimension *d* in the direction normal to the page.)

The significance of the directedness of the area is now clear.

If the vector **S** pointed in the opposite direction, the flux would have the opposite sign.

When defining the flux through a rectangle, it is necessary to define which way the flux is going.

This is what the direction of **S** does — a positive flux is defined as going from the side opposite **S** to the side of **S**.

An analogy with flowing water may be helpful.

Imagine a rectangular channel of cross-sectional area *S* through which water is flowing at velocity *v*.

The flux of water through the channel, which is defined as the volume of water per unit time passing through the cross-sectional area, is $\Phi_w = vS$.

The water velocity takes the place of the gravitational field in this case, and its direction is here assumed to be normal to the rectangular cross-section of the channel.

The *field* thus expresses an *intensity* (e. g., the velocity of the water or the strength of the gravitational field), while the flux expresses an *amount* (the volume of water per unit time in the fluid dynamical case).

The gravitational flux is thus the *amount of some gravitational influence*, while the gravitational field is its strength.

We now try to more clearly understand to what this *amount* really refers.

We need to briefly consider the case in which the gravitational field varies from one point to another on the rectangular surface.

In this case a proper calculation of the flux through the surface cannot be made using equation (13.3) directly.

Instead, we must break the surface into a grid of sub-surfaces.

If the grid is sufficiently fine, the gravitational field will be nearly constant over each sub-surface and equation (13.3) can be applied separately to each of these.

The total flux is then the sum of all the individual fluxes.

There is actually no need for the area in figure 13.2 to be rectangular.

We can calculate the gravitational flux through the surface of a sphere of radius R with a mass M at the center.

As illustrated in figure 13.4, the gravitational field points inward toward the mass.

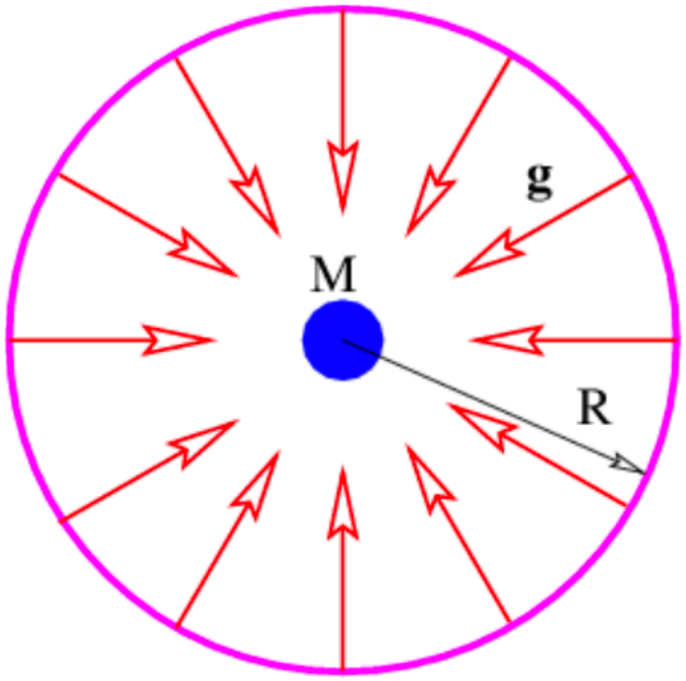


Figure 13.4: Calculation of the gravitational flux through the surface of a sphere with a mass at the center.

It has magnitude $g = GM/R^2$, so if we desire to calculate the gravitational flux *out* of the sphere, we must introduce a minus sign.

Finally, the area of a sphere of radius R is $S = 4\pi R^2$, so the flux is

$$\Phi_g = -gS = -(GM/R^2)(4\pi R^2) = -4\pi GM. \tag{13.4}$$

Notice that this flux doesn't depend on how big the sphere is — the factor of R^2 in the area cancels with the factor of $1/R^2$ in the gravitational field.

This is a hint that something profound is going on.

The size of the surface enclosing the mass is unimportant, and neither is its *shape* — the answer is always the same — the gravitational flux outward through any closed surface surrounding a mass M is just $\Phi_g = -4\pi GM$!

This is an example of *Gauss's law* applied to gravity.

It is possible to formally prove this result using arguments like those posed in figure 13.3, but perhaps the easiest way to understand this result is via the analogy with the flow of water.

If we think of the mass as something which destroys water at a certain rate, then there must be an inward flow of water through the surfaces in the left and center examples in figure 13.5.

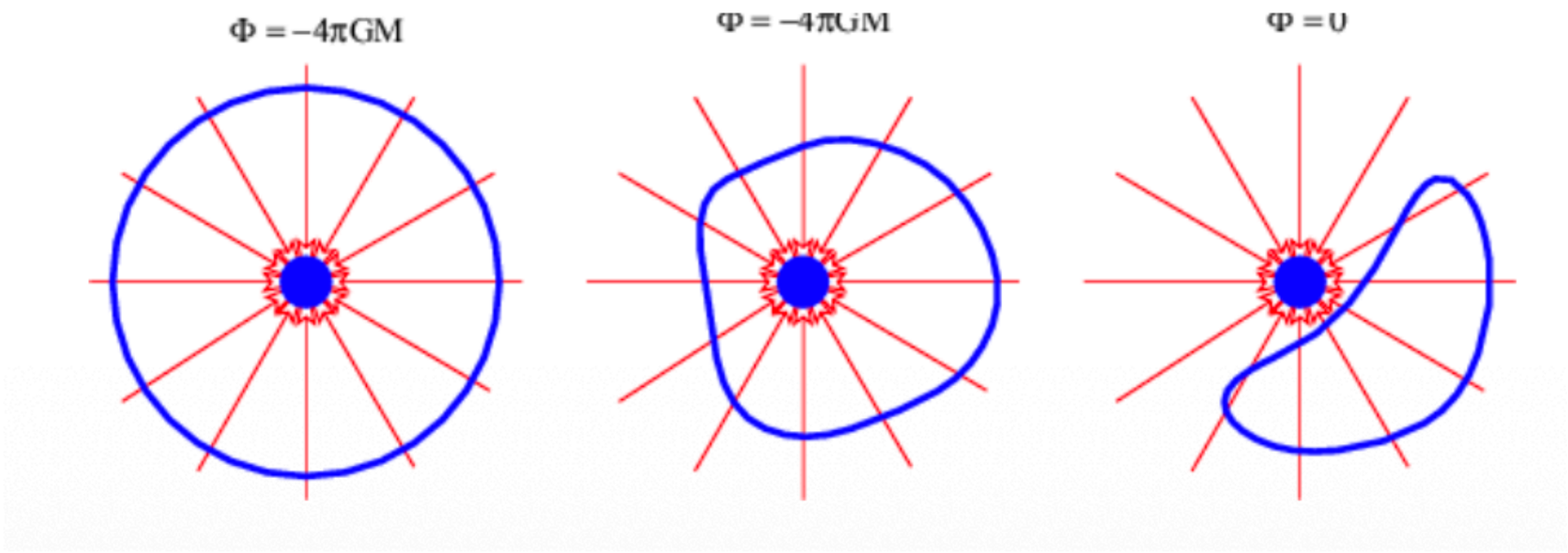


Figure 13.5: Three cases of a mass M and a closed surface. In the left and center examples the mass is inside the closed surface and the outward flux through the surface is $\Phi_g = -4\pi GM$. In the right example the mass is outside the surface and the outward flux through the surface is zero.

Furthermore, the volume of water per unit time flowing inward through these surfaces is the same in the two examples, because the rate at which water is being destroyed is the same.

In the right case the mass is not contained inside the surface and though water flows into the volume bounded by the surface, it also flows out the other side, resulting in a net outward (or inward) volume flux through the surface of zero.

13.4 Flux from Multiple Masses

Gauss's law extends trivially to more than one mass.

As figure 13.6 shows, the outward flux through a closed surface is just

$$\Phi_g = -4\pi G \sum_{inside} M_i \quad \text{(Gauss's law)}. \tag{13.5}$$

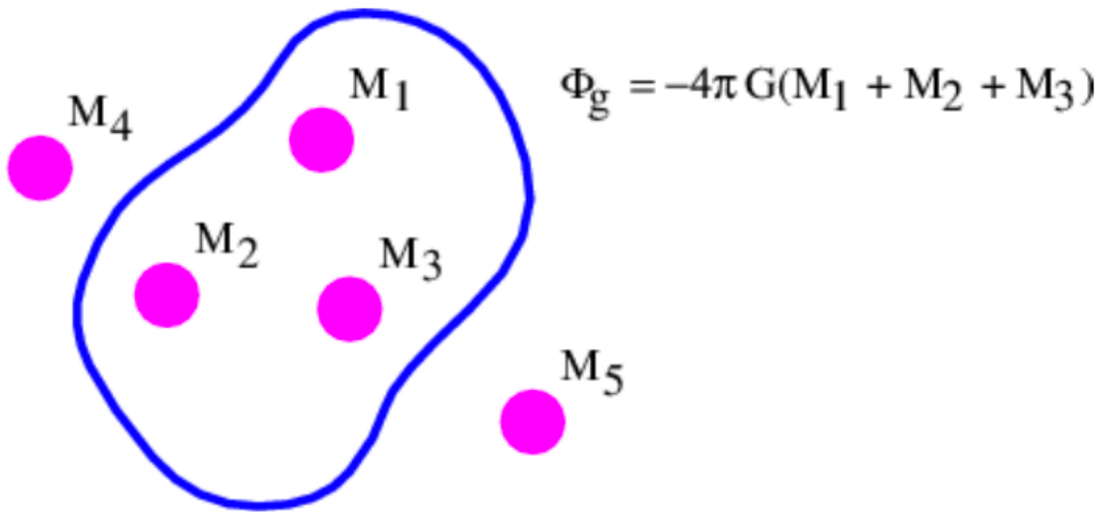


Figure 13.6: Gauss's law applied to more than one mass. The masses M_1, M_2, and M_3 contribute to the outward gravitational flux through the surface shown. The masses M_4 and M_5 don't contribute.

In other words, all masses inside the closed surface contribute to the flux, while no masses outside the surface contribute.

This is the most general statement of Gauss's law as it applies to gravity.

An important application of Gauss's law is to show that the gravitational field outside of a spherically symmetric extended mass M is exactly the same as if all the mass were concentrated at a point at the center of the sphere.

The proof goes as follows: Imagine a sphere concentric with the center of the extended mass, but with larger radius.

The gravitational flux from the mass is just $\Phi = -4\pi GM$ as before

However, because of the assumed spherical symmetry, we know that the gravitational field points normally inward at every point on the spherical surface and is equal in magnitude everywhere on the sphere.

Thus we can infer that $\Phi_g = -4\pi R^2 g$, where R is the radius of the sphere and g is the magnitude of the gravitational field at radius R.

From these two equations we immediately infer that the field magnitude is

$$g = \frac{GM}{R^2}. \tag{13.6}$$

Expressing this in vector form for arbitrary radius r, and remembering that the gravitational field points inward, we find that

$$\mathbf{g} = -\frac{GM\mathbf{r}}{r^3}, \tag{13.7}$$

which is precisely the equation for **g** resulting from a point mass M.

Recall that **r** points from the mass to the test point.

13.5 Effects of Relativity

So far our discussion of gravity has been completely non-relativistic.

We will not explore in detail how the theory of gravity changes in a completely relativistic treatment.

As we noted earlier in the course, Einstein's general theory of relativity covers this, and the mathematics are formidable.

We confine ourselves to two comments:

- As noted previously, gravity is locally equivalent to being in an accelerated reference frame. However, unlike the simple example which we studied earlier, there is in general no universal frame of reference that is everywhere inertial to which we can transform.

- Space is even more non-Euclidean in general relativity than in special relativity. In particular, there is no such thing as a straight line in the geometry of general relativistic spacetime. This is true because spacetime itself is curved. An example of a curved space is the surface of a sphere. Clearly, a straight line cannot be embedded in this space. The closest equivalent to a straight line in a curved space is a *geodesic curve*. On a sphere great circles are geodesic curves. In general relativity, objects subject only to the force of gravity move along geodesic curves.

One potentially observable prediction of relativity is the existence of gravitational waves.

Imagine two stars revolving around each other.

The gravitational field from these stars will change periodically due to this motion.

However, this change propagates outward only at the speed of light.

As a result, ripples in the field, or gravitational waves, spread outward from the revolving stars.

Efforts are currently under way to detect gravitational waves produced by violent cosmic events such as the explosion of a supernova or the merging of black holes.

LIGO experiment did so in 2015 and it continues.

13.6 Kepler's Laws

Johannes Kepler, using data compiled by Tycho Brahe, inferred three laws governing the motions of planets in the solar system:

1. Planets move in elliptical orbits with the sun at one focus.
2. Equal areas are swept out in equal times by the line connecting the sun and the planet.
3. The square of the period of revolution of the planet around the sun is proportional to the cube of the semi-major axis of the ellipse.

These laws were instrumental in the development of modern mechanics and the universal law of gravitation by Isaac Newton.

Showing that the first law is consistent with Newtonian mechanics is mathematically more difficult than we can undertake in this course.

However, the second law turns out to be a simple consequence of the conservation of angular momentum.

Figure 13.7 shows an elliptical orbit with the area swept out as a planet moves from position 1 to position 2.

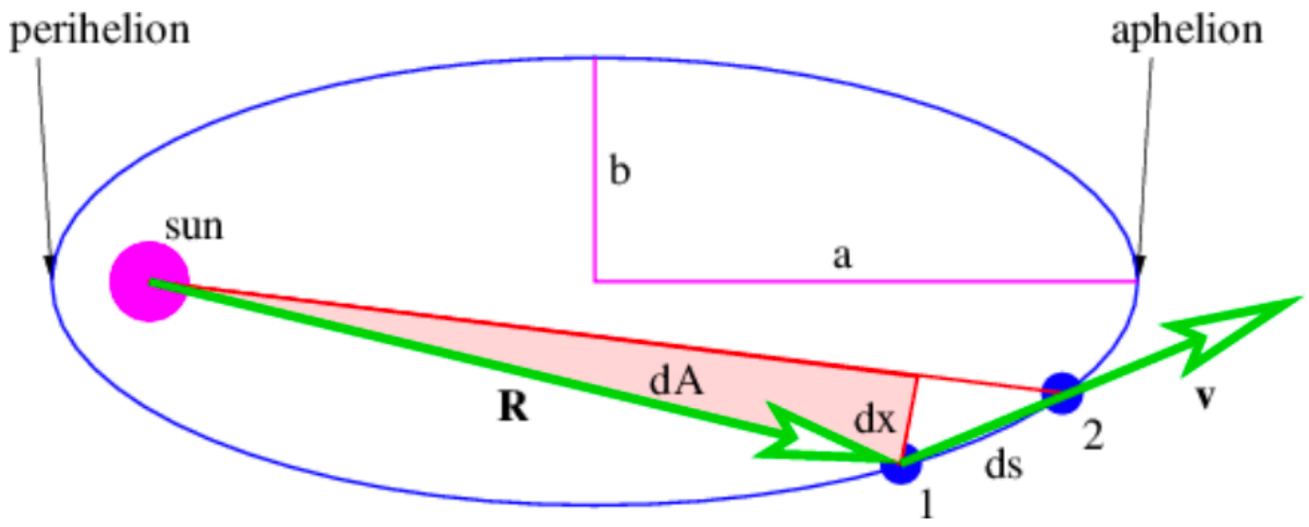


Figure 13.7: Illustration of elliptical orbit of a planet with the sun at the left focus. The *semi-major* and *semi-minor axes* are denoted by a and b. The shaded triangular area element is needed for the discussion of Kepler's second law. *Perihelion* and *aphelion* are respectively the points on the orbit nearest and farthest from the sun. Note that at perihelion and aphelion the velocity is purely tangential, i. e., the velocity component along the radius vector is zero.

We estimate this area as $dA = Rdx/2$, where we have ignored the small unshaded part of the area to the right of the shaded triangle.

The distance traveled by the planet in time dt is ds, so the magnitude of the velocity is $v = ds/dt$.

However, in computing the angular momentum, we need the tangential component of the velocity, i. e., the component normal to the radius vector $\mathbf{R}$.

This is simply $v_t = dx/dt$.

The angular momentum is $L = mRv_t = mRdx/dt$, where m is the mass of the planet.

Combining this with the formula for dA results in

$$\frac{dA}{dt} = \frac{L}{2m}. \tag{13.8}$$

Since gravitation is a central force, angular momentum is conserved, which means that dA/dt is constant.

Thus, we have shown that conservation of angular momentum is equivalent to Kepler's second law.

Kepler's third law turns out to be a consequence of the universal law of gravitation.

We can prove this for circular orbits.

We know that a planet moving in a circular orbit around the sun is accelerating toward the sun with the centripetal acceleration $a = v^2/R$, where v is the speed of the planet's motion in its orbit and R is the orbit's radius.

This acceleration is caused by the gravitational force, so we can equate the force divided by the planetary mass to a, resulting in

$$\frac{v^2}{R} = \frac{GM}{R^2}, \tag{13.9}$$

where M is the mass of the sun.

This may be solved for v:

$$v = \left(\frac{GM}{R}\right)^{1/2}. \tag{13.10}$$

Eliminating v in favor of the period of revolution $T = 2\pi R/v$ results in

$$T^2 = \frac{4\pi^2 R^3}{GM}. \tag{13.11}$$

This agrees with Kepler's third law since the semi-major axis of a circle is simply the radius R.

13.7 Use of Conservation Laws

The gravitational force is conservative, so two point masses M and m separated by a distance r have a potential energy:

$$U = -\frac{GMm}{r}. \tag{13.12}$$

It is easily verified that differentiation recovers the gravitational force.

The conservation of energy and angular momentum in planetary motions can be used to solve many practical problems involving motion under the influence of gravity.

For instance, suppose a bullet is shot straight upward from the surface of the moon.

One might ask what initial velocity is needed to insure that the bullet will escape from the gravity of the moon.

Since total energy E is conserved, the sum of the initial kinetic and potential energies must equal the sum of the final kinetic and potential energies:

$$E = K_{initial} + U_{initial} = K_{final} + U_{final}\,. \tag{13.13}$$

For the bullet to escape the moon, its kinetic energy must remain positive no matter how far it gets from the moon.

Since the potential energy is always negative, asymptoting to zero at infinite distance (i. e., $U_{final} = 0$), the minimum total energy consistent with this condition is zero.

For zero total energy we have

$$\frac{mv_{initial}^2}{2} = K_{initial} = -U_{initial} = +\frac{GMm}{R}, \tag{13.14}$$

where m is the mass of the bullet, M is the mass of the moon, R is the radius of the moon, and $v_{initial}$ is the minimum initial velocity required for the bullet to escape.

Solving for $v_{initial}$ yields

$$v_{initial} = \left(\frac{2GM}{R}\right)^{1/2}\,. \tag{13.15}$$

This is called the *escape velocity*.

Notice that the escape velocity from a given radius is a factor of $2^{1/2}$ larger than the velocity needed for a circular orbit at that radius (see equation (13.10)).

An object is energetically bound to the sun if its kinetic plus potential energy is less than zero.

In this case the object follows an elliptical orbit around the sun as shown by Kepler.

However, if the kinetic plus potential energy is zero, the object follows a parabolic orbit, and if it is greater than zero, a hyperbolic orbit results.

In the latter two cases the sun also resides at a focus of the parabola or hyperbola.

Figure 13.8 shows a typical hyperbolic orbit.

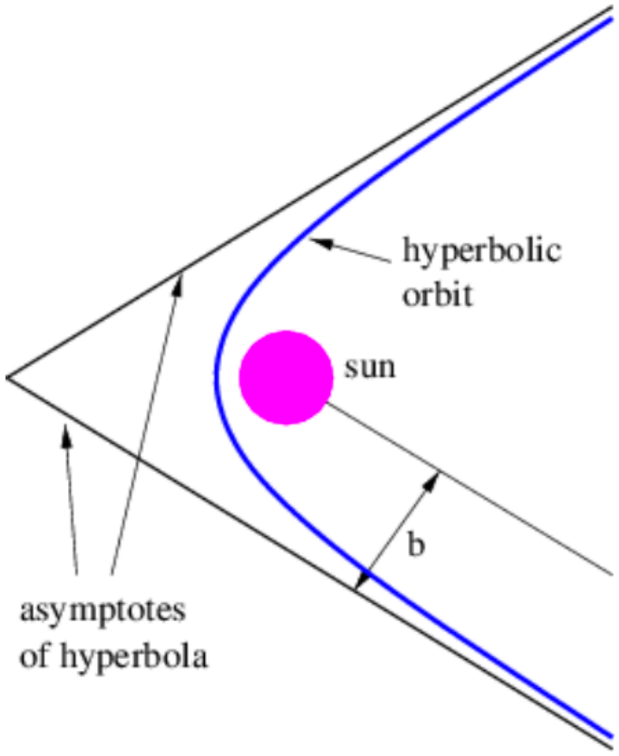


Figure 13.8: Example of a hyperbolic orbit of a positive energy object passing by the sun. The sun sits at the focus of the hyperbola. The quantity *b* is called the *impact parameter*.

The impact parameter, defined in this figure, is the closest the object would have come to the center of the sun if it hadn't been deflected by gravity.

Sometimes energy and angular momentum conservation can be used together to solve problems.

For instance, suppose we know the energy and angular momentum of an asteroid of mass m and we wish to infer the maximum and minimum distances of the asteroid from the sun, the so-called aphelion and perihelion distances.

Since the asteroid is gravitationally bound to the sun, it is convenient to characterize the total energy by $E_b = -E$, the so-called *binding energy*.

If v is the orbital speed of the asteroid and r is its distance from the sun, then the binding energy can be written in terms of the kinetic and potential energies:

$$-E_b = \frac{mv^2}{2} - \frac{GMm}{r}. \tag{13.16}$$

The magnitude of the angular momentum of the asteroid is $L = mv_t r$, where v_t is the tangential component of the asteroid's velocity.

At aphelion and perihelion, the radial part of the velocity of the asteroid is zero and the speed equals the tangential component of the velocity, $v = v_t$.

Thus, *at aphelion and perihelion* we can eliminate v in favor of the angular momentum:

$$-E_b = \frac{L^2}{2mr^2} - \frac{GMm}{r} \quad \text{(aphelion and perihelion)}. \tag{3.17}$$

This can be rearranged into a quadratic equation

$$r^2 - \frac{GMm}{E_b} r + \frac{L^2}{2mE_b} = 0, \tag{3.18}$$

which can be solved to yield

$$r = \frac{1}{2}\left[\frac{GMm}{E_b} \pm \left(\frac{G^2M^2m^2}{E_b^2} - \frac{2L^2}{mE_b}\right)^{1/2}\right]. \tag{13.19}$$

The larger of the two solutions yields the aphelion value of the radius while the smaller yields the perihelion.

Equation (13.19) tells us something else interesting.

The quantity inside the square root cannot be negative, which means that we must have

$$L^2 \leq \frac{G^2M^2m^3}{2E_b}. \tag{13.20}$$

In other words, for a given value of the binding energy E_b there is a maximum value for the angular momentum.

This maximum value makes the square root zero, which means that the aphelion and the perihelion are the same — i. e., the orbit is circular.

Thus, among all orbits with a given binding energy, the circular orbit has the maximum angular momentum.

Chapter 14

Forces in Relativity

In this chapter we ask an apparently simple question: How can the idea of potential energy be extended to the relativistic case?

The answer to this question is unexpectedly complex, but it leads us to immensely fruitful results.

In particular, it prompts us to investigate the idea of potential momentum, which results ultimately in *gauge theory*, of which electromagnetism is an example.

Along the way we show that conservation of four-momentum has an unexpected consequence — the idea of force at a distance is inconsistent with the theory of relativity.

This means that momentum and energy must be carried between interacting particles by another type of particle that we call an *intermediary particle*.

These particles are *virtual* in the sense that they don't have their real-world mass when acting in this role.

In relativistic quantum mechanics, we find that particles can take on negative energies.

Feynman's interpretation of this fact is discussed, which leads us to a model for antiparticles.

14.1 Potential Momentum

For a free, non-relativistic particle of mass m, the total energy E equals the kinetic energy K and is related to the momentum $\mathbf{\Pi}$ of the particle by

$$E = K = \frac{|\mathbf{\Pi}|^2}{2m} \quad \text{(free, non-relativistic)}. \tag{14.1}$$

Note that we have ignored the contribution of the rest energy to the total energy here.

In the non-relativistic case, the momentum is $\mathbf{\Pi} = m\mathbf{v}$, where $\mathbf{v}$ is the particle velocity.

If the particle is not free, but is subject to forces associated with a potential energy $U(x,y,z)$, then equation (14.1) must be modified to account for the contribution of U to the total energy:

$$E - U = K = \frac{|\mathbf{\Pi}|^2}{2m} \quad \text{(non-free, non-relativistic)}. \tag{14.2}$$

The force on the particle is related to the potential energy by

$$\mathbf{F} = -\left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z}\right). \tag{14.3}$$

For a free, relativistic particle, we have

$$E = (|\mathbf{\Pi}|^2 c^2 + m^2 c^4)^{1/2} \quad \text{(free, relativistic)}. \tag{14.4}$$

The obvious way to add forces to the relativistic case is by rewriting equation (14.4) with a potential energy, in analogy with equation (14.2):

$$E - U = (|\mathbf{\Pi}|^2 c^2 + m^2 c^4)^{1/2} \qquad \text{(incomplete!)}. \tag{14.5}$$

Unfortunately, equation (14.5) is incomplete, because we have subtracted U from the energy E without subtracting a corresponding term from the momentum $\mathbf{\Pi}$ as well.

However, $\Pi = (\mathbf{\Pi}, E/c)$ is a four-vector, so an equation with something subtracted from just one of the components of this four-vector is not relativistically invariant.

In other words, equation (14.5) doesn't obey the principle of relativity, and therefore cannot be correct!

How can we fix this problem?

One way is to define a new four-vector with U/c being its timelike part and some new vector $\mathbf{Q}$ being its spacelike part:

$$\underline{Q} \equiv (\mathbf{Q}, U/c) \quad \text{(potential four-momentum)}. \tag{14.6}$$

We then subtract $\mathbf{Q}$ from the momentum $\mathbf{\Pi}$.

When we do this, equation (14.5) becomes

$$E - U = (|\mathbf{\Pi} - \mathbf{Q}|^2 c^2 + m^2 c^4)^{1/2} \quad \text{(non-free, relativistic)}. \tag{14.7}$$

The quantity $\mathbf{Q}$ is called the potential momentum and Q is the potential four-momentum.

Some additional terminology is useful.

We define

$$\mathbf{p} \equiv \mathbf{\Pi} - \mathbf{Q} \quad \text{(kinetic momentum)} \tag{14.8}$$

as the *kinetic momentum* for reasons discussed below. I

n order to avoid confusion, we rename $\mathbf{\Pi}$ the *total* momentum.

Thus, the total momentum equals the kinetic plus the potential momentum, in analogy with energy.

So far, we have shown that the introduction of a potential momentum complements the potential energy so as to make the energy-momentum relationship for a particle relativistically invariant.

However, we as yet have no idea what causes potential momentum nor what it does to the affected particle.

We shall put off answering the former question and address only the latter at this point.

A hint comes from the corresponding behavior of energy.

The *total energy* of a particle is related to the quantum mechanical frequency ω of the particle, and the *total momentum* is related to its wave vector $\mathbf{k}$:

$$E = \hbar\omega \qquad \mathbf{\Pi} = \hbar\mathbf{k}. \tag{14.9}$$

However, the *kinetic energy* and the *kinetic momentum* are related to the particle's velocity **v**:

$$E - U = \frac{mc^2}{(1 - v^2/c^2)^{1/2}} \qquad \mathbf{p} = \mathbf{\Pi} - \mathbf{Q} = \frac{m\mathbf{v}}{(1 - v^2/c^2)^{1/2}}, \tag{14.10}$$

where $v = |\mathbf{v}|$.

The relationship between kinetic momentum and velocity can be proven by dividing equation (14.7) by $\hbar$ to obtain a dispersion relation and then computing the group velocity, which we equate to the particle velocity.

However, we will not do this here.

14.2 Aharonov-Bohm Effect

Let us now study a phenomenon that depends on the existence of potential momentum.

If the potential energy of a particle is zero and both the kinetic and potential momenta point in the $\pm x$ direction, the total energy equation (14.7) for the particle becomes

$$E = [(\Pi - Q)^2 c^2 + m^2 c^4]^{1/2} = (p^2 c^2 + m^2 c^4)^{1/2}. \tag{14.11}$$

Since its total energy E is conserved, the magnitude of the kinetic momentum p of the particle doesn't change according to the above equation.

Thus, if a region of non-zero potential momentum is encountered, the total momentum of the particle must change so as to keep the kinetic momentum constant.

This results in a change in the wavelength of the matter wave associated with the particle.

In particular, if the potential momentum points in the same direction as the kinetic momentum, the total momentum is increased and the wavelength decreases, while a potential momentum pointing in the direction opposite the kinetic momentum results in an increase in wavelength.

Figure 14.1 illustrates what might happen to a particle moving through a channel that splits into two sub-channels for an interval.

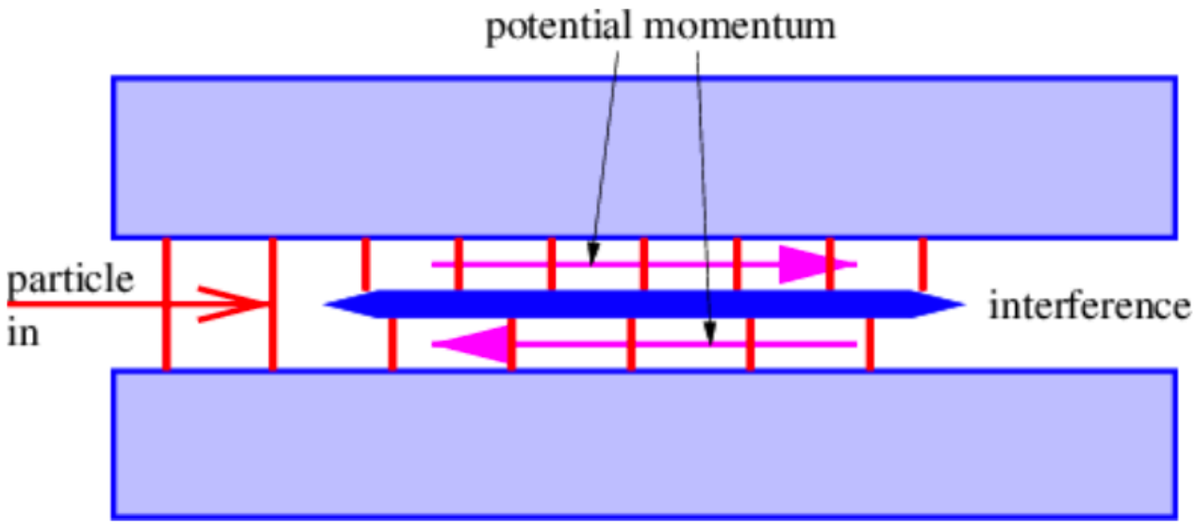


Figure 14.1: Setup for the Aharonov-Bohm effect. The particle moves through a channel which has a divided segment with non-zero potential momenta pointing in opposite directions in the two sub-channels. The vertical line segments show the wave fronts for the particle.

If we arrange to have non-zero potential momenta pointing in opposite directions in the sub-channels, the wavelength of the particle will be different in the two regions.

At the end of the interval, the waves recombine, interfering constructively or destructively, depending on the magnitude of the phase difference between them.

If destructive interference occurs, then the particle cannot pass.

The potential momentum thus acts as a valve controlling the flow of particles through the channel.

This is an example of the *Aharonov-Bohm effect*.

14.3 Forces from Potential Momentum

In the Aharonov-Bohm effect, the potential momentum didn't result in any force on the particle — its only manifestation was to change the particle's wavelength.

In such situations the potential momentum's presence is only revealed by quantum mechanical effects.

The potential momentum has more of an influence on the non-quantum world when the problem is two or three-dimensional or when the potential momentum is changing with time.

The total force on a particle due to all possible effects involving the potential energy and the potential momentum is given by

$$\mathbf{F} = -\left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z}\right) - \frac{\partial \mathbf{Q}}{\partial t} + \mathbf{v} \times \mathbf{P}, \tag{14.12}$$

where $\mathbf{v}$ is the particle velocity and $\mathbf{P}$ is a vector obtained from the potential momentum vector as follows:

$$\mathbf{P} \equiv \left(\frac{\partial Q_z}{\partial y} - \frac{\partial Q_y}{\partial z}, \frac{\partial Q_x}{\partial z} - \frac{\partial Q_z}{\partial x}, \frac{\partial Q_y}{\partial x} - \frac{\partial Q_x}{\partial y}\right). \tag{14.13}$$

This is unexpectedly complicated.

However, equation (14.12) consists of three parts.

The first part involves derivatives of the potential energy and is exactly the same as in the non-relativistic case.

The new effects are confined to the second and third parts, $-\partial \mathbf{Q}/\partial t$ and $\mathbf{v} \times \mathbf{P}$.

A full derivation of these equations involves rather complex mathematics.

However, it is possible to understand the origin of these additional contributions to the force by looking at a couple of simple examples.

14.3.1 Refraction Effect

A matter wave impingent on a discontinuity in potential momentum is refracted, just as it is refracted by a discontinuity in potential energy.

Refraction of a matter wave packet means that the velocity of the associated particle changes as it moves across the interface.

This means that the particle undergoes an acceleration, implying that it is subject to a force.

As in the case of Snell's law for optics, the frequency of a matter wave doesn't change as it crosses such a discontinuity in potential momentum.

Furthermore, neither does the component of the wave vector parallel to the discontinuity.

These two conditions together ensure phase continuity at the interface.

Figure 14.2 shows an example of what happens when a wave encounters a series of parallel slabs with increasing values of **Q**.

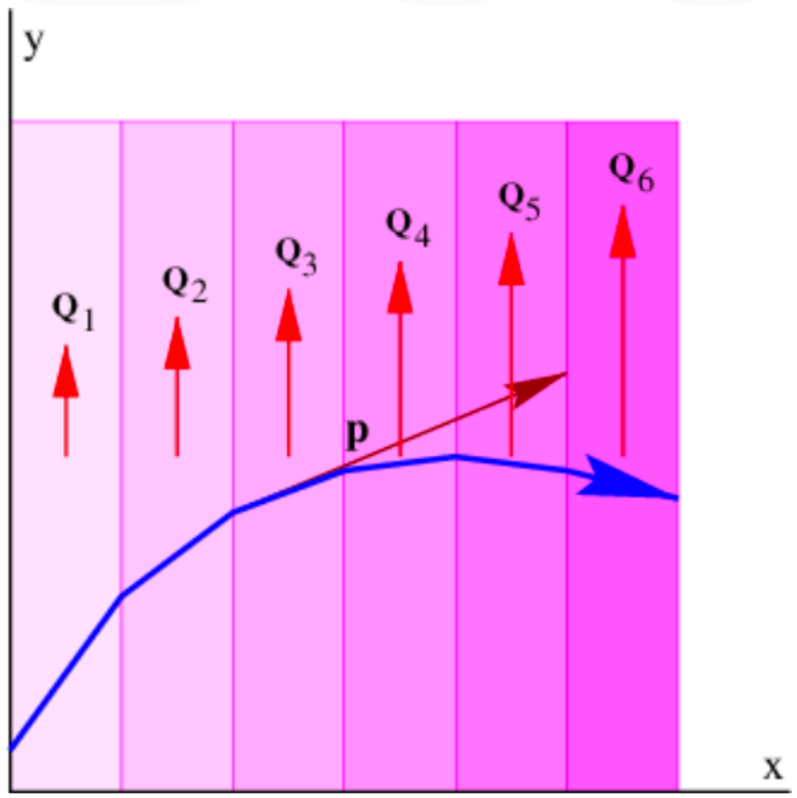


Figure 14.2: Trajectory of a wave packet through a region of variable potential momentum **Q**. **Q** points in the *y* direction and increases in magnitude by steps with increasing *x*. The kinetic momentum **p** indicates the direction of motion of the associated particle at each point along the trajectory.

The *y* component of the wave vector doesn't change as the wave crosses each of the interfaces between slabs, for reasons discussed above.

Hence, $\Pi_y = \hbar k_y$ doesn't change either, which means that $d\Pi_y/dx = 0$.

The *y* component of kinetic momentum, $p_y = \Pi_y - Q_y$, must therefore decrease as Q_y increases, as illustrated in figure 14.2.

Newton's second law tells us that the *y* component of the force on the particle associated with the wave is just the time derivative of the *y* component of the kinetic momentum:

$$F_y = \frac{dp_y}{dt} = \frac{dp_y}{dx}\frac{dx}{dt} = \frac{dp_y}{dx}u_x = -\frac{dQ_y}{dx}u_x. \qquad (14.14)$$

In the last step of this equation we used the fact that $d\Pi_y/dx = 0$.

The x component of the force can be obtained by similar reasoning, using the additional information that the speed, and hence the magnitude of the kinetic momentum, $p^2 = p_x^2 + p_y^2$, doesn't change under he influence of the potential momentum:

$$F_x = \frac{dp_x}{dt} = \frac{dp_x}{dx}\frac{dx}{dt} = -\frac{p_y}{(p^2 - p_y^2)^{1/2}}\frac{dp_y}{dx}u_x = \frac{p_y}{p_x}\frac{dQ_y}{dx}u_x = \frac{dQ_y}{dx}u_y. \qquad (14.15)$$

Aside from assuming that $p^2 = constant$, we have used the relationships $p_x = (p^2 - p_y^2)^{1/2}$ and $p_y/p_x = u_y/u_x$.

Equations (14.14) and (14.15) constitute a special case of equations (14.12) and (14.13) which is valid when **Q** points in the y direction and is a function only of x.

14.3.2 Time-Varying Potential Momentum

The necessity for the term $-\partial\mathbf{Q}/\partial t$ in equation (14.12) is easily understood from the following argument, which is illustrated in figure 14.3.

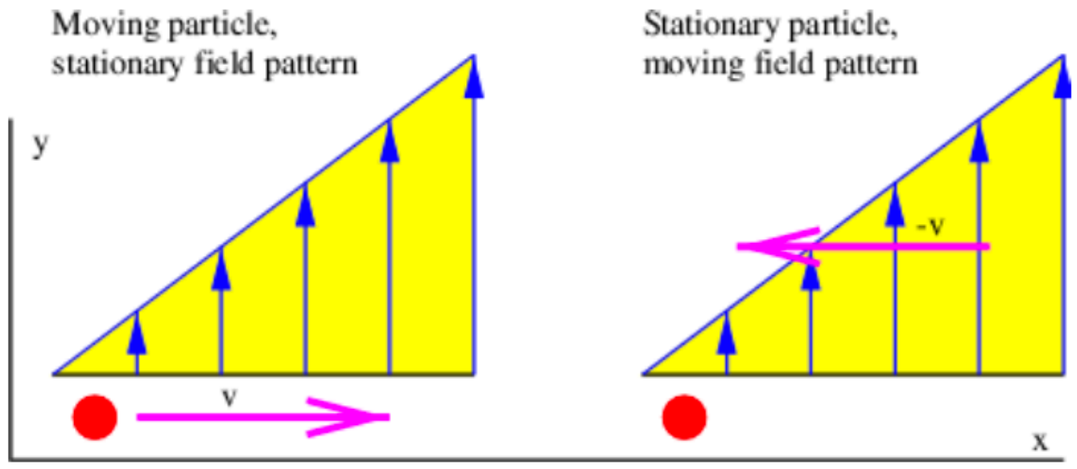


Figure 14.3: A moving particle and a stationary pattern of potential momentum **Q** must be equivalent to a stationary particle and a moving pattern of potential momentum according to the principle of relativity.

The example in the previous section showed that a particle moving in the $+x$ direction with velocity u_x through a field of increasing Q_y (left panel of figure 14.3) experiences a force in the $-y$ direction equal to $F_y = -(dQ_y/dx)u_x$.

However, viewing this same process from a reference frame in which the particle is stationary (right panel of this figure), we see that the potential momentum at the position of the particle increases with time at the rate dQ_y/dt.

The particle is not moving in this reference frame, so the term $\mathbf{v} \times \mathbf{P} = 0$.

However, the stationary particle must still experience the above force in this reference frame in order to satisfy the principle of relativity.

Noting that $dQ_y/dt = (dQ_y/dx)u_x$, we see that equation (14.12) provides this force via the term $-\partial\mathbf{Q}/\partial t$ in the reference frame moving with the particle.

Thus, the time derivative term in equation (14.12) is needed to maintain the principle of relativity.

The same force occurs in the two different reference frames but originates from the term $\mathbf{v} \times \mathbf{P}$ in the original reference frame and the term $-\partial\mathbf{Q}/\partial t$ in the frame moving with the particle.

Arguments similar to these were actually made by Einstein in his original 1905 paper on relativity.

14.4 Lorenz Condition

It turns out that the four components of the potential four-momentum are not independent, but are subject to the condition

$$\frac{\partial Q_x}{\partial x} + \frac{\partial Q_y}{\partial y} + \frac{\partial Q_z}{\partial z} + \frac{1}{c^2}\frac{\partial U}{\partial t} = 0.$$

This is called the *Lorenz condition*.

The physical meaning of this condition will become clear when we study electromagnetism.

14.5 Gauge Theories and Other Theories

The theory of potential momentum is only one of three ways in which the idea of potential energy can be extended to the relativistic case.

This theory is called *gauge theory* for obscure historical reasons.

Gauge theory is important because electromagnetism as well as the theories of weak and strong sub-nuclear interactions are all of this type.

Gravity is the only fundamental force that does not take the form of a gauge theory.

Instead, gravity takes the form of one of two other possible relativistic extensions of potential energy.

This theory is called general relativity.

The gravitational force in general relativity can be interpreted geometrically as a consequence of the curvature of spacetime.

Mathematically, it is far too difficult to pursue here.

The third relativistic extension of potential energy considers potential energy to be a field which alters the rest energy of particles.

High energy physicists believe that the elementary particles gain their mass by this mechanism.

The field is called the *Higgs field* after the English physicist who first proposed this theory, Peter Higgs.

This theory has now been(2012) experimentally verified.

14.6 Conservation of Four-Momentum

We earlier introduced the ideas of energy and momentum conservation.

In other words, if we have a number of particles isolated from the rest of the universe, each with momentum $\mathbf{p}_i$ and energy E_i, then particles may be created and destroyed and they may collide with each other.

In these interactions the energy and momentum of each particle may change, but the sum total of all the energy and the sum total of all the momentum remains constant with time:

$$E = \sum_i E_i = const \qquad \mathbf{p} = \sum_i \mathbf{p}_i = const. \tag{14.17}$$

The expression is simpler in terms of four-momentum:

$$\underline{p} = \sum_i \underline{p}_i = const. \tag{14.18}$$

At this point a statement such as the one above should ring alarm bells.

Just what does it mean to say that the total energy and momentum remain constant with time in the context of relativity?

Which time?

The time in which *reference frame*?

Figure 14.4 illustrates the problem.

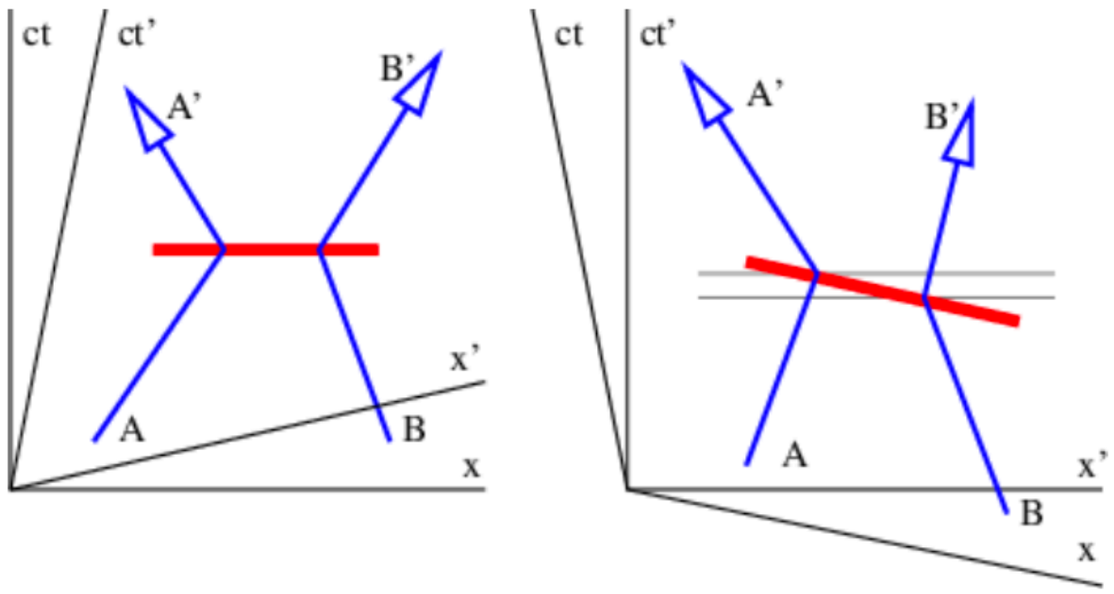


Figure 14.4: The trouble with action at a distance. View of the remote exchange of four-momentum from the point of view of two different coordinate systems. The fat line in both pictures is the line of simultaneity in the unprimed frame which is coincident with the exchange of four-momentum between the two particles.

Suppose two particles exchange four-momentum remotely at the time indicated by the fat horizontal bar in the left panel of figure 14.4.

Conservation of four-momentum implies that

$$\underline{p}_A + \underline{p}_B = \underline{p}'_A + \underline{p}'_B, \tag{14.19}$$

where the subscripted letters correspond to the particle labels in figure 14.4.

Primed values refer to the momentum after the exchange while no primes indicates values before the exchange.

Now view the exchange from the reference frame in the right panel of figure 14.4.

A problem with four-momentum conservation exists in the region between the thin horizontal lines.

In this region particle B has already transferred its four-momentum, but it has yet to be received by particle A.

In other words, four-momentum is *not* conserved in this reference frame!

This problem is so serious that we must eliminate the concept of force at a distance from the repertoire of physics.

The only way to have particles interact remotely and still conserve four-momentum in all reference frames is to assume that all remote interactions are mediated by *another particle*, as indicated in figure 14.5.

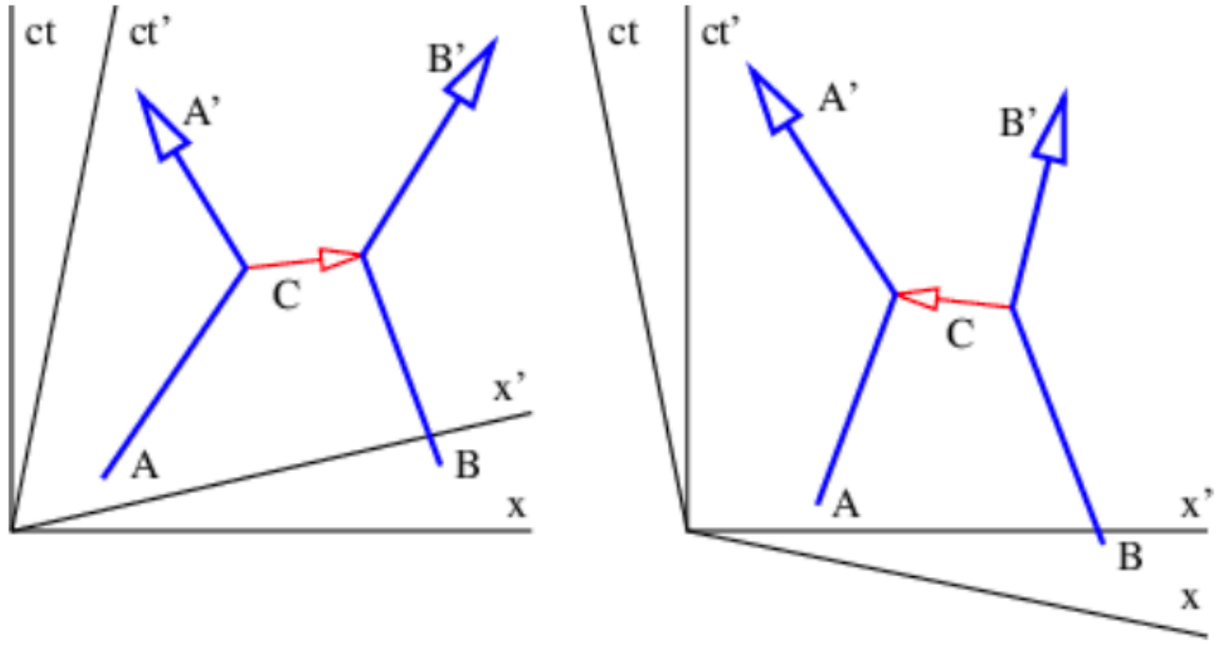


Figure 14.5: Mediation of action at a distance by a third particle. Notice that the world line of the intermediary particle has a slope less than unity, which means that it is nominally moving faster than the speed of light.

In other words, momentum and energy are transferred from particle A to particle B in a two step process.

First, particle A emits particle C in a manner which conserves the four-momentum.

Second, particle C is absorbed by particle B in a similarly conservative interaction.

Four-momentum is conserved at all times in all reference frames in this picture.

14.7 Virtual Particles

Another problem is evident from figure 14.5.

As drawn, the velocity of the intermediary particle exceeds the speed of light.

This is reflected in the fact that different reference frames yield contradictory results as to whether the intermediary particle moves from A to B or B to A.

These difficulties turn out to be much less severe than those arising from non-locality. Let us address them in sequence.

For sake of definiteness, let us view the emission of particle C by particle A in a reference frame in which the velocity of particle A is just reversed in the emission process.

In this case the four-momentum before the emission is $\underline{p}_A = (p, E/c)$, where $E = (p^2c^2 + m^2c^4)^{1/2}$.

After the emission we have $\underline{p}'_A = (-p, E/c)$.

Conservation of four-momentum in the emission process requires that

$$\underline{p}_A = \underline{p}'_A + \underline{q} \tag{14.20}$$

where q is the four-momentum of particle C.

From the above assumptions it is clear that

$$\underline{q} = \underline{p}_A - \underline{p}'_A = (p + p, E/c - E/c) = (2p, 0). \tag{14.21}$$

Suppose that the real, measured mass of particle C is m_C.

This conflicts with the apparent or virtual mass of this particle in its flight from A to B, which is

$$m' = (-\underline{q} \cdot \underline{q})^{1/2}/c \equiv iq/c, \tag{14.22}$$

where $q \equiv |\underline{q}|$ is the *momentum transfer*.

Note that the apparent mass is imaginary because the four-momentum is spacelike.

Classically, this discrepancy in the apparent and actual masses of the particle C would simply indicate that the process wasn't possible.

However, recall that the uncertainty principle allows there to be an uncertainty in the mass if it doesn't persist for too long in terms of the proper time interval along the particle's world line.

The statement of this law is $\Delta\mu\Delta\tau \approx 1$.

Expressed in terms of mass, this becomes

$$\Delta m \Delta \tau \approx \hbar / c^2. \tag{14.23}$$

Let us convert the proper time to an interval since the world line of particle C is horizontal in the reference frame in which we are viewing it.

Ignoring the factor of i, $\Delta\tau = \Delta I / c$.

We finally compute the absolute value of the mass discrepancy as follows:
$|m_C - iq/c| = [(m_C - iq/c)(m_C + iq/c)]^{1/2} = (m_C^2 + q^2/c^2)^{1/2}$.

Solving for I yields the approximate maximum invariant interval that particle C can move from its source point while keeping its erroneous mass hidden by the uncertainty principle:

$$\Delta I \approx \frac{\hbar}{(m_C^2 c^2 + q^2)^{1/2}}. \tag{14.24}$$

A particle forced into having an apparent mass different from its actual mass is called a *virtual particle*.

The interaction shown in figure 14.5 can only take place if particles A and B come closer to each other than the distance ΔI.

This argument thus produces an estimate for the "range" of an interaction with momentum transfer $2p$ and intermediary particle mass m_C.

Two distinct possibilities exist.

If the intermediary particle is massless (a photon, for instance), then the range of the interaction is inversely related to the momentum transfer: $\Delta I \approx \hbar/q$.

Thus, small momentum transfers can occur at large distances.

An interaction of this type is called "long range".

On the other hand, if the intermediary particle has mass, the range is simply $\Delta I \approx \hbar/m_C c$ when $q \ll m_C c$.

The range is thus constant and inversely proportional to the mass of the intermediary particle for low momentum transfers.

For large momentum transfer, i. e., when $q \gg m_C c$ the range decreases from this value with increasing momentum transfer, as in the case of a massless intermediary particle.

14.7.1 Virtual Particles and Gauge Theory

According to quantum mechanics, particles are represented by waves.

The absolute square of the wave amplitude represents the probability of finding the particle.

In gauge theory the potential four-momentum performs this role for the virtual particles intermediary interactions.

Thus a larger potential four-momentum at some point means a higher probability of finding the related virtual particles at that point.

14.8 Negative Energies and Antiparticles

Figure 14.5 illustrates another oddity in the role of intermediary particles in collisions.

In the unprimedframe, particle C appears to be emitted by particle A and absorbed by particle B.

In the primed frame the reverse is true; it appears to be emitted by B and absorbed by A.

These judgements are based on the fact that the A vertex occurs earlier than the B vertex in the unprimed frame, while the B vertex occurs earlier in the primed frame.

However, since these distinctions are based on time ordering in different reference frames of events separated by a spacelike interval, they are inherently not relativistically invariant.

Since the principle of relativity states that physical laws are the same in all inertial reference frames, we have a conceptual problem to overcome.

A related problem has to do with the computation of energy from mass and momentum.

The solution of equation $E^2 = p^2c^2 + m^2c^4$ for the energy has a sign ambiguity that we have so far ignored:

$$E = \pm(p^2c^2 + m^2c^4)^{1/2}. \tag{14.25}$$

A natural tendency would be to omit the minus sign and just consider positive energies.

However, this would be a mistake — experience with quantum mechanics indicates that *both* solutions must be considered.

Richard Feynman won the Nobel Prize in physics largely for developing a consistent interpretation of the above negative energy solutions, which we now relate.

Notice that the four-momentum points backward in time in a spacetime diagram if the energy is negative.

Feynman suggested that a particle with four-momentum p is equivalent to the corresponding *antiparticle* with four-momentum $-p$.

Thus, we interpret a particle with momentum $\mathbf{p}$ and energy $E < 0$ as an antiparticle with momentum $-\mathbf{p}$ and energy $-E > 0$.

Antiparticles are known to exist for all particles.

If a particle and its antiparticle meet, they can annihilate, creating one or more other particles.

Correspondingly, if energy is provided in the right form, a particle-antiparticle pair can be created.

Suppose a particular kind of particle, call it an A particle, produces a B particle when it annihilates with its antiparticle A.

This is illustrated in the left panel in figure 14.6.

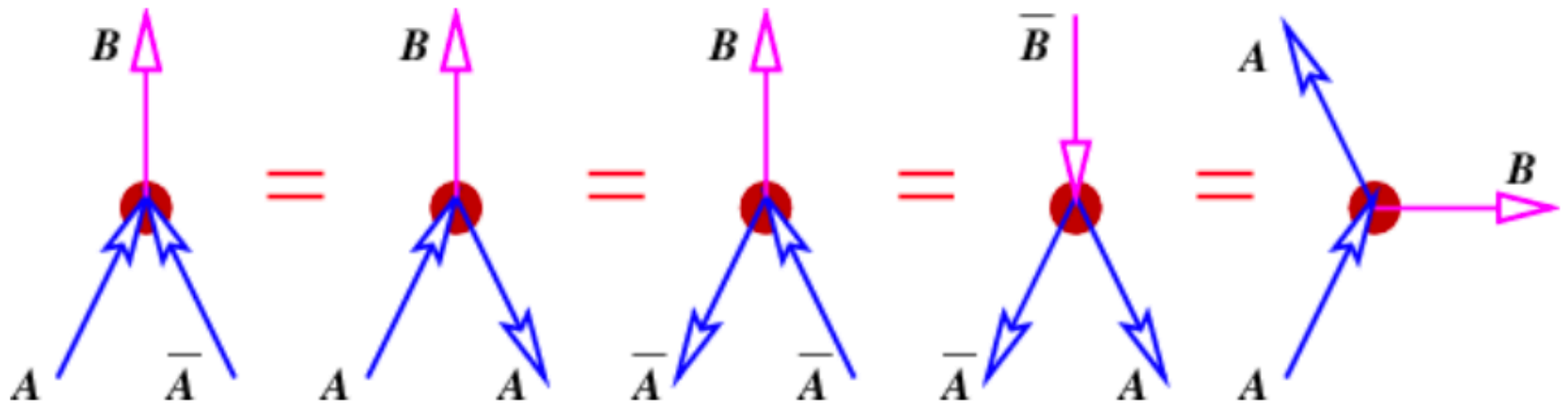


Figure 14.6: Equivalence of different processes according to Feynman's picture.

In Feynman's view, this process is equivalent to the scattering of an A particle backward in time by a B particle, the scattering of an A backward in time by a B particle, the creation of an AA pair moving backward in time by a B particle (an anti$B$), and the emission of a B particle by an A particle moving forward in time.

The statement "moving backward in time" has stimulated generations of physics students to contemplate the possibility that Feynman's picture makes time travel possible.

As far as we know, this is not so.

The key phrase is *equivalent to*.

In other words, causality still works forward in time as we have come to expect.

The real utility of the "backward in time" picture is that it makes calculations easier, since processes which are normally thought of as being very different turn out to have the same mathematical form.

Returning to the ambiguity shown in figure 14.5, it turns out that it does not matter whether the picture in the left or right panel is chosen.

According to the Feynman view the two processes are equivalent if one small correction is made — if the intermediary particle going from left to right is a *C* particle, then the intermediary particle going from right to left in the other picture is a *C* particle, or an anti*C*.

It is immaterial whether the arrow representing either the *C* or the *C* points forward or backward in time.

The key point is that if an arrow points *into* a vertex, the four-momentum of that particle contributes to the *input* side of the momentum-energy budget for that vertex.

If an arrow points *away from* a vertex, then the four-momentum contributes to the *output* side.

Chapter 15

Electromagnetic Forces

In this chapter we begin the study of electromagnetism.

The forces on charged particles due to electromagnetic fields are introduced and related to the general case of force on a particle by a gauge field.

The principles of electric motors and generators are then addressed as an example of such forces in action.

15.1 Electromagnetic Four-Potential

Electromagnetism is a gauge theory.

Particles that have a property called *electric charge* are subject to forces exerted by the gauge fields of electromagnetism.

The potential four-momentum $\underline{Q} = (\mathbf{Q}, U/c)$ of a particle with charge q in the presence of the electromagnetic four-potential $\underline{a}$ is just

$$\underline{Q} = q\underline{a}. \tag{15.1}$$

In the simplest case the four-potential represents the amplitude for finding the intermediary particle associated with the electromagnetic gauge field.

This particle has zero mass and is called the *photon*.

If more than one photon is present, the interpretation of $\underline{a}$ becomes more complicated.

This issue will be considered later.

The four-potential has space and time components $\mathbf{A}$ and φc such that $\underline{a} = (\mathbf{A}, \varphi/c)$.

The quantity $\mathbf{A}$ is called the *vector potential* and φ is called the *scalar potential*.

The scalar and vector potential are related to the potential energy U and potential momentum $\mathbf{Q}$ of a particle of charge q by

$$U = q\phi \qquad \mathbf{Q} = q\mathbf{A}. \tag{15.2}$$

The Lorenz condition written in terms of **A** and φ is

$$\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0. \tag{15.3}$$

15.2 Electric and Magnetic Fields and Forces

Electric and magnetic fields manifest themselves observationally by the forces that they cause.

These vector quantities are related to the scalar and vector potentials as follows:

$$\mathbf{E} = -\left(\frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z}\right) - \frac{\partial \mathbf{A}}{\partial t} \quad \text{(electric field)} \tag{15.4}$$

$$\mathbf{B} \equiv \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) \quad \text{(magnetic field)}. \tag{15.5}$$

Note that arbitrary scalar and vector constants may be added respectively to φ and **A** without changing either the electric or magnetic fields, since the latter are functions only of space and time derivatives of the former.

This is a simple example of the concept of *gauge invariance* in action.

We will see later that not just a constant, but any time-independent vector function $\mathbf{A}'(x,y,z)$ may be added to **A** with similar null results, as long as $\partial A_x'/\partial y = \partial A_y'/\partial x$, etc.

Gauge invariance is an important part of gauge theory, but a full understanding depends on more sophisticated mathematics than currently at our disposal.

By comparison of equations (15.4) and (15.5) with the general expression for force in gauge theory, we find that the electromagnetic force on a particle with charge q is

$$\begin{aligned}
\mathbf{F}_{em} &= -\left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z}\right) - \frac{\partial \mathbf{Q}}{\partial t} + \mathbf{v} \times \mathbf{P} \\
&= -q\left(\frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z}\right) - q\frac{\partial \mathbf{A}}{\partial t} + q\mathbf{v} \times \mathbf{B} \\
&= q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \quad \text{(Lorentz force)}
\end{aligned} \tag{15.6}$$

where $\mathbf{v}$ is the velocity of the particle and where we have used equations (15.2) and (15.4).

For historical reasons this is called the *Lorentz force*.

15.3 Charged Particle Motion

We now explore some examples of the motion of charged particles under the influence of electric and magnetic fields.

15.3.1 Particle in Constant Electric Field

Suppose a particle with charge q is exposed to a constant electric field E in the x direction.

The x component of the force on the particle is thus $F_x = qE_x$.

From Newton's second law the acceleration in the x direction is therefore $a_x = F_x/m = qE_x/m$ where m is the mass of the particle.

The behavior of the particle is the same as if it were exposed to a constant gravitational field equal to qE_x/m.

15.3.2 Particle in Conservative Electric Field

If $\partial \mathbf{A}/\partial t = 0$, then the electric force on a charged particle is

$$\mathbf{F}_{electric} = -q\left(\frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z}\right), \qquad \frac{\partial \mathbf{A}}{\partial t} = 0. \tag{15.7}$$

This force is conservative, with potential energy $U = q\varphi$.

Recalling that the total energy, $E = K + U$, of a particle under the influence of a conservative force remains constant with time, we can infer that the change in the kinetic energy with position of the particle is just minus the change in the potential energy: $\Delta K = -\Delta U$.

Notice in particular that if the particle returns to its initial position, the change in the potential energy is zero and the kinetic energy recovers its initial value.

If $\partial \mathbf{A}/\partial t \neq 0$, then there is the possibility that the electric force is not conservative.

Recall that the magnetic field is derived from $\mathbf{A}$.

Interestingly, a necessary and sufficient criterion for a non-conservative electric force is that the magnetic field be changing with time.

This result was first inferred experimentally by the English physicist Michael Faraday in 1831 and at nearly the same time by the American physicist Joseph Henry.

It will be further explored later in this chapter.

15.3.3 Torque on an Electric Dipole

Let us now imagine a "dumbbell" consisting of positive and negative charges of equal magnitude separated by a distance d, as shown in figure 15.1.

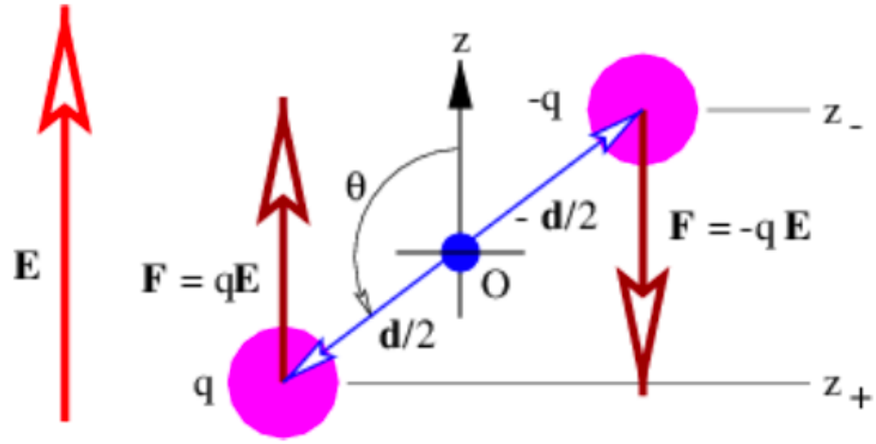


Figure 15.1: Definition sketch for an electric dipole. Two charges, q and -q are connected by an uncharged bar of length d. The vectors $\mathbf{d}/2$ and $-\mathbf{d}/2$ give the positions of the two charges relative to the central point between them. The two forces are due to the electric field **E**.

If there is a uniform electric field **E**, the positive charge experiences a force $q\mathbf{E}$, while the negative charge experiences a force $-q\mathbf{E}$.

The net force on the dumbbell is thus zero.

The torque acting on the dumbbell is not zero.

The total torque acting about the origin in figure 15.1 is the sum of the torques acting on the two charges:

$$\boldsymbol{\tau} = (-q)(-\mathbf{d}/2) \times \mathbf{E} + (q)(\mathbf{d}/2) \times (\mathbf{E}) = q\mathbf{d} \times \mathbf{E}. \tag{15.8}$$

The vector **d** can be thought of as having a length equal to the distance between the two charges and a direction going from the negative to the positive charge.

The quantity $\mathbf{p} = q\mathbf{d}$ is called the *electric dipole moment*.

Don't confuse it with the momentum!

The torque is just

$$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}. \tag{15.9}$$

This shows that the torque depends on the dipole moment, or the product of the charge and the separation.

Thus, halving the separation and doubling the charge results in the same dipole moment.

The tendency of the torque is to rotate the dipole so that the dipole moment **p** is parallel to the electric field **E**.

The magnitude of the torque is given by

$$\tau = pE\sin(\theta), \tag{15.10}$$

where the angle θ is defined in figure 15.1 and $p = |\mathbf{p}|$ is the magnitude of the electric dipole moment.

The potential energy of the dipole is computed as follows: The scalar potential associated with the electric field is $\varphi = -Ez$ where E is the magnitude of the field, assumed to point in the $+z$ direction.

Thus, the potential energy of a single particle with charge q is $U = q\varphi = -qEz$.

The total potential energy of the dipole is the sum of the potential energies of the individual charges:

$$\begin{aligned} U &= (+q)(-Ez_+) + (-q)(-Ez_-) = -qE(z_+ - z_-) \\ &= -qEd\cos(\theta) = -pE\cos(\theta) = -\mathbf{p}\cdot\mathbf{E}, \end{aligned} \tag{15.11}$$

where z_+ and z_- are the z positions of the positive and negative charges.

The equating of $z_+ - z_-$ to $d\cos(\theta)$ may be verified by examining the geometry of figure 15.1.

The tendency of the electric field to align the dipole moment with itself is confirmed by the potential energy formula.

The potential energy is lowest when the dipole moment is aligned with the field and highest when the two are anti-aligned.

15.3.4 Particle in Constant Magnetic Field

The magnetic force on a particle with charge q moving with velocity $\mathbf{v}$ is $\mathbf{F}_{\text{magnetic}} = q\mathbf{v} \times \mathbf{B}$, where $\mathbf{B}$ is the magnetic field.

The magnetic force is directed perpendicular to both the magnetic field and the particle's velocity.

Because of the latter point, no work is done on the particle by the magnetic field.

Thus, by itself the magnetic force cannot change the magnitude of the particle's velocity, though it can change its direction.

If the magnetic field is constant, the magnitude of the magnetic force on the particle is also constant and has the value $F_{\text{magnetic}} = qvB\sin(\theta)$ where $v = |\mathbf{v}|$, $B = |\mathbf{B}|$, and θ is the angle between $\mathbf{v}$ and $\mathbf{B}$.

If the initial velocity is perpendicular to the magnetic field, then $\sin(\theta) = 1$ and the force is just $F_{\text{magnetic}} = \text{qvB}$.

The particle simply moves in a circle with the magnetic force directed toward the center of the circle.

This force divided by the mass m must equal the particle's centripetal acceleration: $v^2/R = a = F_{\text{magnetic}}/\text{m} = qvB/m$ in the non-relativistic case, where R is the radius of the circle.

Solving for R yields

$$R = mv/(qB). \tag{15.12}$$

The angular frequency of revolution is

$$\omega = v/R = qB/m \quad \text{(cyclotron frequency)}. \tag{15.13}$$

Notice that this frequency is a constant independent of the radius of the particle's orbit or its velocity.

This is called the *cyclotron frequency*.

If the initial velocity is not perpendicular to the magnetic field, then the particle still has a circular component of motion in the plane normal to the field, but also drifts at constant speed in the direction of the field.

The net result is a spiral motion in the direction of the magnetic field, as illustrated in figure 15.2.

The radius of the circle is $R = mv_p/(qB)$ in this case, where v_p is the component of **v** perpendicular to the magnetic field.

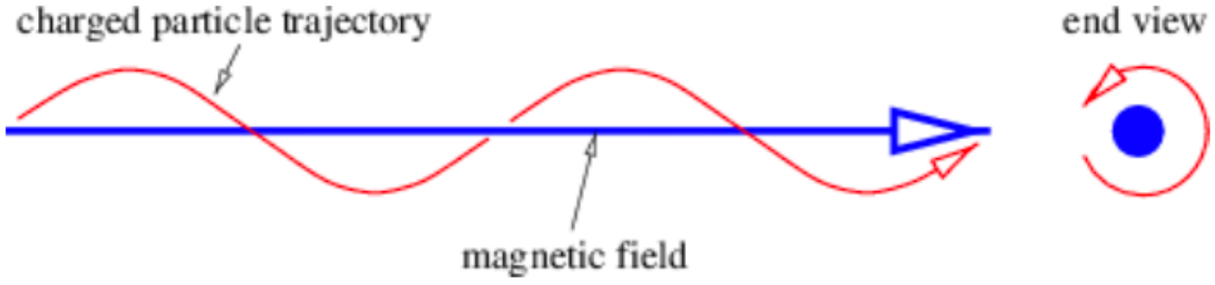


Figure 15.2: Spiraling motion of a charged particle in the direction of the magnetic field. This is composed of a circular motion about the field vector plus a translation along the field.

15.3.5 Crossed Electric and Magnetic Fields

If we have perpendicular electric and magnetic fields as shown in figure 15.3, then it is possible for a charged particle to move such that the electric and magnetic forces simply cancel each other out.

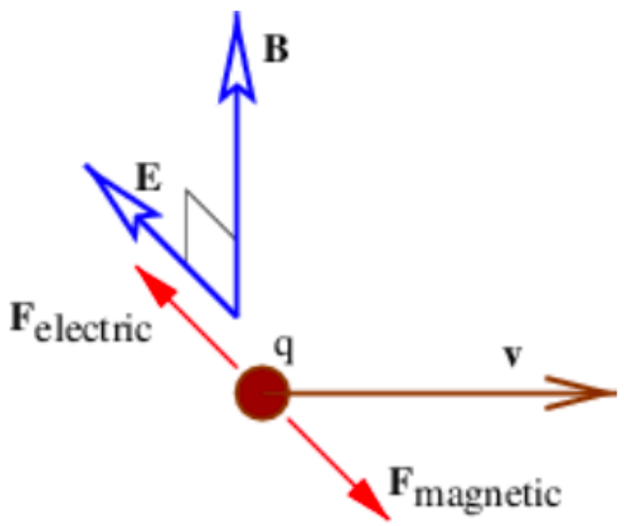


Figure 15.3: With crossed electric **E** and magnetic **B** fields (i. e., fields perpendicular to each other), a charged particle can move at a constant velocity **v** with magnitude equal to $v = |\mathbf{E}|/|\mathbf{B}|$ and direction perpendicular to both **E** and **B**. This is because the electric and magnetic forces, $\mathbf{F}_{\text{electric}} = q\mathbf{E}$ and $\mathbf{F}_{\text{magnetic}} = q\mathbf{v} \times \mathbf{B}$, balance each other in this case.

From the Lorentz force equation (15.6), the condition for this happening is $\mathbf{E} + \mathbf{v} \times \mathbf{B} = 0$.

If **E** and **B** are perpendicular, then this equation requires **v** to point in the direction of $\mathbf{E} \times \mathbf{B}$ (i. e., normal to both vectors) with the magnitude $v = |\mathbf{E}|/|\mathbf{B}|$.

This, of course, is not the only possible motion under these circumstances, just the simplest.

It is interesting to consider this situation from the point of view of a reference frame that is moving with the charged particle.

In this reference frame the particle is stationary and therefore not subject to the magnetic force.

Since the particle is not accelerating, the net force, which in this frame consists only of the electric force, is zero.

Hence, the electric field must be zero in the moving reference frame.

This argument shows that the electric field perceived in one reference frame is not necessarily the same as the electric field perceived in another frame.

Figure 15.4 shows why this is so.

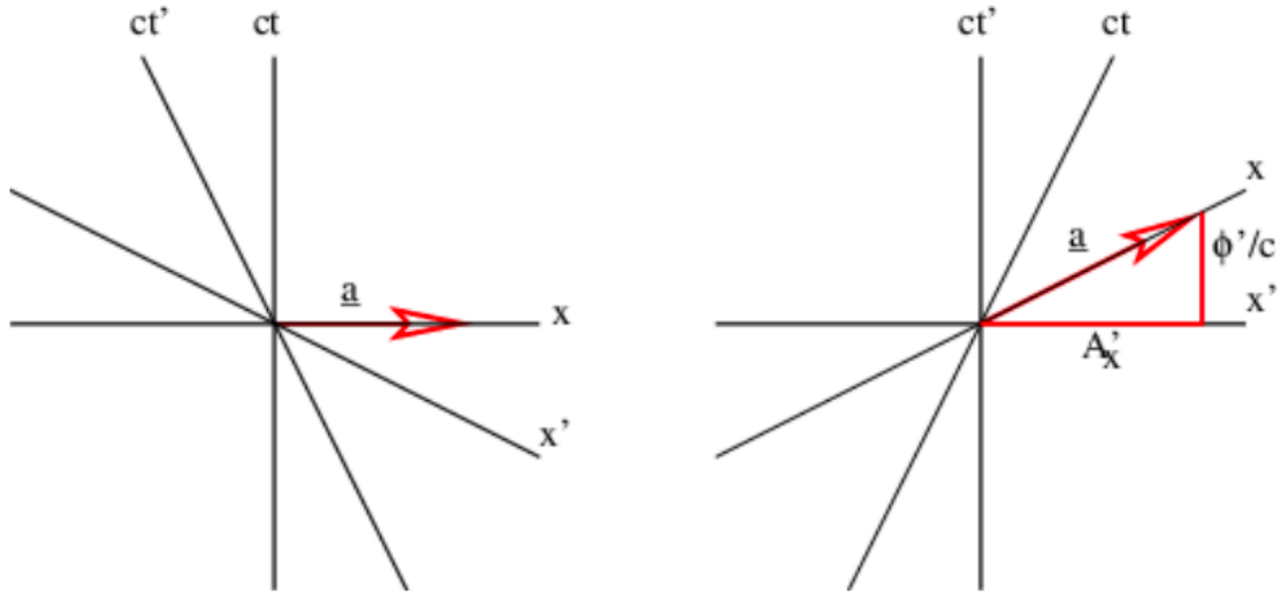


Figure 15.4: The four-potential *a* and its components in two different reference frames. In the unprimed frame the four-potential is purely space-like. The primed frame is moving in the *x* direction at speed *U* relative to the unprimed frame. The four-potential points along the *x* axis.

The left panel shows the situation in the reference frame moving to the right, which is the unprimed frame in this picture.

The charged particle is stationary in this reference frame.

The four-potential is purely spacelike, having no time component φ/c.

Assuming that a is constant in time, there is no electric field, and hence no electric force.

Since the particle is stationary in this frame, there is also no magnetic force.

However, in the primed reference frame, which is moving to the left relative to the unprimed frame and therefore is equivalent to the original reference frame in which the particle is moving to the right, the four-potential has a time component, which means that a scalar potential and hence an electric field is present.

15.4 Forces on Currents in Conductors

So far we have talked mainly about point charges moving in free space.

However, many practical applications of electromagnetism have charges moving through a *conductor* such as copper.

A conductor is a material in which electrically charged particles can freely move.

An *insulator* is a material in which charged particles are fixed in place.

Practical conductors are often surrounded by insulators in order to confine the motion of charge to particular paths.

The *current* through a wire is defined as the amount of charge passing through the wire per unit time.

When defining current, one needs to decide which direction constitutes a positive current for the problem at hand, i. e., the direction in which the positive charge is moving.

If the current consists of particles carrying negative charge, then the direction of the current is opposite the direction of the motion of the particles.

Metals tend to be good conductors, while glass, plastic, and other non-metallic materials are usually insulators.

All materials contain both positive and negative charges.

In metals, negatively charged electrons can escape from atoms and are free to move about the material.

When atoms lose one or more electrons, they become positively charged.

Atoms tend to be fixed in place.

Since the electron charge is negative, the current in a wire actually has a direction opposite the direction of motion of the electrons, as noted above (see Figure 15.5).

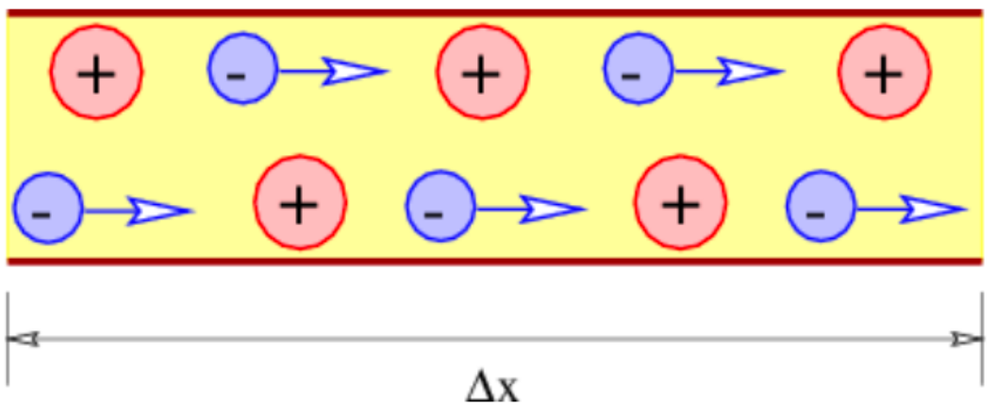


Figure 15.5: Stationary positive charge and negative charge moving to the right with speed *v* in blown up segment of wire.

If a conductor is in the form of a wire, we can compute the magnetic force on the wire if we know the number of mobile particles per unit length of wire N, the charge on each particle q, and the speed v with which they are moving down the wire.

The total force on a length of wire L is $\mathbf{F} = qNLv\mathbf{n} \times \mathbf{B}$, where $\mathbf{n}$ is a unit vector pointing in the direction of motion of the particles through the wire.

The quantity $i \equiv qNv$ is called the *current* in the wire.

It equals the amount of charge per unit time flowing down the wire.

Written in terms of the current, the force on a length L of the wire is

$$\mathbf{F} = iL\mathbf{n} \times \mathbf{B}. \tag{15.14}$$

15.5 Torque on a Magnetic Dipole and Electric Motors

Figure 15.6 shows a rectangular loop of wire mounted on an axle in a magnetic field.

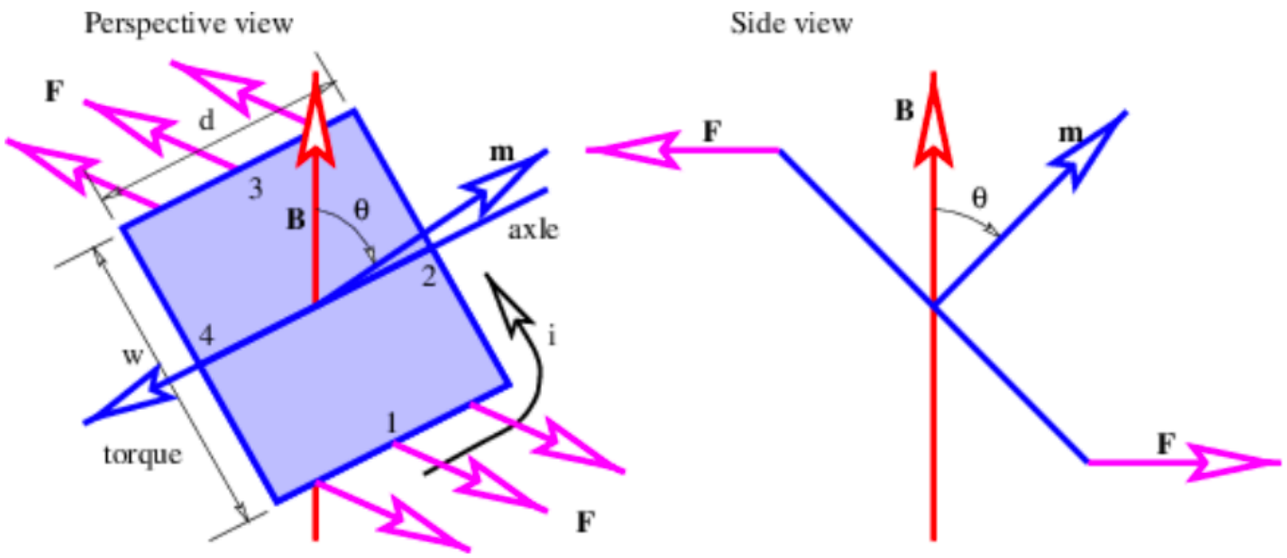


Figure 15.6: Perspective and side views of a rectangular loop of wire mounted on an axle in a magnetic field. Forces on the currents in loop segments 1 and 3 generate a torque about the axle.

A current i exists in the loop as shown.

The currents in loop segments 2 and 4 experience a force parallel to the axle.

These forces generate no net torque.

However, the magnetic forces on loop segments 1 and 3 are each $F = idB$ in magnitude, where $B = |\mathbf{B}|$ is the magnitude of the magnetic field.

Together these forces generate a counterclockwise torque about the axle equal to
$\tau = 2F(w/2)\sin(\theta) = iwdB\sin(\theta)$.

This can be represented in vector form as

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}, \tag{15.15}$$

where $\mathbf{m}$ is a vector with magnitude iwd and direction normal to the loop as shown in figure 15.6.

The vector $\mathbf{m}$ is called the *magnetic dipole moment*.

The loop can actually be any shape, not just rectangular.

In the general case the magnitude of the magnetic moment equals the current i times the area S of the loop:

$$|\mathbf{m}| = iS \quad (\text{magnetic dipole moment}). \tag{15.16}$$

In the above example the area is $S = wd$.

The direction of **m** is determined by the right hand rule; curl the fingers on your right hand around the loop in the direction of the current and your thumb points in the direction of **m**.

In analogy with the electric dipole in an electric field, the potential energy of a magnetic dipole in a magnetic field is

$$U = -\mathbf{m} \cdot \mathbf{B}. \tag{15.17}$$

Figure 15.6 illustrates the principle of an electric motor.

A motor consists of multiple loops of wire on an axle carrying a current in a magnetic field.

The torque on the axle turns the loops so that the magnetic moment is parallel to the field.

The angular momentum of the loops carries the rotation of the axle through the zero torque region, which occurs when the magnetic moment is either perfectly parallel or perfectly anti-parallel (i. e., pointing in the opposite direction) to the field.

At this point either the magnetic field is reversed by some mechanism or the magnetic dipole is reversed by making the current circulate around the loops in the opposite direction.

The torque due to the magnetic force then turns the axle through another half-turn, whereupon the field or the magnetic moment is again reversed, and so on.

15.6 Electric Generators and Faraday's Law

As was shown earlier, the electric field is derived from two different sources, spatial derivatives of the scalar potential and time derivatives of the vector potential:

$$E_x = -\frac{\partial \phi}{\partial x} - \frac{\partial A_x}{\partial t}, \quad E_y = -\frac{\partial \phi}{\partial y} - \frac{\partial A_y}{\partial t}, \quad E_z = -\frac{\partial \phi}{\partial z} - \frac{\partial A_z}{\partial t}. \tag{15.18}$$

In time independent situations the vector potential part drops out and we are left with a dependence only on the scalar potential.

In this case a particle with charge q has an electrostatic potential energy $U = q\varphi$, which means that the electric force is conservative.

However, in the time dependent situation there is no guarantee that the part of the electric field derived from the vector potential will be conservative.

An example of a non-conservative electric field occurs when we have

$$\mathbf{A} = (Cyt, -Cxt, 0), \qquad \phi = 0 \tag{15.19}$$

where C is a constant.

In this case the electric and magnetic fields are

$$\mathbf{E} = (-Cy, Cx, 0), \qquad \mathbf{B} = (0, 0, -2Ct). \tag{15.20}$$

The magnetic field points in the -z direction and increases in magnitude with time.

The electric field vectors are shown in figure 15.7.

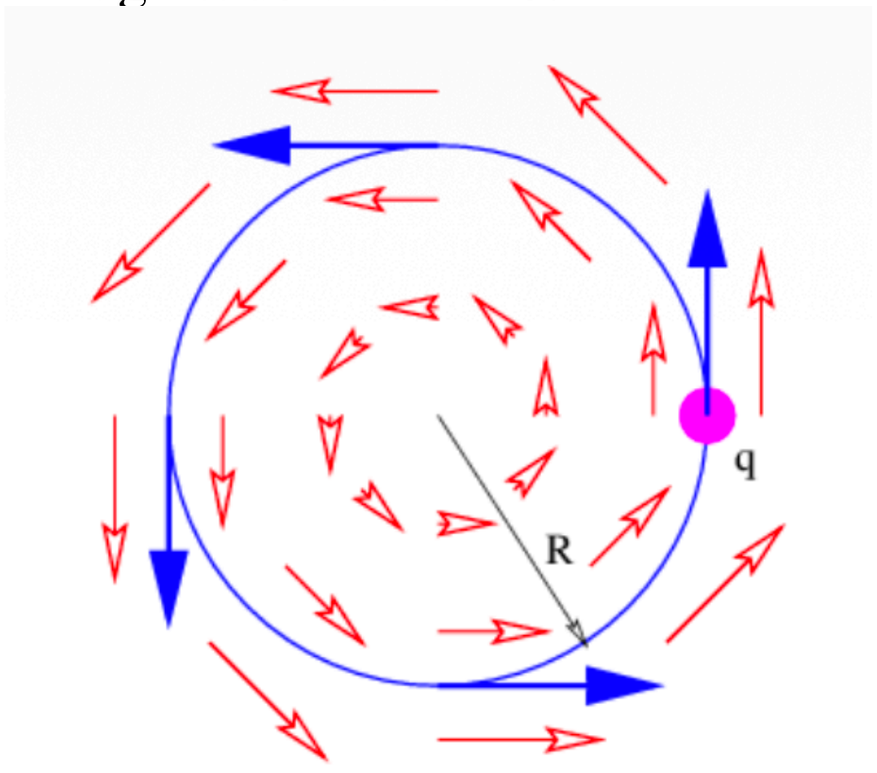


Figure 15.7: Illustration of the electric field pattern $\mathbf{E} = (-C_y, C_x, 0)$ (small arrows). A charged particle moving in a circle as shown continually gains energy.

Notice that a positively charged particle moving in a counterclockwise circle as shown is continually being accelerated in the direction of motion, and is therefore continually gaining energy.

This is impossible with a conservative force.

How much energy is gained by a particle with charge q moving in a complete circle of radius R under the above circumstances?

The magnitude of the electric field at this radius is $E = CR$, so the force on the particle is $F = qCR$.

The circumference of the circle is $2\pi R$, so the total work done by the electric field in one revolution is just $\Delta W = 2\pi RF = 2\pi qCR^2 = 2qCS$, where $S = \pi R^2$ is the area of the circle.

Let us define $\Delta V = \Delta W / q = 2CS$.

For historical reasons this is called the *electromotive force* or EMF.

This is deceptive terminology, because in fact ΔV doesn't have the dimensions of force — it is really just the work per unit charge done on a particle making a single loop around the circle in figure 15.7.

Recall that the z component of the magnetic field in this case is $B = -2Ct$.

Note that the time derivative of the magnetic field is just $\partial B_z / \partial t = -2C$.

Comparison with the equation for electromotive force shows us that

$$\Delta V = -\frac{\partial B_z}{\partial t} S = -\frac{\partial B_z S}{\partial t}, \tag{15.21}$$

where the area is brought inside the time derivative since it is constant in time.

This is a special case of a general law in electromagnetism called *Faraday's law*.

Notice that the argument of the time derivative in the above equation is the component of **B** perpendicular to the plane of the loop.

The loop area multiplied by the normal component of **B** is the *magnetic flux* through the loop: $\Phi_B \equiv B_{normal} S$.

Faraday's law is expressed most compactly as

$$\Delta V = -\frac{d\Phi_B}{dt} \quad \text{(Faraday's law)}, \tag{15.22}$$

and it turns out to be valid for arbitrary loops and arbitrary magnetic field configurations, not just for the simple loop we have been investigating.

The most general statement of the law is that the *EMF around a* closed loop equals minus the time rate of change of magnetic flux through the loop.

The minus sign in equation (15.22) means the following: If the fingers on your right hand curl around the loop in the direction *opposite* to the direction that causes a positive charge to gain energy, then your thumb points in the direction of the *time rate of change* of the magnetic flux passing through the loop.

This is illustrated in figure 15.8.

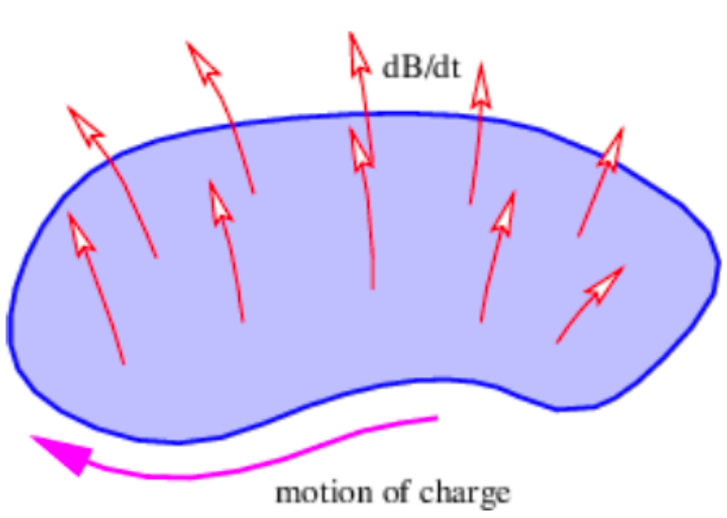


Figure 15.8: Sketch illustrating Faraday's law. The arrows passing through the loop indicate the direction of the *time rate of change of the magnetic field*. The arrow going around the loop indicates the direction a positive charge would be pushed by the electric field.

The electric generator is perhaps the best known application of Faraday's law.

Figure 15.9 shows a rectangular loop of wire fixed to an axle that rotates at an angular rate Ω.

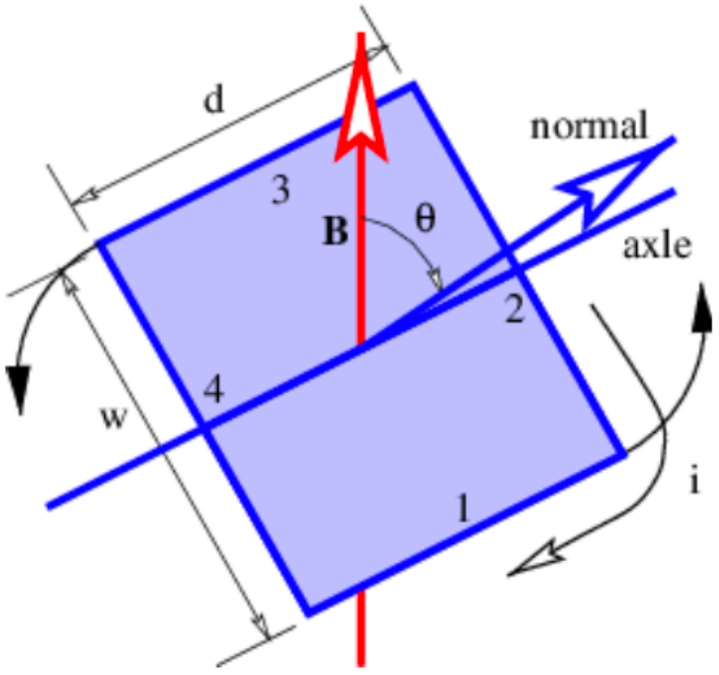


Figure 15.9: Rotating wire loop in a magnetic field. At the instant illustrated the magnetic flux is increasing with time, which means that an EMF tends to drive a current as illustrated.

The magnetic flux through the loop thus varies with time according to $\Phi_B = wdB\cos(\theta) = wdB\cos(\Omega t)$.

The EMF around the loop is thus

$$\Delta V = -\frac{d\Phi_B}{dt} = \Omega wdB\sin(\Omega t). \tag{15.23}$$

In a real generator there are many loops forming a coil of wire and the ends of the coil are brought out through the axle so that the resulting current can be tapped for practical use.

15.7 EMF and Scalar Potential

The EMF ΔV has the same units as the scalar potential φ.

What is the difference between the two quantities?

Both represent work done per unit charge by the electric field on a particle moving through the field.

However, recall that the electric field is composed of two parts:

$$\mathbf{E} = -\left(\frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z}\right) - \frac{\partial \mathbf{A}}{\partial t}. \qquad (15.24)$$

$\Delta\varphi = \varphi_2 - \varphi_1$ is *minus* the work done on the particle in going from point 1 to point 2 by the part of the electric field associated with the scalar potential.

ΔV is (*plus*) the work done by the part of the electric field associated with the time derivative of the vector potential.

This sign difference is consistent with the general result that the work on a particle for a conservative force equals minus the change in potential energy.

Aside from the different sign conventions, there is one other fundamental difference between the two quantities: $\Delta\varphi$ is always zero for closed paths, i. e., paths in which the particle returns to its initial point.

This is because point 1 is then the same as point 2, so $\varphi_1 = \varphi_2$.

This condition doesn't necessarily apply to the EMF.

ΔV often is non-zero for closed particle paths.

The electric generator that we have just discussed is an important case in point.

The total work done per unit charge by the electric field on a charged particle moving along some path is thus ΔV - $\Delta\varphi$.

The $\Delta\varphi$ term drops out if the path is closed.

Chapter 16

Generation of Electromagnetic Fields

In this chapter we investigate how charge produces electric and magnetic fields.

We first introduce Coulomb's law, which is the basis for everything else in the section.

We then discuss Gauss's law for the electric and magnetic field, drawing on what we learned while using it on the gravitational field. Coulomb's law and the theory of relativity together show that magnetic fields are generated by moving charge.

We then use this fact to compute the magnetic fields from some simple charge distributions.

We finish with a discussion of electromagnetic waves.

16.1 Coulomb's Law and the Electric Field

A stationary point electric charge q is known to produce a scalar potential

$$\phi = \frac{q}{4\pi\epsilon_0 r} \tag{16.1}$$

a distance r from the charge.

The constant $\epsilon_0 = 8.85 \times 10^{-12}$ C^2 N^{-1} m^{-2} is called the *permittivity of free space*.

The vector potential produced by a stationary charge is zero.

The potential energy between two stationary charges is equal to the scalar potential produced by one charge multiplied by the value of the other charge:

$$U = \frac{q_1 q_2}{4\pi\epsilon_0 r}. \tag{16.2}$$

Notice that it doesn't make any difference whether one multiplies the scalar potential from charge 1 by charge 2 or vice versa – the result is the same.

Since $r = (x^2 + y^2 + z^2)^{1/2}$, the electric field produced by a charge is

$$\mathbf{E} = -\left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right) = \frac{q\mathbf{r}}{4\pi\epsilon_0 r^3} \tag{16.3}$$

where $\mathbf{r} = (x, y, z)$ is the vector from the charge to the point where the electric field is being measured.

The magnetic field is zero since the vector potential is zero.

The force between two stationary charges separated by a distance r is the value of one charge multiplied by the electric field produced by the other charge.

Thus the magnitude of the force is

$$F = \frac{q_1 q_2}{4\pi\epsilon_0 r^2} \quad \text{(Coulomb's law)}, \tag{16.4}$$

with the force being repulsive if the charges are of the same sign, and attractive if the signs are opposite.

This is called Coulomb's law.

Equation (16.4) is the electric equivalent of Newton's universal law of gravitation.

Replacing mass by charge and G by $-1/(4\pi\epsilon_0)$ in the equation for the gravitational force between two point masses gives us equation (16.4).

The most important aspect of this result is that both the gravitational and electrostatic forces decrease as the square of the distance between the particles.

16.2 Gauss's Law for Electricity

The electric flux is defined in analogy to the gravitational flux as

$$\Phi_E = \mathbf{S} \cdot \mathbf{E} \quad \text{(electric flux)} \tag{16.5}$$

where **S** is the directed area through which the flux passes.

This is strictly true only for small, flat areas **S** over which the component of **E** normal to **S** can be assumed constant.

Since the electric field obeys an inverse square law, Gauss's law applies to the electric flux Φ_E just as it applies to the gravitational flux.

In particular, since the magnitude of the outward electric field a distance r from a charge q is $E = q/(4\pi\epsilon_0 r^2)$, the electric flux through a sphere of radius r (and area $4\pi r^2$) concentric with the charge is $ES = [q/(4\pi\epsilon_0 r^2)] \times (4\pi r^2) = q/\epsilon_0$.

This generalizes to an arbitrary distribution of charge as in the gravitational case:

$$\Phi_E = q_{inside}/\epsilon_0 \quad \text{(Gauss's law for electricity)}, \tag{16.6}$$

where Φ_E in this equation is the outward electric flux through a closed surface and q_{inside} is the net charge inside this surface.

This is an expression of Gauss's law for the electric field.

Since Gauss's law for electricity and for gravitation are so similar, we can use all our insights from studying gravity on the electric field case.

16.2.1 Sheet of Charge

Figure 16.1 shows how to set up the Gaussian surface to obtain the electric field emanating from an infinite sheet of charge.

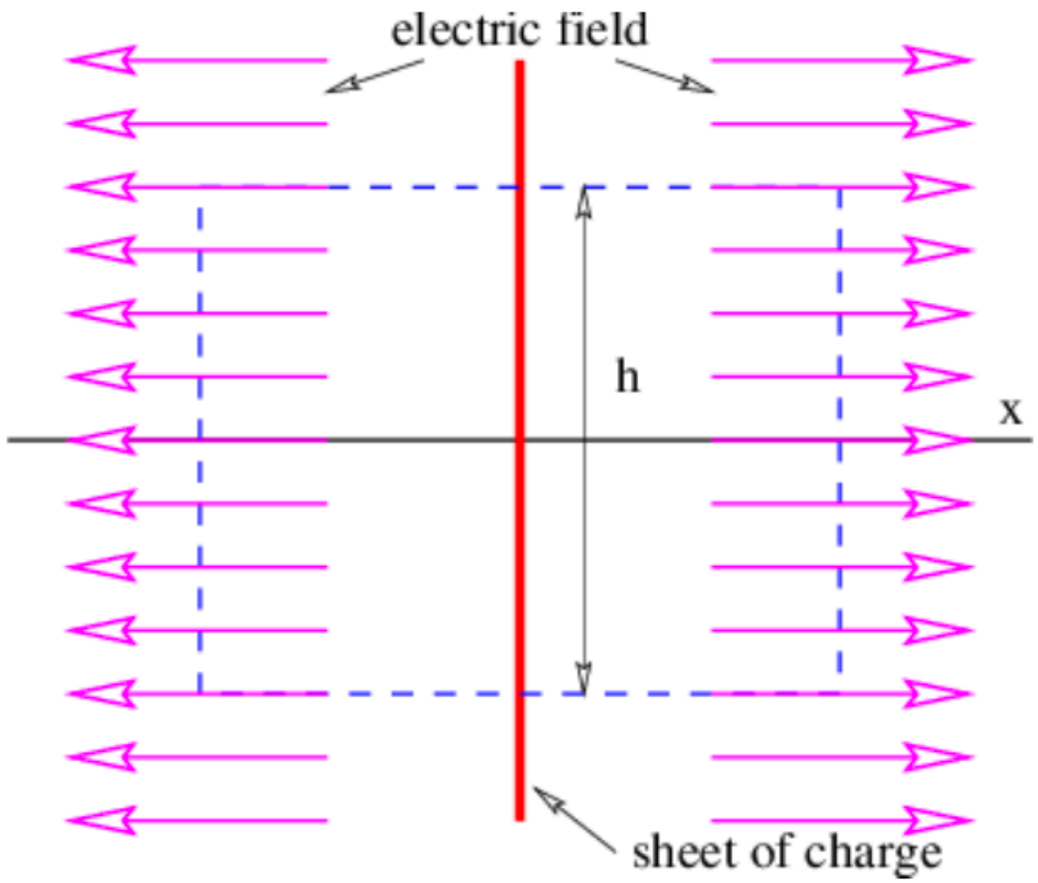


Figure 16.1: Definition sketch for use of Gauss's law to obtain the electric field due to an infinite sheet of surface charge. The dashed line shows the Gaussian box, which is of height h and depth d into the page

We assume a charge density of σ Coulombs per square meter, which means that the amount of charge inside the box is $q_{inside} = \sigma hd$, where the box has height h and depth d into the page.

The total electric flux out of the left and right faces of the box is $\Phi_E = 2Ehd$, where E is the magnitude of the electric field on these surfaces.

The field is assumed to point away from the charge, and hence out of the box on both faces.

Due to the assumed direction of the electric field, there is no electric flux out of any of the other faces of the box.

Applying Gauss's law, we infer that $2Ehd = \sigma hd/\epsilon_0$, which means that the electric field emanating from a sheet of charge with charge density per unit area σ is

$$E = \frac{\sigma}{2\epsilon_0}. \tag{16.7}$$

The scalar potential associated with this electric field is easily obtained by realizing that equation (16.7) gives the x component of this field — the other components are zero.

Using $E = -\partial\varphi/\partial x$, we infer that

$$\phi = -\frac{\sigma|x|}{2\epsilon_0}. \tag{16.8}$$

The absolute value signs around x take account of the fact that the direction of the electric field for negative x is opposite that for positive x.

16.2.2 Line of Charge

Similar reasoning is used to obtain the electric field due to a line of charge.

A sketch of the expected electric field vectors and a Gaussian cylinder coaxial with the line of charge is shown in figure 16.2.

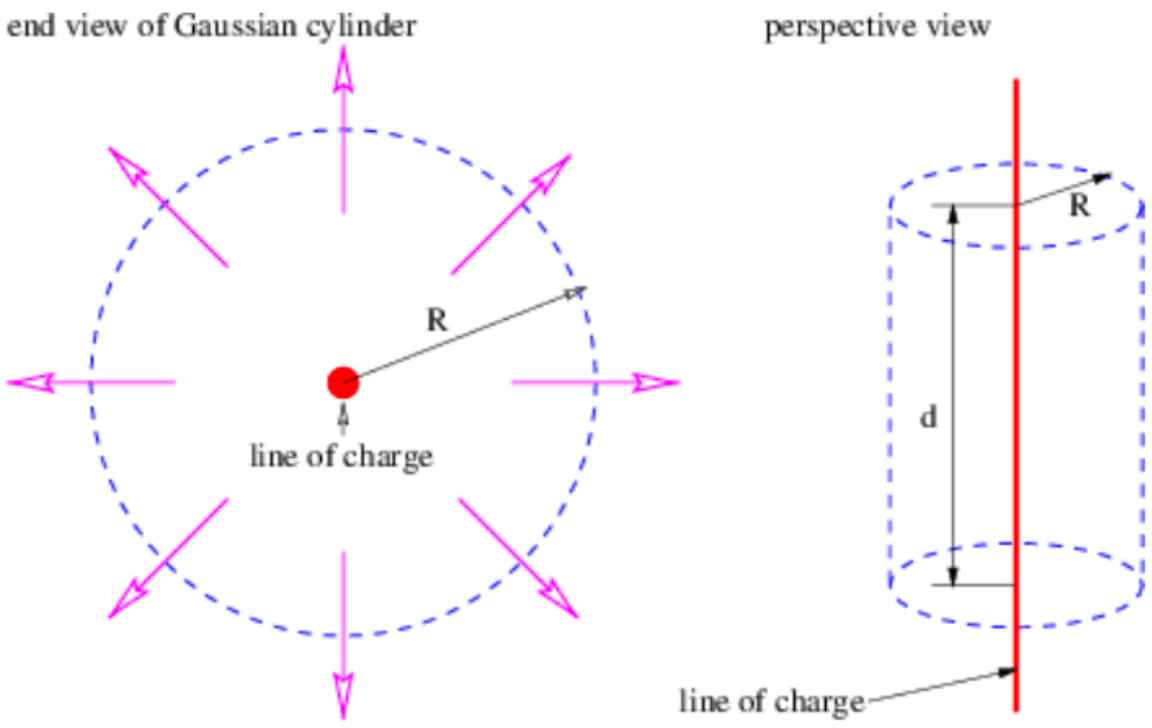


Figure 16.2: Definition sketch for use of Gauss's law to obtain the electric field due to an infinite line charge oriented normal to the page. The dashed line shows the Gaussian cylinder, which is of radius R and length d into the page. The outward-pointing arrows show the electric field.

If the charge per unit length is λ, the amount of charge inside the cylinder is $q_{inside} = \lambda d$, where d is the length of the cylinder.

The outward electric flux at radius r is $\Phi_E = 2\pi r d E$. Gauss's law therefore tells us that the electric field at radius r is just

$$E = \frac{\lambda}{2\pi\epsilon_0 r}. \tag{16.9}$$

In this case $E = -\partial\varphi/\partial r$, so that the scalar potential is

$$\phi = -\frac{\lambda}{2\pi\epsilon_0}\ln(r). \tag{16.10}$$

16.3 Gauss's Law for Magnetism

By analogy with Gauss's law for the electric field, we could write a Gauss's law for the magnetic field as follows:

$$\Phi_B = C q_{magnetic\ inside}, \tag{16.11}$$

where Φ_B is the outward magnetic flux through a closed surface, C is a constant, and $q_{magnetic\ inside}$ is the "magnetic charge" inside the closed surface.

Extensive searches have been made for magnetic charge, generally called a *magnetic monopole*.

However, none has ever been found.

Thus, Gauss's law for magnetism can be written

$$\Phi_B = 0 \qquad \text{(Gauss's law for magnetism)}. \tag{16.12}$$

This of course doesn't preclude non-zero values of the magnetic flux through open surfaces, as illustrated in figure 16.3.

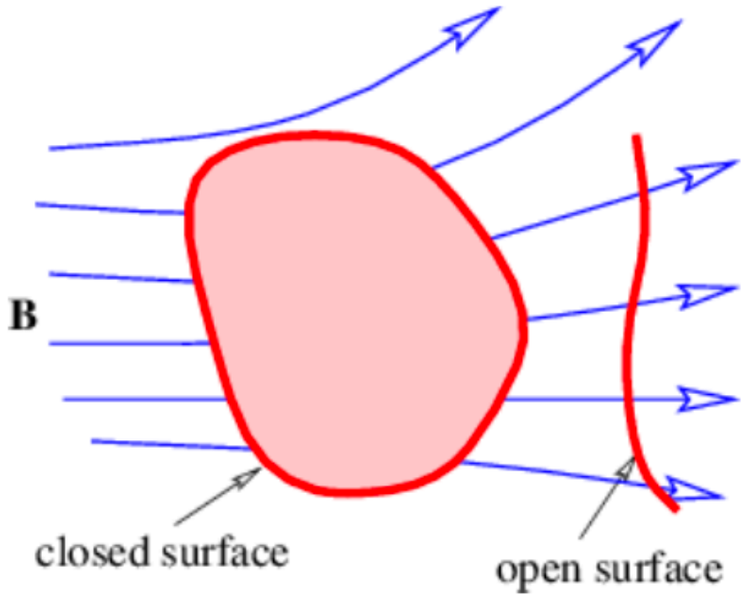


Figure 16.3: Illustration for Gauss's law for magnetism. The net flux out of the closed surface is zero, but the flux through the open surface is not.

16.4 Coulomb's Law and Relativity

The equation (16.1) for the scalar potential of a point charge is valid only in the reference frame in which the charge q is stationary.

By symmetry, the vector potential must be zero.

Since φ is actually the timelike component of the four-potential, we infer that the four-potential due to a charge is tangent to the world line of the charged particle.

A consequence of the above argument is that a moving charge produces a magnetic field, since the four-potential must have spacelike components in this case.

16.5 Moving Charge and Magnetic Fields

We have shown that electric charge generates both electric and magnetic fields, but the latter result only from moving charge.

If we have the scalar potential due to a static configuration of charge, we can use this result to find the magnetic field if this charge is set in motion.

Since the four-potential is tangent to the particle's world line, and hence is parallel to the time axis in the reference frame in which the charged particle is stationary, we know how to resolve the space and time components of the four-potential in the reference frame in which the charge is moving.

Figure 16.4 illustrates this process.

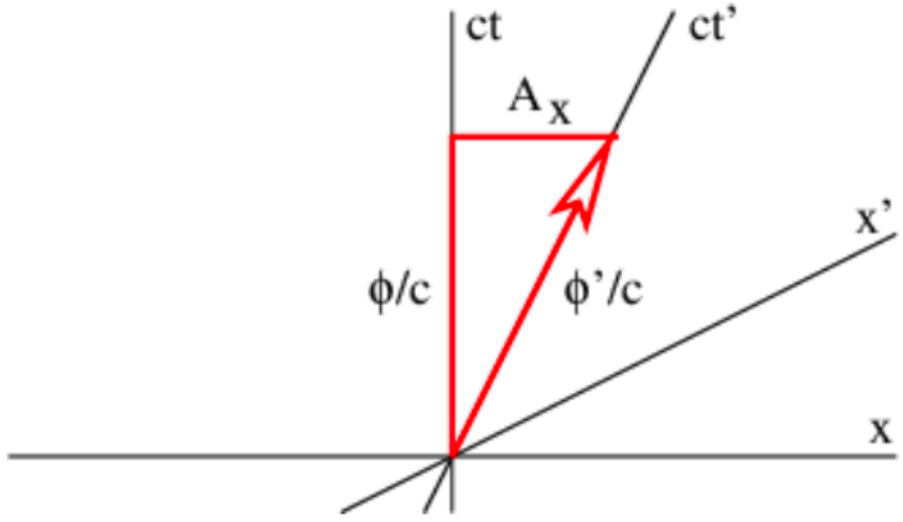


Figure 16.4: Finding the space and time components of the four-potential produced by a particle moving at the velocity of the primed reference frame. The ct' axis is the world line of the charged particle that generates the four-potential.

For a particle moving in the $+x$ direction at speed v, the slope of the time axis in the primed frame is just c/v.

The four-potential vector has this same slope, which means that the space and time components of the four-potential must now appear as shown in figure 16.4.

If the scalar potential in the primed frame is φ', then in the unprimed frame it is φ, and the x component of the vector potential is A_x.

Using the spacetime Pythagorean theorem, $\varphi'^2/c^2 = \varphi^2/c^2 - A_x^2$, and relating slope of the ct' axis to the components of the four-potential, $c/v = (\varphi/c)/A_x$, it is possible to show that

$$\phi = \gamma\phi' \qquad A_x = v\gamma\phi'/c^2 \tag{16.13}$$

where

$$\gamma = \frac{1}{(1 - v^2/c^2)^{1/2}}. \tag{16.14}$$

Thus, the principles of special relativity allow us to obtain the full four-potential for a moving configuration of charge if the scalar potential is known for the charge when it is stationary.

From this we can derive the electric and magnetic fields for the moving charge.

16.5.1 Moving Line of Charge

As an example of this procedure, let us see if we can determine the magnetic field from a line of charge with linear charge density in its own rest frame of λ', aligned along the z axis.

The line of charge is moving in a direction parallel to itself.

From equation (16.10) we see that the scalar potential a distance r from the z axis is

$$\phi' = -\frac{\lambda'}{2\pi\epsilon_0}\ln(r) \tag{16.15}$$

in a reference frame moving with the charge.

The z component of the vector potential in the stationary frame is therefore

$$A_z = -\frac{\lambda' v\gamma}{2\pi\epsilon_0 c^2}\ln(r) \tag{16.16}$$

by equation (16.13), with all other components being zero.

This is illustrated in figure 16.5.

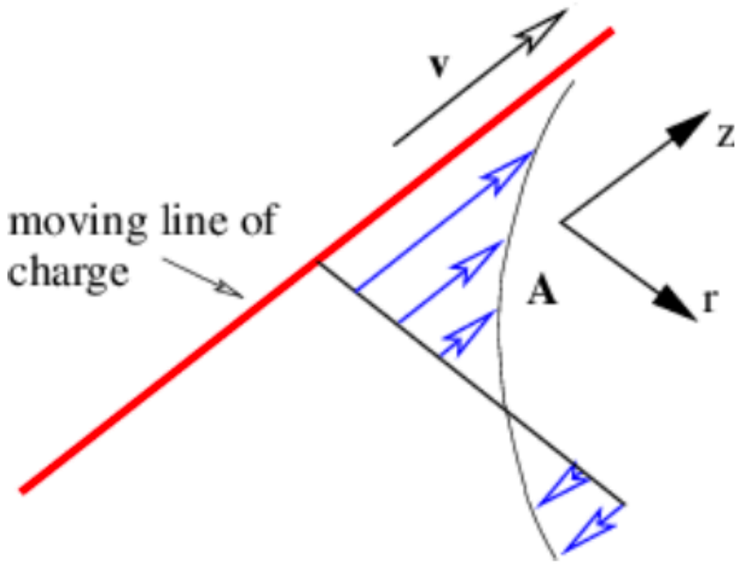


Figure 16.5: Vector potential from a moving line of charge. The distribution of vector potential around the line is cylindrically symmetric.

We infer that

$$B_x = \frac{\partial A_z}{\partial y} = -\frac{\lambda' v\gamma y}{2\pi\epsilon_0 c^2 r^2} \quad B_y = -\frac{\partial A_z}{\partial x} = \frac{\lambda' v\gamma x}{2\pi\epsilon_0 c^2 r^2} \quad B_z = 0, \tag{16.17}$$

where we have used $r^2 = x^2 + y^2$.

The resulting field is illustrated in figure 16.6.

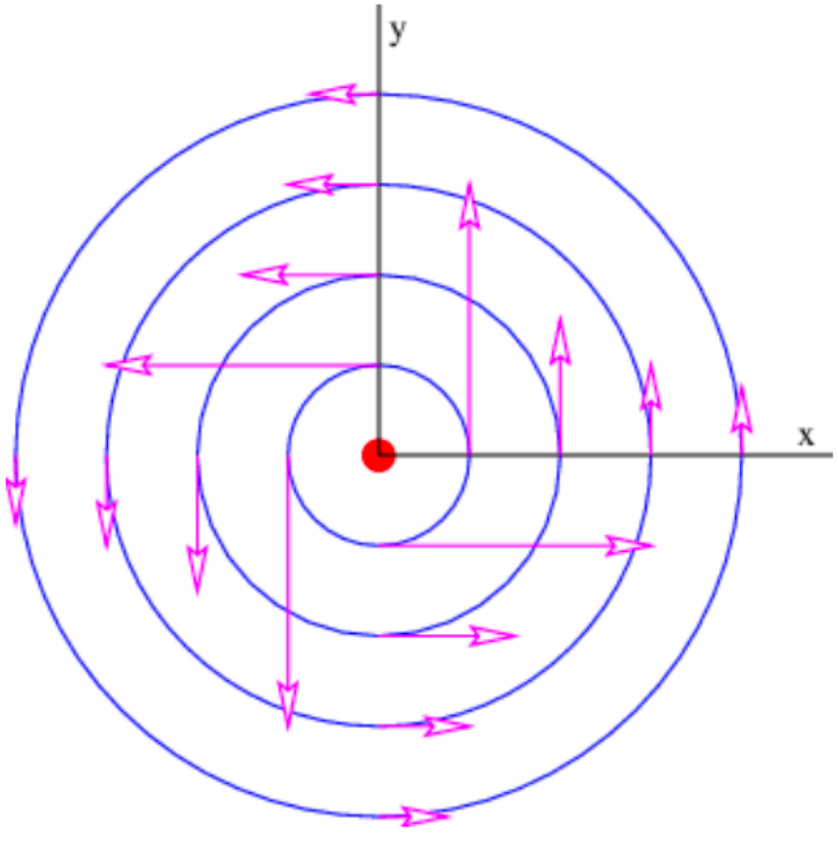


Figure 16.6: Magnetic field from a moving line of charge. The charge is moving along the z axis out of the page.

The field lines circle around the line of moving charge and the magnitude of the magnetic field is

$$B = (B_x^2 + B_y^2)^{1/2} = \frac{\lambda' v \gamma}{2\pi\epsilon_0 c^2 r}. \tag{16.18}$$

There is an interesting relativistic effect on the charge density λ', which is defined in the co-moving or primed reference frame.

In the unprimed frame the charges are moving at speed v and therefore undergo a Lorentz contraction in the z direction.

This decreases the charge spacing by a factor of γ and therefore increases the charge density as perceived in the unprimed frame to a value $\lambda = \gamma\lambda'$.

We also define a new constant $\mu_0 \equiv 1/(\epsilon_0 c^2)$.

This is called the *permeability of free space*.

This constant has the *assigned* value $\mu_0 = 4\pi \times 10^{-7}$ N s^2 C^{-2}.

The value of $\epsilon_0 = 1/(\mu_0 c^2)$ is actually derived from this assigned value and the measured value of the speed of light.

The reasons for this particular way of dealing with the constants of electromagnetism are obscure, but have to do with making it easy to relate the values of constants to the experiments used in determining them.

With the above substitutions, the magnetic field equation becomes

$$B = \frac{\mu_0 \lambda v}{2\pi r}. \tag{16.19}$$

The combination λv is called the *current* and is symbolized by i.

The current is the charge per unit time passing a point and is a fundamental quantity in electric circuits.

The magnetic field written in terms of the current flowing along the z axis is

$$B = \frac{\mu_0 i}{2\pi r} \quad \text{(straight wire)}. \tag{16.20}$$

16.5.2 Moving Sheet of Charge

As another example we consider a uniform infinite sheet of charge in the *x*-*y* plane with charge density σ'.

The charge is moving in the $+x$ direction with speed v.

As we showed in the section on Gauss's law for electricity, the electric field for this sheet of charge in the co-moving reference frame is in the z direction and has the value

$$E'_z = \frac{\sigma'}{2\epsilon_0}\mathrm{sgn}(z) \tag{16.21}$$

where we define

$$\mathrm{sgn}(z) \equiv \begin{cases} -1 & z < 0 \\ 0 & z = 0 \\ 1 & z > 0 \end{cases} . \tag{16.22}$$

The sgn(z) function is used to indicate that the electric field points upward above the sheet of charge and downward below it (see figure 16.7).

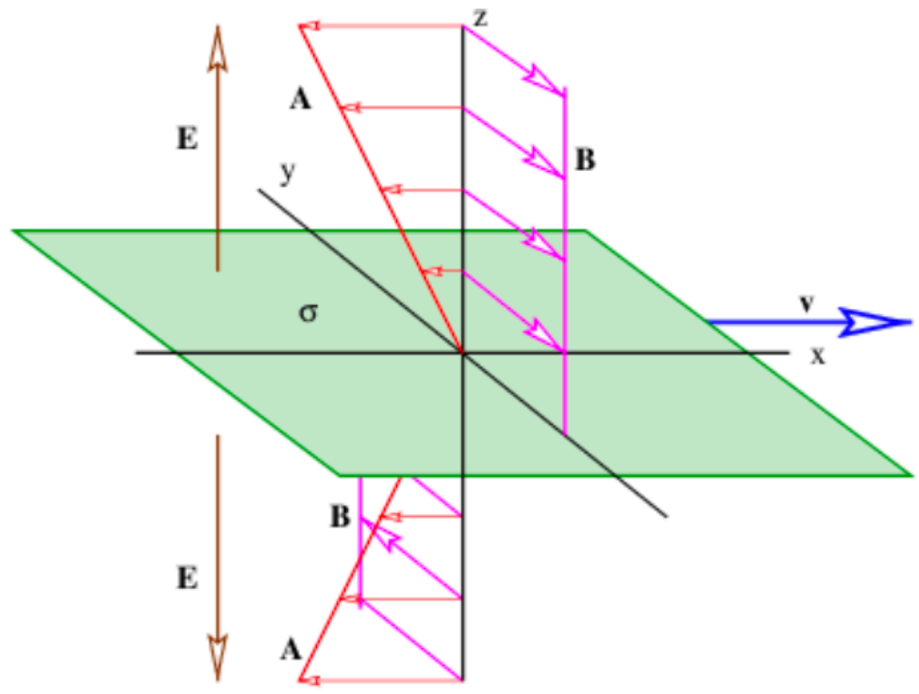


Figure 16.7: Vector potential **A**, electric field **E**, and magnetic field **B** from a moving sheet of charge. The charge is moving in the x direction.

In the stationary reference frame in which the sheet of charge is moving in the x direction, the scalar potential and the x component of the vector potential are

$$\phi = -\frac{\gamma\sigma'|z|}{2\epsilon_0} = -\frac{\sigma|z|}{2\epsilon_0} \qquad A_x = -\frac{v\gamma\sigma'|z|}{2\epsilon_0 c^2} = -\frac{v\sigma|z|}{2\epsilon_0 c^2}, \tag{16.24}$$

according to equation (16.13), where $\sigma = \gamma\sigma'$ is the charge density in the stationary frame.

The other components of the vector potential are zero.

We calculate the magnetic field as

$$B_x = 0 \qquad B_y = \frac{dA_x}{dz} = -\frac{v\sigma}{2\epsilon_0 c^2}\mathrm{sgn}(z) \qquad B_z = 0 \tag{16.25}$$

where sgn(z) is defined as before.

The vector potential and the magnetic field are shown in figure 16.7.

Note that the magnetic field points normal to the direction of motion of the charge but parallel to the sheet.

It points in opposite directions on opposite sides of the sheet of charge.

16.6 Electromagnetic Radiation

We have found so far that stationary charge produces an electric field while moving charge produces a magnetic field.

It turns out that *accelerated* charge produces *electromagnetic radiation*.

Electromagnetic radiation is nothing more than one or more photons that have zero mass, and are therefore real, not virtual.

Acceleration of a charged particle is needed to produce radiation because of the conservation of energy and momentum.

The left panel of figure 16.8 shows why.

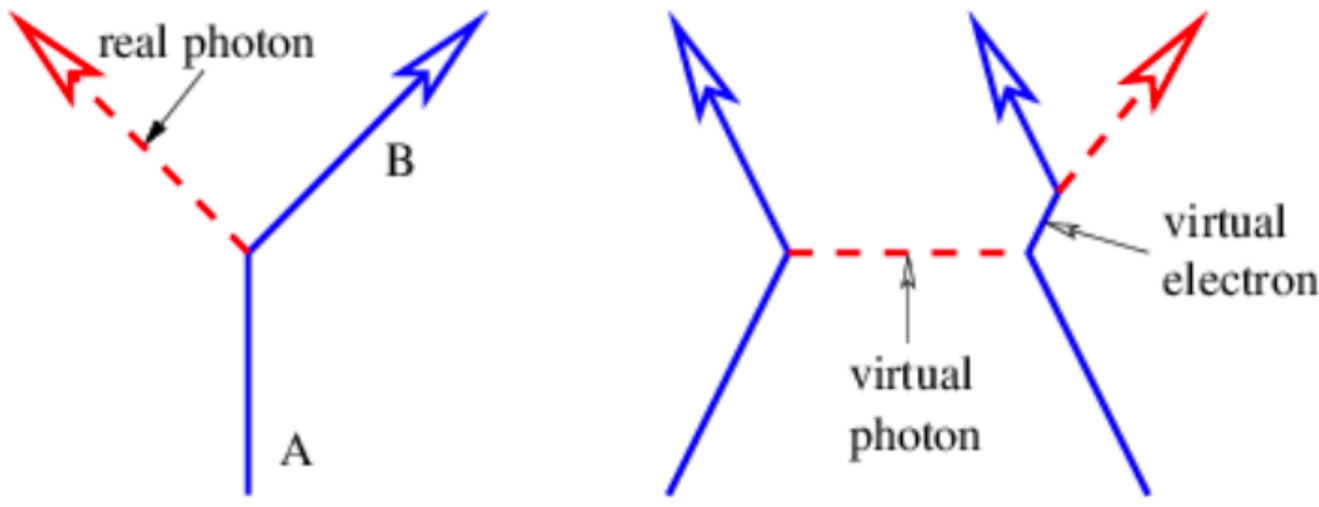


Figure 16.8: Feynman diagrams for two processes that potentially might produce real photons and hence electromagnetic radiation. The process in the left panel turns out to be impossible if the masses of particles A and B are the same, for reasons discussed in the text. The process in the right panel occurs commonly. Solid lines represent electrons while dashed lines represent photons. Particles are taken to be real unless otherwise labeled.

Since a photon carries off energy and momentum, conservation means that the energy and momentum of the emitting particle change due to the emission of a photon.

This corresponds in classical mechanics to an acceleration.

The process in the left panel of figure 16.8 actually cannot occur if particles A and B have the same mass.

If the mass of the outgoing particle B is less than the mass of the incoming particle A, then this reaction can and does occur.

An example is the decay of an atom from a higher energy state to a lower energy state (and hence lower mass), accompanied by the emission of a photon.

Another type of reaction that can generate radiation occurs when two charged particles (say, electrons) collide, as illustrated in the right panel of figure 16.8.

In an elastic collision both electrons are real both before and after the photon transfer.

However, it is possible for one of the electrons to have a virtual mass that is greater than the normal electron mass after the collision, which means that it is free to decay to a real electron plus a real photon.

We now try to understand the characteristics of free electromagnetic radiation.

In our studies of waves we found it easiest to examine plane waves.

We will follow this path here, writing the four-potential for an electromagnetic plane wave moving in the x direction as

$$\underline{a} = (\mathbf{A}, \phi/c) = (\mathbf{A_0}, \phi_0/c)\cos(k_x x - \omega t), \tag{16.26}$$

where $a_0 = (\mathbf{A}_0, \varphi_0/c)$ is a constant four-vector representing the direction and maximum amplitude of the four-potential, and k_x and ω are the wavenumber and the angular frequency of the wave.

Since the real photon is massless, we have $\omega = k_x c$ in this case.

Virtual photons are not subject to this constraint.

By substituting **A** and φ from equation (16.26) into the Lorenz condition, we find that

$$k_x A_x - \omega\phi/c^2 = 0 \quad \text{(Lorenz condition for plane wave).} \tag{16.27}$$

Thus, the Lorenz condition requires that the scalar potential ϕ be related to the *x* or *longitudinal component* of the vector potential, A_x, i. e., the component pointing in the direction of wave propagation.

The *transverse components*, A_y *and* A_z*, are unconstrained by the Lorenz condition, since they don't* depend on *y* and *z*.

Using equations for the electric and magnetic field, as well as equations (16.26) and (16.27), we can now find **E** and **B** in an electromagnetic plane wave:

$$\mathbf{B} = (0, k_x A_{0z}, -k_x A_{0y}) \sin(k_x x - \omega t) \tag{16.28}$$

$$\mathbf{E} = (k_x \phi_0 - \omega A_{0x}, -\omega A_{0y}, -\omega A_{0z}) \sin(k_x x - \omega t). \tag{16.29}$$

The electric field has a longitudinal or *x* component proportional to $k_x \phi_0 - \omega A_{0x} = -\omega(\phi_0/c - A_{0x})$.

However, comparison with equation (16.27) shows that $E_x = 0$ as long as $\omega / k_x = c$, i. e., as long as the photons travel at the speed of light, *c*.

Thus, virtual photons, i.e., those that have a non-zero mass and therefore travel at a speed other than that of light, can have a non-zero longitudinal component of the electric field, but real photons cannot.

The dot product of the electric and magnetic fields in a plane wave is $\mathbf{E}\cdot\mathbf{B} = 0$, as can be verified from equations (16.28) and (16.29).

This means that **E** and **B** are perpendicular to each other.

Furthermore, both **E** and **B** are perpendicular to the direction of wave motion for real photons.

Figure 16.9 shows the electric and magnetic fields for real photons in the special case where $A_z = 0$.

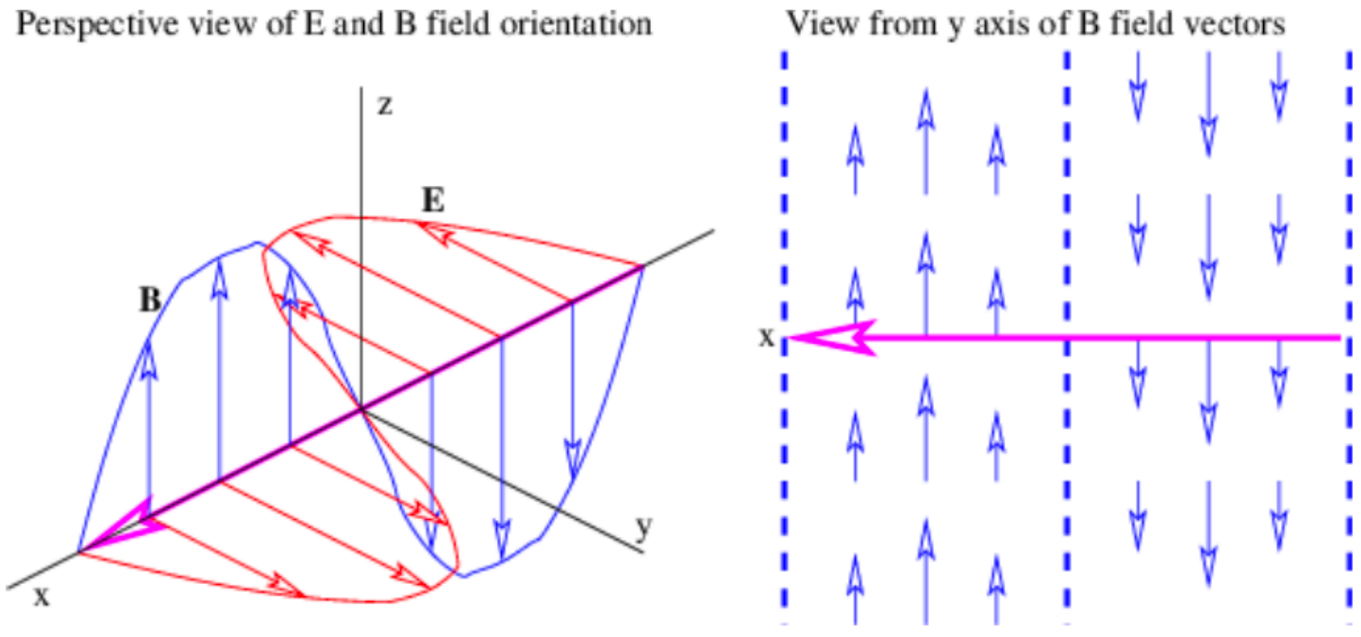


Figure 16.9: Electric and magnetic fields in a horizontally polarized plane wave, i. e., with $A_z = 0$, moving in the direction of the large arrow. The left panel shows how the electric and magnetic fields point, while the right panel shows the distribution of the magnetic field in space.

The electric field points in the same direction as the *transverse* part of the vector potential, while the magnetic field points in the other transverse direction.

The ratio of the magnitudes of the electric and magnetic fields is easily inferred from equations (16.28) and (16.29):

$$\frac{|\mathbf{E}|}{|\mathbf{B}|} = \frac{\omega(A_y^2 + A_z^2)^{1/2}\sin(k_x x - \omega t)}{k_x(A_z^2 + A_y^2)^{1/2}\sin(k_x x - \omega t)} = \frac{\omega}{k_x} = c. \tag{16.30}$$

Notice that the electric and magnetic fields for a wave do not depend on the longitudinal component of the vector potential, A_x.

This is because the Lorenz condition forces A_x to cancel with the term containing ϕ in the expression for E_x.

16.7 The Lorenz Condition

We are now in a position to see what the Lorenz condition means.

For an isolated stationary charge, the scalar potential is given by equation (16.1) and the vector potential **A** is zero.

The Lorenz condition reduces to

$$\frac{1}{c^2}\frac{\partial \phi}{\partial t} = \frac{1}{4\pi\epsilon_0 r c^2}\frac{dq}{dt} = 0. \tag{16.31}$$

From this we see that the Lorenz condition applied to the four-potential for a point charge is equivalent to the statement that *the charge on a point particle is conserved*, i. e., it doesn't change with time.

This is extended to any stationary distribution of charge by the superposition principle.

We thus see that the Lorenz condition is closely related to charge conservation for the four-potential of any charge distribution in the reference frame in which the charge is stationary.

If we can further show that the Lorenz condition is an equation that is equally valid in all reference frames, then we will have demonstrated that it is true for the four-potential produced by moving charged particles as well.

If the Lorenz condition is valid in one reference frame, it is valid in all frames for the special case of a plane electromagnetic wave.

This follows from substituting the four-potential for a plane wave into the Lorenz condition, as was done in equation (16.27) in the previous section.

In this case the Lorenz condition reduces to $\underline{k} \cdot \underline{a} = 0$.

Since the dot product of two four-vectors is a relativistic scalar, the Lorenz condition is equally valid in all frames.

Chapter 17

Capacitors, Inductors, and Resistors

Various electronic devices are considered in this chapter.

This is useful not only for understanding these devices but also for revealing new aspects of electromagnetism.

The capacitor is first discussed and Ampère's law is introduced.

The theory of magnetic inductance is then developed.

Ohm's law and the resistor are discussed.

The energy associated with electric and magnetic fields is calculated and Kirchhoff's laws for electric circuits are briefly discussed.

17.1 The Capacitor and Ampère's Law

We first discuss a device that is commonly used in electronics, called the capacitor.

We then introduce a new mathematical idea called the *circulation* of a vector field around a loop.

Finally, we use this idea to investigate Ampère's law.

17.1.1 The Capacitor

The *capacitor* is an electronic device for storing charge.

The simplest type is the parallel plate capacitor, illustrated in figure 17.1.

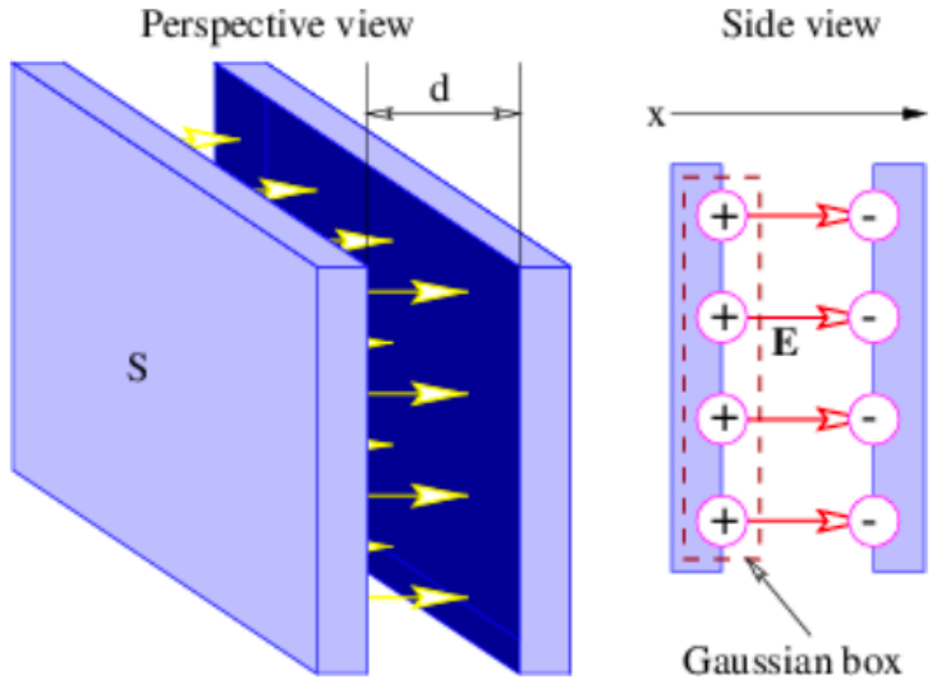


Figure 17.1: Two views of a parallel plate capacitor.

This consists of two conducting plates of area S separated by distance d, with the plate separation being much smaller than the plate dimensions.

Positive charge q resides on one plate, while negative charge -q resides on the other.

The electric field between the plates is $E = \sigma/\epsilon_0$, where the charge per unit area on the inside of the left plate in figure 17.1 is $\sigma = q/S$.

The density on the right plate is just -σ.

All charge is assumed to reside on the inside surfaces and thus contributes to the electric field crossing the gap between the plates.

The above formula for the electric field comes from applying Gauss's law to the sheet of charge on the positive plate.

The factor of $1/2$ present in the equation for an isolated sheet of charge is absent here because all of the electric flux exits the Gaussian surface on the right side — the left side of the Gaussian box is inside the conductor where the electric field is zero, at least in a static situation.

There is no vector potential in this case, so the electric field is related solely to the scalar potential φ.

Integrating $E_x = -\partial\varphi/\partial x$ across the gap between the conducting plates, we find that the potential difference between the plates is $\Delta\varphi = Ed = qd/(\epsilon_0 S)$, since E is known to be constant in this case.

This x equation indicates that the potential difference $\Delta\varphi$ is proportional to the charge q on the left plate of the capacitor in figure 17.1.

The constant of proportionality is $d/(\epsilon_0 S)$, and the inverse of this constant is called the *capacitance*:

$$C = \frac{\epsilon_0 S}{d} \quad \text{(parallel plate capacitor)}. \tag{17.1}$$

The relationship between potential difference, charge, and capacitance is thus

$$\Delta\phi = q/C \quad \text{or} \quad C = q/\Delta\phi. \tag{17.2}$$

The equation for the capacitance of the illustrated parallel plates contains just a fundamental constant (ϵ_0) and geometrical factors (area of plates, spacing between them), and represents the amount of charge the parallel plate capacitor can store per unit potential difference between the plates.

A word about signs: The higher potential is always on the plate of the capacitor that has the positive charge.

Note that equation (17.1) is valid only for a parallel plate capacitor.

Capacitors come in many different geometries and the formula for the capacitance of a capacitor with a different geometry will differ from this equation.

However, equation (17.2) is valid for *any* capacitor.

We now show that a capacitor that is charging or discharging has a magnetic field between the plates.

Figure 17.2 shows a parallel plate capacitor with a current i flowing into the left plate and out of the right plate.

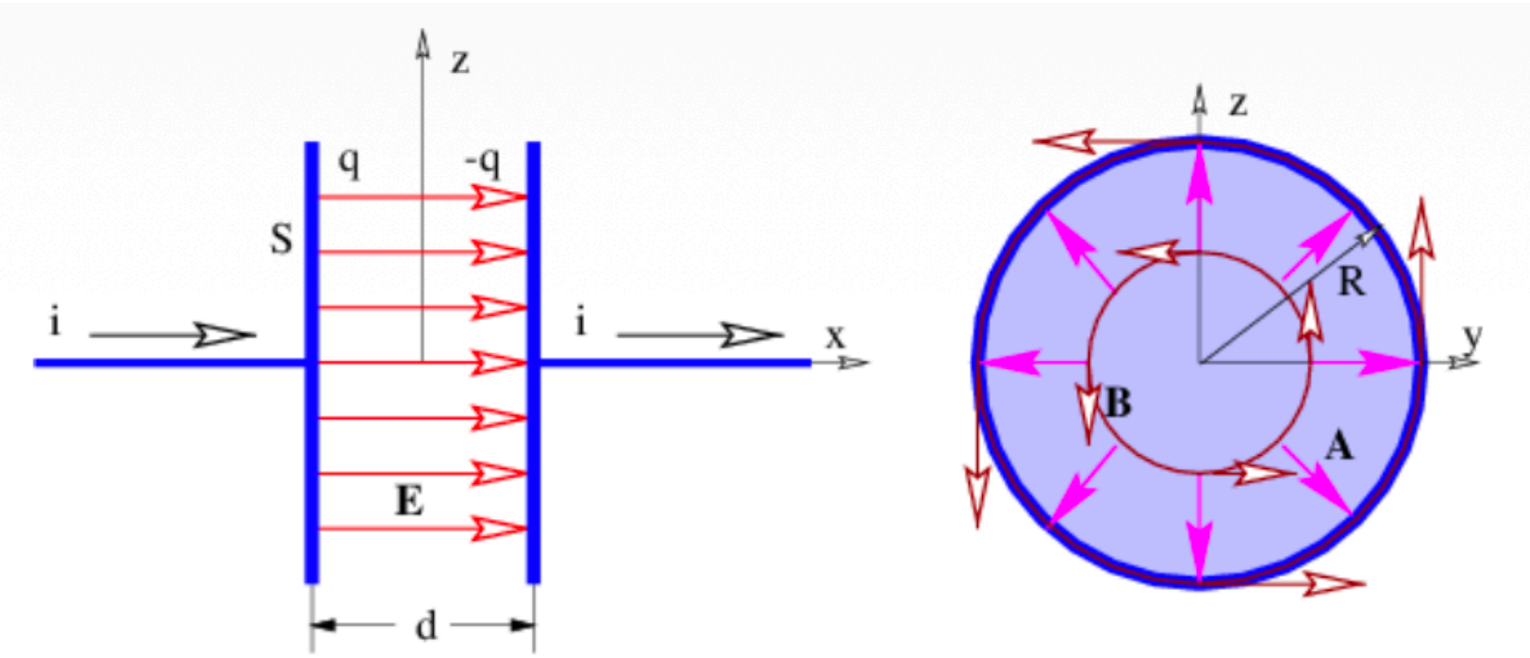


Figure 17.2: Parallel plate capacitor with circular plates in a circuit with current i flowing into the left plate and out of the right plate. The magnetic field that occurs when the charge on the capacitor is increasing with time is shown at right as vectors tangent to circles. The radially outward vectors represent the vector potential giving rise to this magnetic field in the region where $x > 0$. The vector potential points radially inward for $x < 0$. The y axis is into the page in the left panel while the x axis is out of the page in the right panel.

This current is necessarily accompanied by an electric field that is changing with time:
$E_x = q/(\epsilon_0 S) = it/(\epsilon_0 S)$.

Such an electric field can be derived from a scalar potential that is a function of time: $\varphi = -\, itx/(\epsilon_0 S)$.

However, the Lorenz condition

$$\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0 \qquad (17.3)$$

demands that some component of the vector potential **A** be non-zero under these circumstances, since $\partial\varphi/\partial t$ is non-zero.

How much can we infer about the vector potential from the geometry of the capacitor and equation (17.3)?

Substituting $\varphi = -itx/(\epsilon_0 S)$ into this equation results in

$$\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} = \frac{ix}{\epsilon_0 c^2 S}, \tag{17.4}$$

which suggests a number of different possibilities for **A**.

For instance, $\mathbf{A} = (0, ixy/(\epsilon_0 c^2 S), 0)$ and $\mathbf{A} = [0, 0, ixz/(\epsilon_0 c^2 S)]$ both satisfy equation (17.4).

However, neither of these trial choices is satisfactory by itself, as they are not consistent with the cylindrical symmetry of the capacitor about the *x* axis.

A choice of vector potential that is consistent with the shape of the capacitor and satisfies the Lorenz condition is obtained by combining these two trial solutions:

$$\mathbf{A} = [0, ixy/(2\epsilon_0 c^2 S), ixz/(2\epsilon_0 c^2 S)]. \tag{17.5}$$

This vector potential leads to the magnetic field

$$\mathbf{B} = [0, -iz/(2\epsilon_0 c^2 S), iy/(2\epsilon_0 c^2 S)]. \tag{17.6}$$

These fields are illustrated in the right-hand panel of figure 17.2.

17.1.2 Circulation of a Vector Field

We have already seen one example of the circulation[1] of a vector field, though we didn't label it as such.

In chapter 15 we computed the work done on a charge by the electric field as it moves around a closed loop in the context of the electric generator and Faraday's law.

The work done per unit charge, or the EMF, is an example of the *circulation* of a field, in this case the electric field, Γ_E.

Faraday's law can be restated as

$$\Gamma_E = -\frac{d\Phi_B}{dt} \quad \text{(Faraday's law)}. \tag{17.7}$$

In the simple case of a circular loop with the field directed along the loop, the circulation is just the magnitude of the field multiplied by the circumference of the loop, as illustrated in the left panel of figure 17.3.

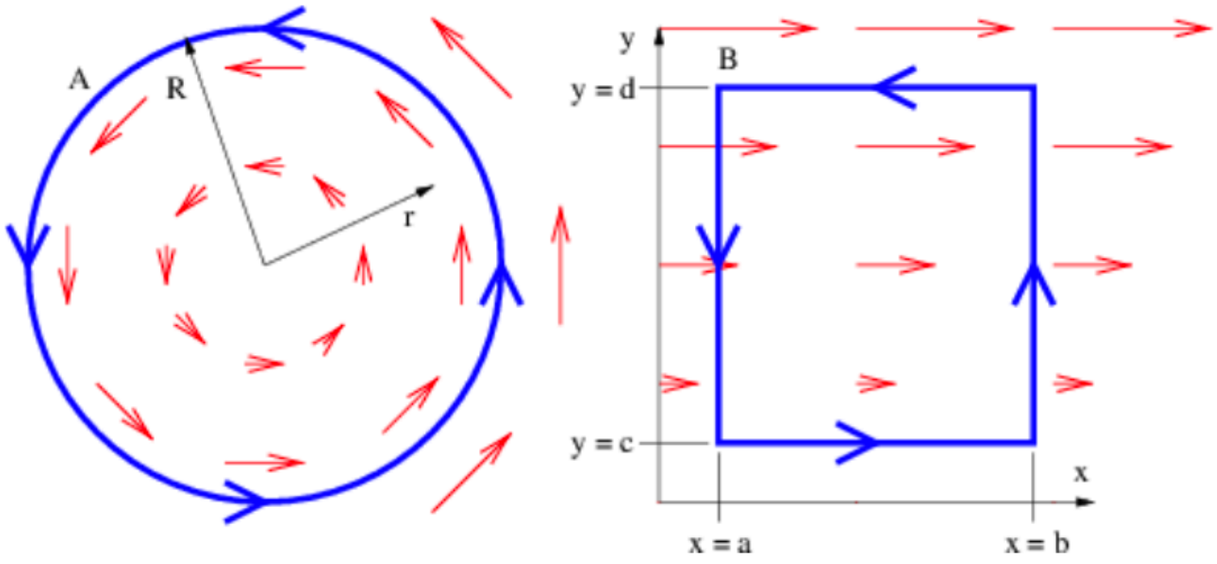


Figure 17.3: Two examples of circulation paths in a vector field.

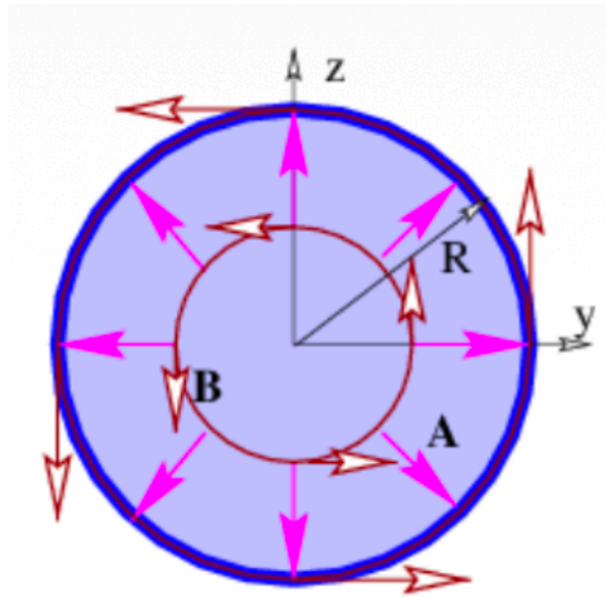
In more complicated cases in which the field points in a direction other than the direction of the loop, just the component in the direction of traversal around the loop enters the circulation.

If this component varies as one progresses around the loop, the calculation must be broken into pieces.

The total circulation is then obtained by adding up the contributions from segments of the loop in which the value of the field component parallel to the motion around the loop is constant.

An example of this type is the calculation of the EMF around a square loop of wire in an electric generator.

Another is illustrated in the right panel of figure 17.3.



17.1.3 Ampère's Law

The magnetic circulation Γ_B around the periphery of the capacitor in the right panel of figure 17.2 is easily computed by taking the magnitude of **B** in equation (17.6).

The magnitude of the magnetic field on the inside of the capacitor is just $B = ir/(2\epsilon_0 c^2 S)$, since $r = (y^2 + z^2)_{1/2}$ in figure 17.2.

Thus, at the periphery of the capacitor, $r = R$, and $B = iR/(2\epsilon_0 c^2 S))$ there.

The area of the capacitor plates is $S = \pi R^2$ and $\epsilon_0 c^2 = 1/\mu_0$, as we discussed previously.

Thus, the magnetic field is $B = \mu_0 i/(2\pi R)$ at the periphery.

If the periphery is traversed in the counter-clockwise direction, the magnetic circulation around the capacitor is $\Gamma_B = 2\pi RB = \mu_0 i$.

Let us now compute the magnetic circulation around a wire carrying a current.

The magnetic field a distance r from a straight wire carrying a current i is $B = \mu_0 i/(2\pi R)$.

The magnetic field points in the direction of a circle concentric with the wire.

The magnetic circulation around the wire is thus $\Gamma_B = 2\pi \mathrm{r} \mathrm{B} = \mu_0 i$.

Notice that the magnetic circulation is found to be the same around the wire and around the periphery of the capacitor.

Furthermore, this circulation depends only on the current in the wire and the constant μ_0.

One further item needs to be calculated, namely the *electric* flux across the gap between the capacitor plates.

This is just the electric field $E = \sigma/\epsilon_0$ multiplied by the area S, or $\Phi_E = S\sigma/\epsilon_0 = q/\epsilon_0$.

The current into the capacitor is the time rate of change on the capacitor, so $i = dq/dt = \epsilon_0 d\Phi_E/dt$.

We are now in a position to understand Ampère's law:

$$\Gamma_B = \mu_0 \left(i + \epsilon_0 \frac{d\Phi_E}{dt} \right) \qquad \text{(Ampère's law).} \tag{17.8}$$

This states that the magnetic circulation around a loop equals the sum of two contributions, (1) μ_0 multiplied by the electric current through the loop and (2) $\mu_0\epsilon_0$ multiplied by the time rate of change of the electric flux through the loop.

In the above example the first term dominates when the loop is around the wire, while the second term acts when the loop is around the gap between the capacitor plates.

Ampère actually formulated an incomplete version of the law named after him — he included only the first term containing the current.

The Scottish physicist James Clerk Maxwell added the second term, based primarily on theoretical reasoning.

Maxwell's additional term solved a serious internal inconsistency in electromagnetic theory — in our terms, the Lorenz condition *requires* a magnetic field to exist if the scalar potential ϕ is time-dependent.

This magnetic field is only predicted by Ampère's law if Maxwell's term is included.

The quantity $\epsilon_0 d\Phi_E/dt$ was called the *displacement current* by Maxwell since it has the dimensions of current and is numerically equal to the current entering the capacitor.

However, it isn't really a current — it is just an electric flux that changes with time!

Gauss's law for electricity and magnetism, Faraday's law, and Ampère's law are collectively called Maxwell's equations.

Together they form the basis for electromagnetism as it developed historically.

However, our formulation of electromagnetism in terms of the four-potential, the dispersion relation for free electromagnetic waves, the Lorenz condition, and Coulomb's law, is precisely equivalent to Maxwell's equations, and is much closer to the modern approach to electromagnetism.

17.2 Magnetic Induction and Inductors

Induction is the tendency of a current in a conductor to maintain itself in the face of changes in the potential difference driving the current.

Figure 17.4 shows a parallel plate *inductor* in which a current i passes through the two plates in opposite directions.

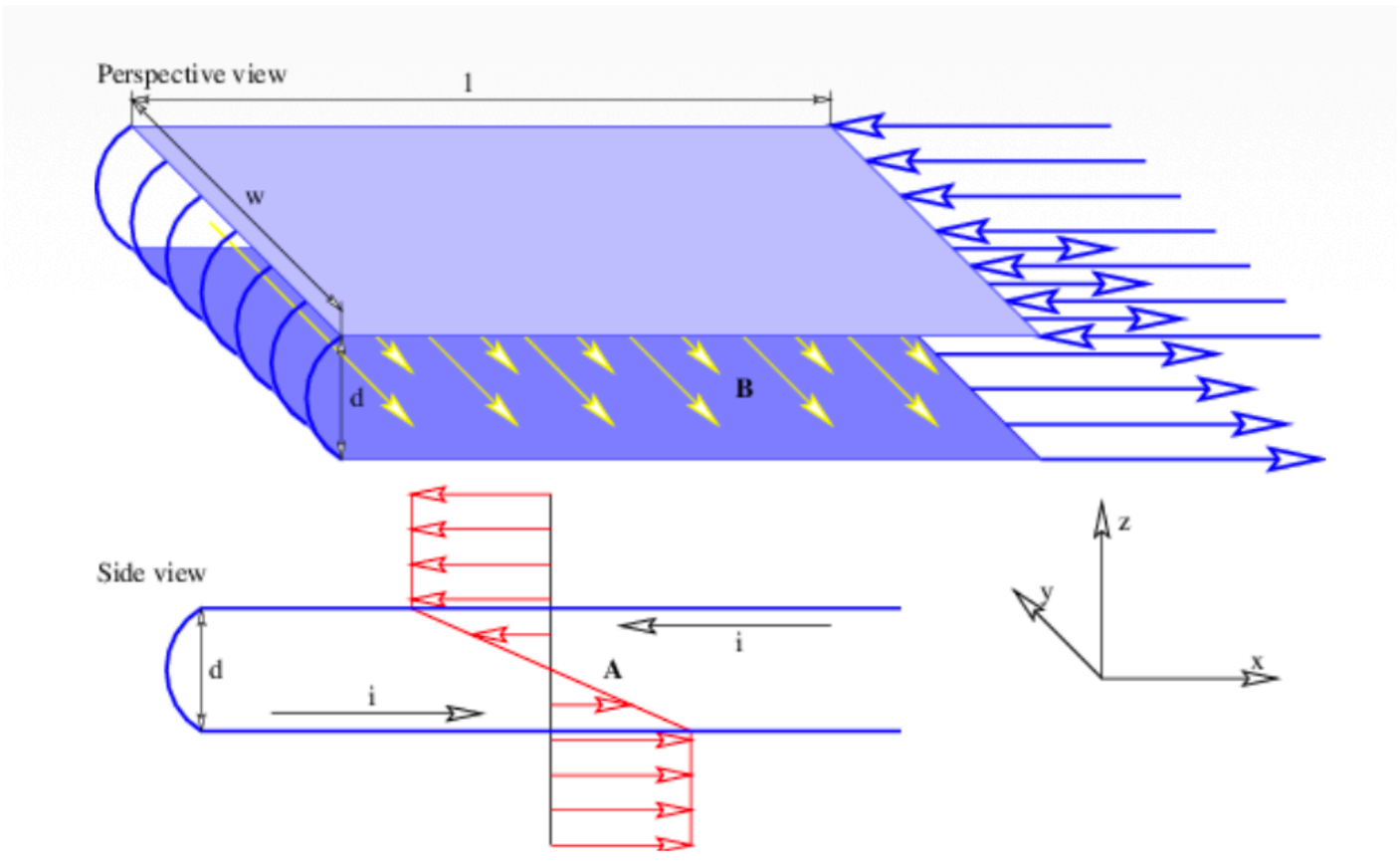


Figure 17.4: Magnetic field and vector potential for two parallel plates carrying equal currents in opposite directions. This is an example of an inductor.

The vector potential between plates of width w and spacing d is

$$\mathbf{A} = \left(-\frac{\mu_0 i z}{w}, 0, 0\right), \tag{17.9}$$

as long as $w \gg d$ (see figure 17.4).

Let us try to understand how this vector potential is constructed from what we already know.

The vector potential for a single current sheet in the x-y plane at $z = 0$ moving in the x direction was computed in the previous chapter as $A_x = -v\sigma|z|/(2\epsilon_0 c^2)$, with $A_y = A_z = 0$.

The quantity σ is the charge per unit area on the sheet and v is the velocity of the charge sheet in the x direction.

We use the relationship $1/(\epsilon_0 c^2) = \mu_0$ and also realize that if each plate has a width w, then the current in each plate is $i = v\sigma w$, which means that we can rewrite $A_x = -\mu_0 i|z|/(2w)$ for a single plate.

To proceed further, we first need to understand that $|z|$ in the above equation is only valid if the charge sheet is at $z = 0$.

If the sheet is located a distance a from the origin, then we must replace $|z|$ by $|z - a|$.

We also need to call on the definition of absolute value to realize that $|z - a| = z - a$ if $z > a$, and $|z - a| = -z + a$ if $z < a$.

Figure 17.5 shows how the profiles of Ax from each of the charge sheets add together to form a combined profile for the two sheets together.

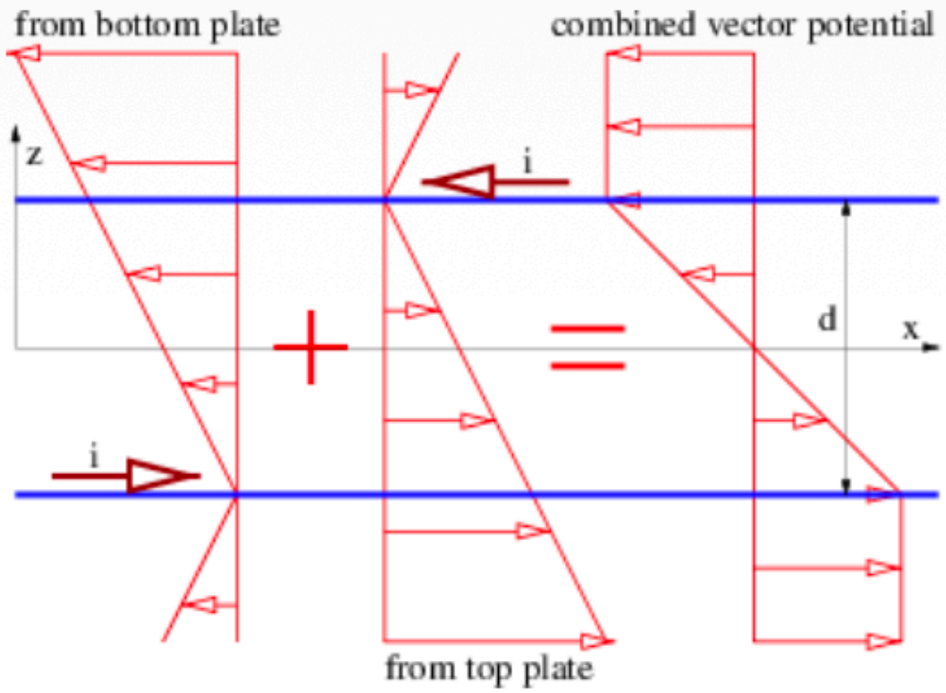


Figure 17.5: Illustration of the addition of the vector potentials from two current sheets with the left- moving current located above the x axis and the right-moving current below. The sum is obtained by the vector addition of the two components. Notice how the vector potential varies with z between the current sheets, but is constant outside of them.

The resulting magnetic field between the plates can be computed from the vector potential:

$$\mathbf{B} = \left(0, -\frac{\mu_0 i}{w}, 0\right). \tag{17.10}$$

Above and below the plates the magnetic field is zero because the vector potential is constant.

Let us now ask what happens when the current through the inductor increases or decreases with time.

Assuming initially that no scalar potential exists, the x component of the electric field in the device is

$$E_x = -\frac{\partial A_x}{\partial t} = \frac{\mu_0 z}{w}\frac{di}{dt}, \tag{17.11}$$

while $E_y = E_z = 0$.

Substituting the z values for each plate, we see that

$$
\begin{aligned}
E_{x-upper} &= +\frac{\mu_0 d}{2w}\frac{di}{dt} \quad \text{(upper plate)} \\
E_{x-lower} &= -\frac{\mu_0 d}{2w}\frac{di}{dt} \quad \text{(lower plate).}
\end{aligned}
\tag{17.12}
$$

The work done by this electric field on a unit charge moving from the right end of the upper plate, around the wire loops at the left end, and back to the right end of the lower plate is $\Delta V = E_{\text{x-upper}}(-l) + E_{\text{x-lower}}(+l) = -(\mu_0 dl/w)(di/dt)$, where l is the length of the plate, as illustrated in figure 17.4.

The minus sign means that the electric field acts so as to oppose a change in the current.

This result is called *Lenz's law*.

Lenz's law is not an independent law, but arises from the minus sign in the statement of Faraday's law.

In order for the current i to flow through the inductor, an external potential difference $\Delta\phi$ must be imposed between the input and output wires of the inductor, which just balances the effects of the internally generated electric field:

$$
\Delta\phi = \frac{\mu_0 ld}{w}\frac{di}{dt} \quad \text{(parallel plate inductor).} \tag{17.13}
$$

If this potential difference is positive, i. e., if the input wire of the inductor is at a higher potential then the output wire, then the current through the inductor will increase with time.

If it is lower, the current will decrease.

As with capacitors, inductors come in many shapes and forms.

The above equation is valid only for a parallel plate inductor, but the relationship

$$\Delta\phi = L\frac{di}{dt} \tag{17.14}$$

is valid for any inductor, assuming that the *inductance L* is known.

Comparison of the above two equations reveals that the inductance for the parallel plate inductor shown in figure 17.4 is just

$$L = \frac{\mu_0 l d}{w} \quad \text{(parallel plate inductor).} \tag{17.15}$$

17.3 Resistance and Resistors

Normal conducting materials require an electric field to keep an electric current flowing through them.

The electric field causes a force on the electrons in the material, which is balanced by the energy loss that occurs when the electrons collide with the atoms forming the material.

Most objects exhibit a linear relationship between the current *i* through them and the potential difference $\Delta\varphi$ applied to them.

This relationship is called *Ohm's law*,

$$\Delta\phi = iR \quad (R \text{ constant}), \tag{17.16}$$

where the constant of proportionality R is called the *resistance*.

The quantity $\Delta\varphi$ is sometimes called the *voltage drop* across the resistor.

For certain materials, such as semiconductors, the resistance depends on the current.

For such materials, the above equation defines resistance, but since the resistance doesn't remain constant when the current changes, these materials don't obey Ohm's law.

Figure 17.6 illustrates a rectangular resistor.

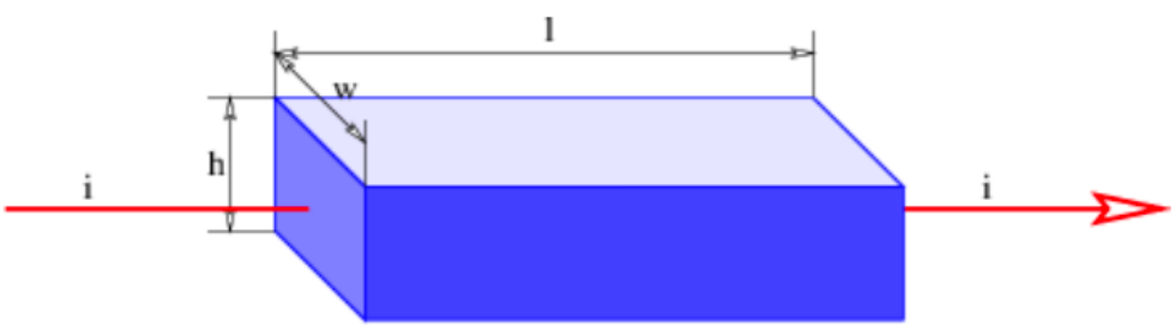


Figure 17.6: Rectangular resistor with a current i flowing through it.

The resistance of such a resistor can be written

$$R = \frac{l}{wh}\rho \tag{17.17}$$

where the *resistivity* ϱ is characteristic only of the material and not its shape or size.

Unlike capacitors and inductors, resistors are dissipative devices.

The work done on a charge q passing through a resistor is just $q\Delta\varphi$.

This energy is converted to heat.

The work done per unit time, which equals the power dissipated by a resistor is therefore

$$P = i\Delta\phi = i^2R = (\Delta\phi)^2/R. \tag{17.18}$$

17.4 Energy of Electric and Magnetic Fields

In this section we calculate the energy stored by a capacitor and an inductor.

It is most profitable to think of the energy in these cases as being stored *in the electric and magnetic fields* produced respectively in the capacitor and the inductor.

From these calculations we compute the energy per unit volume in electric and magnetic fields.

These results turn out to be valid for *any* electric and magnetic fields — not just those inside parallel plate capacitors and inductors!

Let us first consider a capacitor starting in a discharged state at time $t = 0$.

A constant current i is caused to flow through the capacitor by some device such as a battery or a generator, as shown in the left panel of figure 17.7.

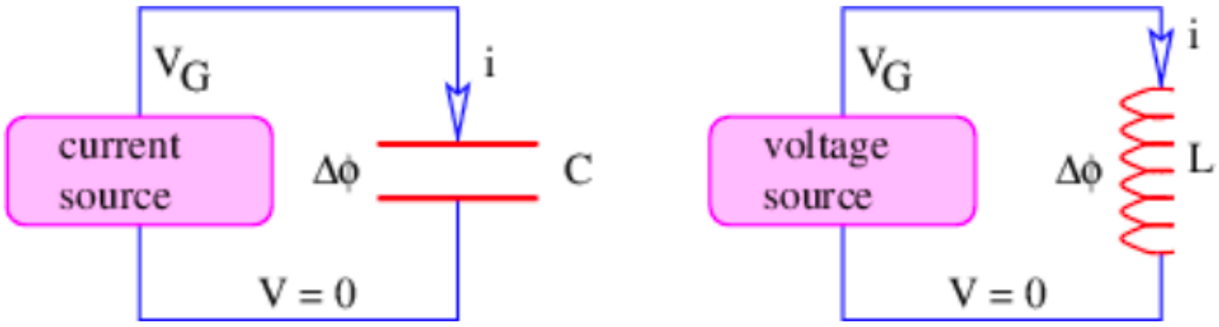


Figure 17.7: Capacitor (left) and inductor (right) being charged respectively by constant sources of current and voltage.

As the capacitor charges up, the potential difference across it increases with time:

$$\Delta\phi = \frac{q}{C} = \frac{it}{C}. \tag{17.19}$$

The EMF supplied by the generator has to increase to match this value.

The generator does work on the positive charges moving around the circuit in the direction indicated by the arrow.

We assume that $\Delta\varphi$ equals the EMF or work per unit charge done by the generator V_G, so the work done in time dt by the generator is $dW = V_G\, dq = V_G\, idt$.

Using the equation for the potential difference across a capacitor, we see that the power input is

$$P = \frac{dW}{dt} = \Delta\phi i = \frac{i^2 t}{C}. \tag{17.20}$$

Integrating this in time yields the total energy U_E supplied to the capacitor by the generator:

$$U_E = \frac{i^2 t^2}{2C} = \frac{q^2}{2C} \quad \text{(capacitor)}. \tag{17.21}$$

Assuming that we have a parallel plate capacitor, let's insert the formula for the capacitance of such a device, $C = \epsilon_0 S/d$.

Let us further recall that the electric field in a parallel plate capacitor is $E = \sigma/\epsilon_0 = q/(\epsilon_0 S)$, so that $q = \epsilon\epsilon_0 ES$ and

$$U_E = \frac{(E\epsilon_0 S)^2}{2(\epsilon_0 S/d)} = \frac{\epsilon_0 E^2 Sd}{2}. \tag{17.22}$$

The combination Sd is just the volume between the capacitor plates.

The *energy density* in the capacitor is therefore

$$u_E = \frac{U_E}{Sd} = \frac{\epsilon_0 E^2}{2} \quad \text{(electric energy density)}. \tag{17.23}$$

This formula for the energy density in the electric field is specific to a parallel plate capacitor.

However, it turns out to be valid for any electric field.

A similar analysis of a current increasing from zero in an inductor yields the energy density in a magnetic field.

Imagine that the generator in the right panel of figure 17.7 produces a constant EMF, V_G, starting at time $t = 0$ when the current is zero.

The work done by the generator in time dt is $dW = V_G\,dq = V_G\,i dt$ so that the power is

$$P = \frac{dW}{dt} = V_G i = Li\frac{di}{dt} = \frac{d}{dt}\left(\frac{Li^2}{2}\right). \tag{17.24}$$

We have assumed that the EMF supplied by the generator, V_G, balances the voltage drop across the inductor: $V_G = \Delta\phi = L(di/dt)$.

If we integrate the above equation in time, we get the energy added to the inductor as a result of increasing the current through it.

Substituting the formula for the inductance of a parallel plate inductor, $L = \mu_0 dl/w$, we arrive at the equation for the energy stored by the inductor:

$$U_B = \frac{Li^2}{2} = \frac{\mu_0 i^2 ld}{2w} \quad \text{(parallel plate inductor)}. \tag{17.25}$$

Finally, using the relationship between the current and the magnetic field in a parallel plate inductor, $B = \mu_0 i/w$, we can eliminate the current i and write

$$U_B = \frac{B^2 lwd}{2\mu_0}. \tag{17.26}$$

The volume between the inductor plates is just dlw, so again we can write an energy density, this time for the magnetic field:

$$u_B = \frac{U_B}{lwd} = \frac{B^2}{2\mu_0} \quad \text{(magnetic energy density)}. \tag{17.27}$$

Though we only proved this equation for the magnetic field inside a parallel plate inductor, it turns out to be true for any magnetic field.

The total energy density is just the sum of the electric and magnetic energy densities:

$$u_T = u_E + u_B = \frac{\epsilon_0 E^2}{2} + \frac{B^2}{2\mu_0}. \tag{17.28}$$

17.5 Kirchhoff's Laws

In the above discussion of energy we made two assumptions about electric circuits, which consist of electronic components connected by wires:

- Currents are conserved. Thus, the current entering one end of a wire connecting two devices is equal to the current leaving the other end, and the current out of the generator in the left panel of figure 17.7 is assumed to equal the current into and out of the capacitor. No electric charge is stored in any of the wires connecting components.

- The electric fields outside of components are conservative, and therefore are derived solely from the scalar potential. Thus, the net work done on a charged particle passing completely around a circuit loop is zero and the positive work done on charge passing through the generator in the right panel of figure 17.7 is exactly balanced by the potential difference across the inductor. As a result, the effects of electric fields outside of components can be represented by electrostatic potentials, or "voltages". Every part of a wire connecting components has the same potential and the effect of each component is to maintain a potential difference between its input and output connections.

These are called *Kirchhoff's laws*.

They are used extensively in electronic circuit design.

Figure 17.8 illustrates a typical circuit with a voltage source VS, which may be a battery or generator, and three circuit components.

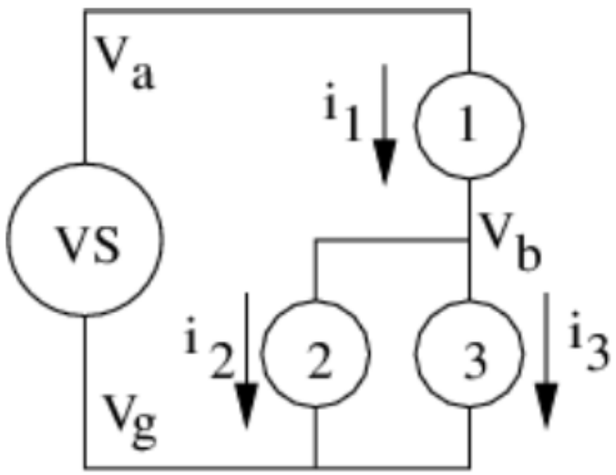


Figure 17.8: A typical circuit to which we apply Kirchhoff's laws.

The voltage source provides a potential difference equal to V_a - V_g, which is equal simply to V_a, since we are assuming that $V_g = 0$.

Since the scalar potential is insensitive to an arbitrary additive constant, we can always set the potential at one point in the circuit to zero to simplify our calculations.

The voltage drop across component 1 is V_a - V_b and the voltage drop across both components 2 and 3 is V_b - V_g.

If the components are resistors, then Ohm's law can be used to relate the voltage drops across the components to the currents through them.

If they are capacitors or inductors, the voltage drop is related respectively to the charge on the capacitor or the time rate of change of current through the inductor.

The time rate of change of the charge on the capacitor can be related to the current through the capacitor.

The final point in figure 17.8 is that current is conserved at junctions, i. e., $i_1 = i_2 + i_3$.

The methods of algebra (for just resistors) or calculus (if there are capacitors or inductors) can then be used to calculate all currents and voltages.

It is important to realize that Kirchhoff's laws are only approximations that hold when the currents and potentials in a circuit change slowly with time.

For steady currents and constant potentials they are precisely true, since imbalances in charge entering and leaving a junction between devices would result in the indefinite buildup of charge in the junction with time and therefore an increasing electrostatic potential, which would violate the steady state assumption.

Furthermore, a non-zero EMF around a closed loop would result in net acceleration of charge around the loop and a constantly increasing current.

If currents and potentials are changing with time, Kirchhoff's laws are approximately valid only if the capacitance, inductance, and resistance of the wires connecting circuit elements are much smaller than the capacitance, inductance, and resistance of the circuit elements themselves.

For very high frequency operation, the effects of these "parasitic" properties are not small and must be included in the design of the circuit.

Chapter 18

Measuring the Very Small

To begin our study of matter we discuss experiments in the late 19th and early 20th centuries that led to proof of the existence of atoms and their constituents.

We then introduce a fundamental idea about the scattering of waves using the diffraction of light by small particles as a prototype.

The famous Geiger-Marsden experiment that led to the idea of the atomic nucleus is discussed.

Finally, we examine some of the crucial experiments done with modern particle accelerators and the physical principles behind them.

18.1 Continuous Matter or Atoms?

From the time of the ancient Greeks there have been debates about the ultimate nature of matter.

One of these debates is whether matter is infinitely divisible or whether it consists of fundamental building blocks that are themselves indivisible.

However, it wasn't until the late 19th century that real progress began to be made on this question.

Advancements in our understanding of matter have largely been coupled to the development of machines to accelerate atomic and sub-atomic particles.

The original accelerator was developed in the 19th century and is called the *Crookes tube (Figure 18.1)*.

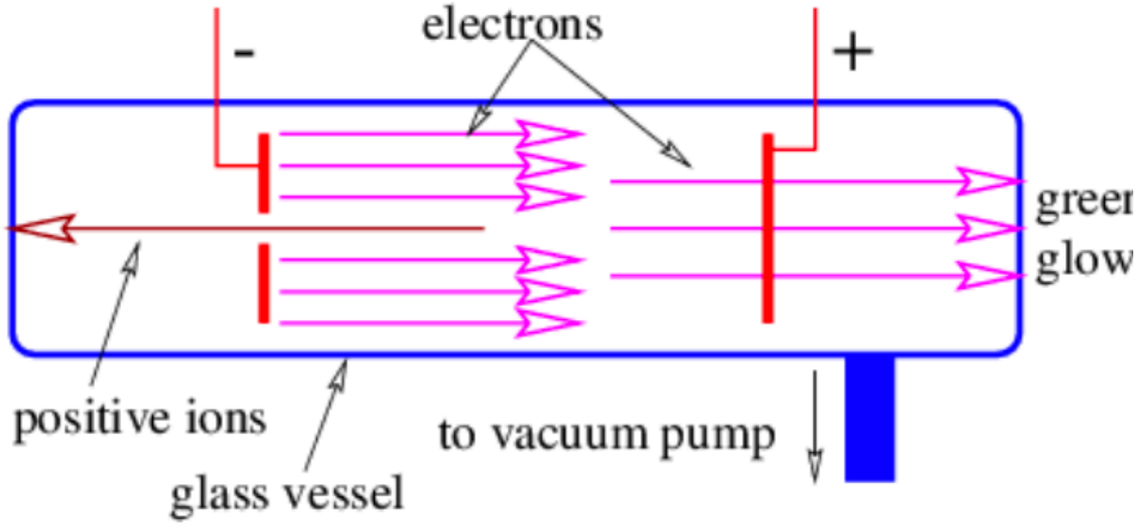


Figure 18.1: Crookes tube, the original particle accelerator. When potentials are applied to the plates as shown, electrons are emitted by the left electrode and accelerated to the right, some of which pass through holes in the right electrode. Positive ions, which are atoms missing one or more electrons, are created by collisions between electrons and residual gas atoms. These are accelerated to the left.

J. J. Thomson measured the charge to mass ratio for both electrons and positive ions in the Crookes tube in the following way: If a potential difference $\Delta\phi$ is applied between the electrodes, then by energy conservation a particle of charge q starting from rest will acquire a kinetic energy moving from electrode to electrode of $K = mv^2/2 = q\Delta\phi$.

Solving for v, we find $v = (2q\Delta\phi/m)^{1/2}$.

If a magnetic field B is then imposed normal to the electron beam after it has passed the positive electrode, the beam bends with a radius of curvature of $R = mv/(qB)$.

Since R and B are known, the charge to mass ratio can be computed by eliminating v and solving for q/m: $q/m = 2\Delta\phi/(BR)^2$.

Thomson found that positive ions typically had charge to mass ratios several thousand times smaller than the electrons.

Furthermore, the ions were positively charged, while the electrons were negatively charged.

If the ions and the electrons have electrical charges equal in magnitude (plausible, since the ions are neutral atoms with at least one electron removed), the ions have to be much more massive than the electrons.

Robert Millikan made the first direct measurement of electric charge.

He did this by suspending electrically charged oil drops in a known electric field against gravity.

The size of an oil drop is directly measured using a microscope, leading to a calculation of its mass, and hence the gravitational force, mg.

This is then balanced against the electric force, qE, leading to $q = mg/E$.

Occasionally an oil drop loses an electron due to photoelectric emission caused by photons from an ultraviolet lamp.

This disrupts the force balance, and causes the oil drop to move up or down.

If the electric field is quickly adjusted, this motion can be arrested.

The *change* in the charge can be related to the *change* in the electric field: $\Delta q = mg\Delta(1/E)$.

If only a single electron is emitted, then Δq is equal to the electronic charge.

Between the work of Thomson and Millikan, the masses and the charges of sub-atomic particles were accurately measured for the first time.

Ironically, this work also showed that the “atom”, which means “indivisible” in Greek, in fact isn’t.

Atoms consist of positive charges with large mass, or protons, in conjunction with low mass electrons of negative charge.

Electrons and protons have opposite charges, so they attract each other to form atoms in this picture.

Geiger and Marsden did an experiment that strongly suggested that atoms consist of very small positively charged atomic nuclei, surrounded by a cloud of circling, negatively charged electrons.

This is called the *Rutherford model* of the atom after Ernest Rutherford.

Chadwick completed our picture of the atom with the discovery of a neutral particle of mass comparable to the proton, called the neutron.

The neutron is a constituent of the atomic nucleus along with the proton.

The number of protons in a nucleus is denoted Z while the number of neutrons is N.

We define $A = Z + N$ to be the total number of *nucleons* (protons plus neutrons). The parameter Z is often called the *atomic number* while A is called the *atomic mass number*.

Marie and Pierre Curie and Henri Becquerel were the first to discover a more fundamental divisibility of atoms in the form of the radioactive decay, though the implications of their results did not become clear until much later.

Radioactive decay of atomic nuclei comes in three common forms: alpha, beta, and gamma decay.

Alpha decay is the spontaneous emission of a helium-4 nucleus, called an *alpha particle* by a heavy nucleus such as uranium or radium.

The alpha particle consists of two protons and two neutrons, so the emission decreases both Z and N by 2.

Beta decay is the emission of an electron or its antiparticle, the positron, by a nucleus, with an accompanying change in the electric charge of the nucleus.

For electron emission Z increases by 1 while N decreases by 1.

The opposite occurs for positron emission.

Gamma decay is the emission of a high energy photon by a nucleus.

The values of Z and N remain unchanged.

The energy released by these decays is typically of order a few million electron volts.

Of the three forms of decay, beta decay is the most interesting, since it involves the transformation of one sub-atomic particle into another.

In the case of neutron decay, a neutron is converted into a proton, an electron, and an antineutrino.

For proton decay, a proton becomes a neutron, a positron, and a neutrino.

Only the neutron form occurs for an isolated particle.

However, the energetics inside atomic nuclei can result in either form, depending on the nucleus in question.

The neutrino is one of the great theoretical predictions of modern physics.

Careful studies of beta decay, which at the time was thought to result only in the emission of a proton and an electron for the neutron form of the reaction, showed apparent non-conservation of energy and angular momentum.

Rather than accept this rather unpalatable conclusion, Wolfgang Pauli proposed that a third particle named a neutrino, or little neutral particle, is emitted in the decay, thus accounting for the missing energy and angular momentum.

The presumed electrical neutrality of the particle explained the difficulty of detecting it.

Over 25 years passed before Frederick Reines and Clyde Cowan from Los Alamos observed this elusive particle.

The three forms of radioactive decay are associated with three of the four known fundamental forces of nature.

Gamma decay is electromagnetic in nature, while alpha decay involves the breaking of bonds produced by the *nuclear* or *strong* force.

Beta decay is a manifestation of the so-called *weak* force.

The fourth force is gravity, which plays a negligible role on the sub-atomic scale, as far as we know.

Beta decay gives us a strong hint that even particles such as protons and neutrons, which make up atomic nuclei, are not "atomic" in the sense of the original Greek, since neutrons can change into protons in beta decay and vice versa.

We now have excellent evidence that protons, neutrons, and many other sub-nuclear particles are made up of particles called *quarks*.

Quarks and electrons are currently thought to be fundamental in that they are supposedly indivisible, and are hence the true "atoms" of the universe.

However, who knows, perhaps someday we will discover that they too are composed of even more fundamental constituents!

18.2 The Ring Around the Moon

Sometimes at night one sees a diffuse disk of light around the moon if it happens to be shining through a thin layer of cloud.

This disk consists of light diffracted by the water or ice particles in the cloud.

The diameter of the disk contains information about the size of the cloud particles doing the diffraction.

In particular, if the particles have diameter d and the light has wavelength λ, then the diffraction half-angle shown in figure 18.2 is approximately

$$\alpha \approx \lambda/(2d). \tag{18.1}$$

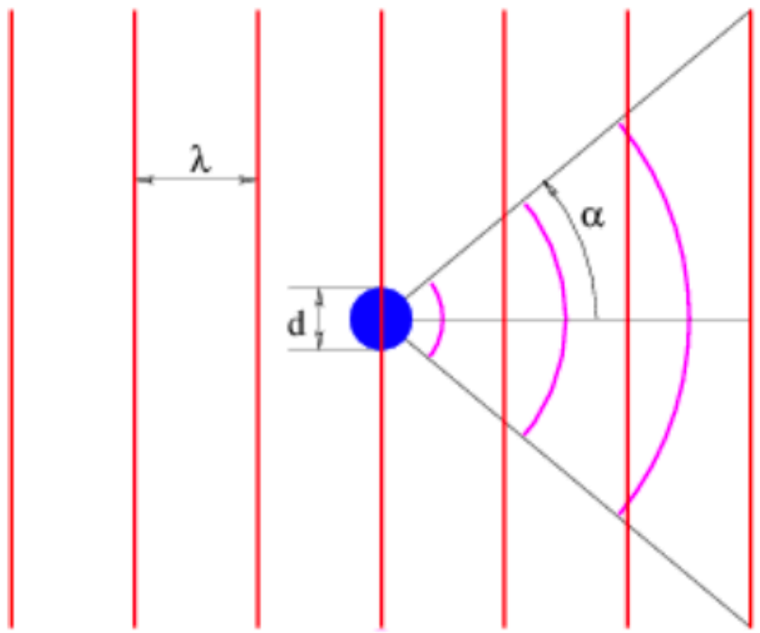


Figure 18.2: Scattering of an incident plane wave by a water droplet. The opening half-angle of the scattered wave is $\alpha \approx \lambda/(2d)$.

This equation comes from the problem of passage of light through a hole or slit of diameter or width d.

This problem was treated in the chapter on waves, and the above formula was concluded to hold in that case.

One can think of the diffraction of light by a particle to be the linear superposition of a plane wave minus the diffraction of light by a hole in a mask, as illustrated in figure 18.2.

The angular spread of the diffracted light is the same in both cases.

The interesting point about equation (18.1) is that the opening angle of the diffraction cone is inversely proportional to the diameter of the diffracting particles.

Thus, for a given wavelength, smaller particles cause diffraction through a wider angle.

Note that when the wavelength exceeds the diameter of the particle by a significant amount, equation (18.1) fails, since scattering through an angle greater than π doesn't make physical sense.

In this case the diffracted photons tend to be isotropic, i. e., they are scattered with equal probability into any direction.

If one wishes to measure the size of an object by observing the diffraction of a wave around the object, the lesson is clear; the wavelength of the wave must be less than or equal to the dimensions of the object — otherwise the scattering of the wave by the object is largely isotropic and equation (18.1) yields no information.

Since wavelength is inversely related to momentum by the de Broglie relationship, this condition implies that the momentum must satisfy

$$p = h/\lambda > h/d \tag{18.2}$$

in order that the size of an object of diameter d be resolved.

18.3 The Geiger-Marsden Experiment

In 1908 Hans Geiger and Ernest Marsden, working with Ernest Rutherford of the Physical Laboratories at the University of Manchester, measured the angular distribution of alpha particles scattered from a thin gold foil in an experiment illustrated in figure 18.3.

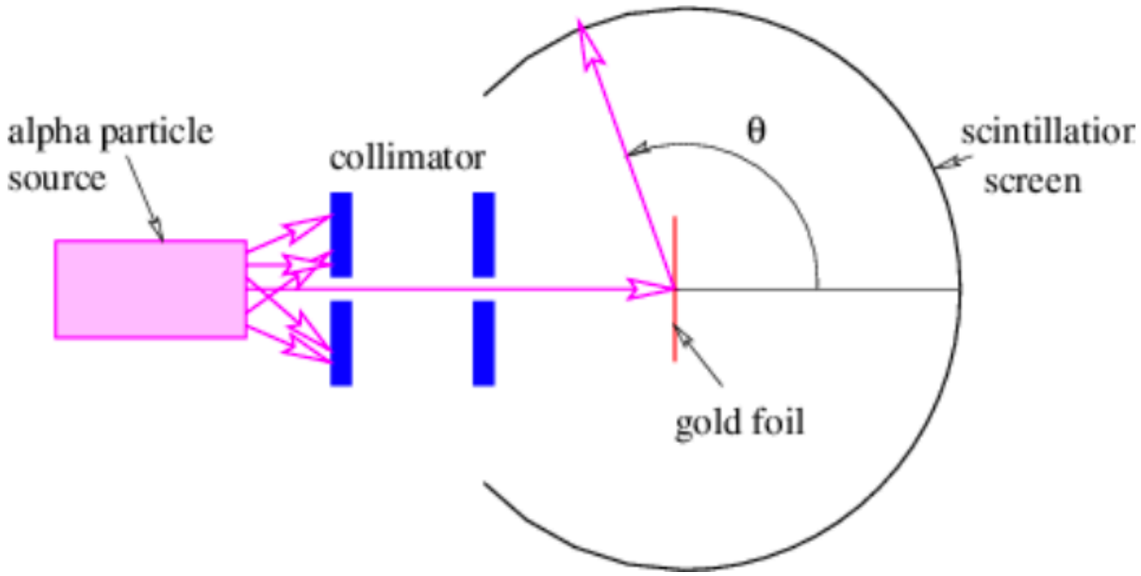


Figure 18.3: Schematic of Geiger-Marsden experiment. The radioactive source produces alpha particles that are collimated into a beam and directed at a gold foil. The alpha particles scatter off the foil and are detected by a flash of light when they hit the scintillation screen.

In order to understand this experiment, we need to compute the de Broglie wavelength of alpha particles resulting from radioactive decay.

Typical alpha particle kinetic energies are of order 5 MeV = 8×10^{-13} J.

Since the alpha particle consists of two protons and two neutrons, its mass is about $M_\alpha = 6.7 \times 10^{-27}$ kg.

This implies a velocity of about $v = 1.1 \times 10^7$ m s^{-1} , a momentum of about $p = mv = 7.4 \times 10^{-20}$ N s, and a de Broglie wavelength of about $\lambda = h/p = 9.0 \times 10^{-15}$ m.

Other evidence indicates that atoms have dimensions of order 10^{-10} m, so the de Broglie wavelength of an alpha particle is about a factor of 10^4 smaller than a typical atomic dimension.

Thus, the typical diffraction scattering angle of alpha particles off of atoms ought to be very small, of order $\alpha = \lambda/(2d) \approx 10^{-4}$ radian $\approx 0.01^o$.

Imagine the surprise of Geiger and Marsden when they found that while most alpha particles suffered only small deflections when passing through the gold foil, a small fraction of the incident particles scattered through large angles, some in excess of 90°!

Ernest Rutherford calculated the probability for an alpha particle, considered to be a positive point charge, to be scattered through various angles by a stationary atomic nucleus, assumed also to be a positive point charge.

The calculation was done classically, though interestingly enough a quantum mechanical calculation gives the same answer.

The relative probability for scattering with a momentum transfer to the alpha particle of **q** is proportional to $|\mathbf{q}|^{-4} \equiv q^{-4}$ according to Rutherford's calculation.

Do not confuse this q with charge!

As figure 18.4 indicates, a larger momentum transfer corresponds to a larger scattering angle.

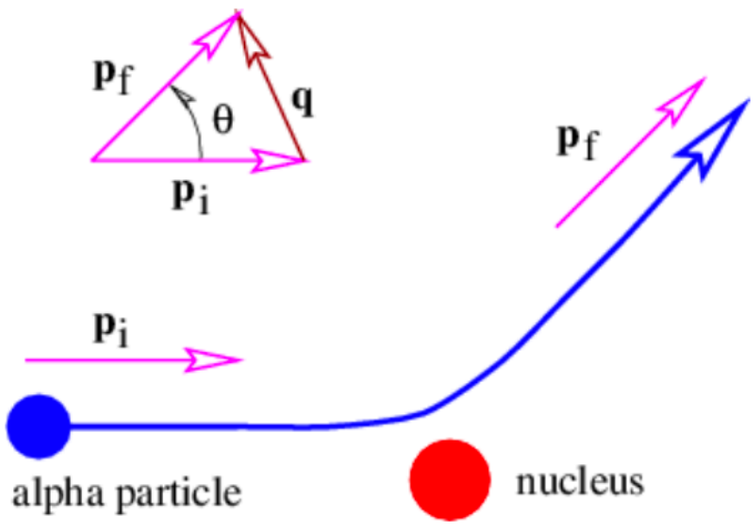


Figure 18.4: Illustration of alpha particle trajectory in Rutherford's model of the atom. The momentum transfer **q** from the nucleus to the alpha particle is equal to the change in the alpha particle's momentum.

The maximum momentum transfer for an incident alpha particle with momentum **p** is 2|**p**|, or just twice the initial momentum.

This corresponds to a head-on collision between the alpha particle and the nucleus followed by a recoil of the alpha particle directly backwards.

Since this collision is elastic, the kinetic energy of the alpha particle after the collision is approximately the same as before, as long as the nucleus is much more massive than the alpha particle.

Rutherford's calculation agreed quite closely with the experimental results of Geiger and Marsden.

Though the probability for scattering through a large angle is small even in the Rutherford theory, it is still much larger than would be expected if there were no small scale atomic nucleus.

18.4 Cosmic Rays and Accelerators

18.4.1 Early Cosmic Ray Results

Earlier we indicated that particles interacted with each other via the exchange of a virtual intermediary particle that interchanges energy, momentum, and other physical properties between the interacting particles.

This idea originated with the Japanese physicist Hideki Yukawa in 1935 in an effort to understand the forces between nucleons.

Yukawa hypothesized that the force that holds nucleons together is associated with the exchange of a boson, i. e., a particle with integer spin, with rest energy $mc_2 \approx 100$ MeV.

The range of this force at low momentum transfers is $I \approx \hbar/(mc) \approx 2 \times 10^{-15}$ m, or comparable to the observed size of an atomic nucleus.

In 1947 two new particles were discovered in cosmic rays, the negatively charged muon with a rest energy of 106 MeV, and the pion, which comes in three varieties, the π^+, the π^-, and the π^0, which respectively have positive, negative, and zero charge.

The rest energies of the π^+ and π^- are 140 MeV while that of the π^0 is 135 MeV.

All of these particles are unstable in that they decay into other, more stable particles in a tiny fraction of a second.

In particular, the negative pion decays into a muon and an antineutrino, while the neutral pion decays into two gamma rays, or high energy photons.

The antineutrino that results from pion decay is actually distinct from the antineutrino emitted in nuclear beta decay; it is called the *mu antineutrino* since it is associated with the muon in the same way that the antineutrino in beta decay is associated with the electron.

To further distinguish between the two, the latter is called the *electron antineutrino*.

The muon itself decays into an electron, a mu neutrino, and an electron antineutrino.

The muon and its associated neutrino are rather peculiar.

In all respects except mass, the muon appears to be identical to the electron.

The physicist I. I. Rabi is reputed to have responded “Who ordered that?” upon learning of the properties of the muon.

Furthermore, the electron neutrino only interacts with the electron and the muon neutrino only interacts with the muon.

This is the first hint that elementary particles occur in families that appear to be replicated at higher energies.

Since the muon is a fermion with spin $1/2$, it can’t be Yukawa’s intermediary particle since all intermediary particles are bosons with integral spin.

Furthermore, as with the electron, it is not subject to the nuclear force.

The pions are more promising candidates for being intermediary particles of the nuclear force, since they are bosons with spin 0.

However, as we shall see, the situation is more complex than Yukawa imagined, and the force between nucleons cannot be so simply treated.

However, Yukawa’s idea of intermediary particle exchange lives on in today’s theories of sub-nuclear particles.

18.4.2 Particle Accelerators

Soon after the discovery of muons and pions in cosmic rays, a whole plethora of unstable particles was uncovered.

Central to these discoveries was the particle accelerator.

In these devices, charged particles, typically electrons or protons, are accelerated to high energy and then smashed into a target.

Detectors of various sorts are used to examine the particles created by the collisions of the accelerated particles and the atomic nuclei with which they collide.

Sometimes an *elastic collision* occurs, in which the accelerated particle simply "bounces off" of the target particle, transferring a good bit of its momentum to this particle.

However, under many circumstances the collision results in the production of new particles that didn't exist before the collision.

This is referred to as an *inelastic collision*.

The simplest type of target is liquid hydrogen since the nucleus consists of a single proton.

The orbital electrons of the target atoms are so light that they are generally just "brushed aside" without greatly affecting the trajectories of the accelerated particles.

However, a variety of targets are used under different circumstances.

18.4.3 Size and Structure of the Nucleus

In the late 1950s and early 1960s Robert Hofstadter of Stanford University extended the Geiger-Marsden experiment to much shorter de Broglie wavelengths using high energy electrons from an accelerator, rather than alpha particles, as the probe.

The type of results obtained by Hofstadter are shown in figure 18.5.

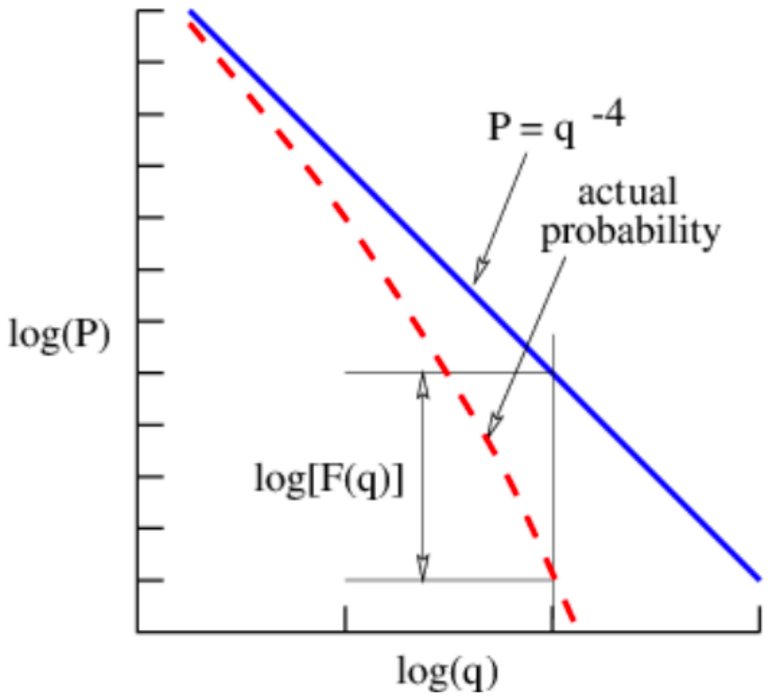


Figure 18.5: Schematic illustration of Robert Hofstadter's results for scattering of electrons off of atomic nuclei. The solid line shows the relative probability (in log-log coordinates) of elastic scattering as a function of the momentum transfer. The dashed curve illustrates the observed probability distribution. The difference between the curves is the logarithm of the form factor, $F(q)$.

After accounting for some effects having to do with the electron spin, these experiments should agree with the Rutherford formula if the nucleus is truly a point particle.

However, the actual results show probabilities that drop off more rapidly with increasing momentum transfer q than is predicted by the Rutherford model.

The ratio of the actual to the Rutherford probability distributions is called the *form factor*, $F(q)$, for this process:

$$P_{obs}(q) = F(q)P_{Ruth}(q) \propto F(q)q^{-4}. \tag{18.3}$$

Taking the logarithm of this equation results in

$$\log[P_{obs}] = \log[F(q)] - 4\log(q) + const. \tag{18.4}$$

These results are related to the fact that the nucleus is actually of finite size.

The diffraction effects discussed in the section on the scattering of moonlight come into play here, in that little scattering takes place for scattering angles larger than roughly $\lambda/(2d)$, where λ is the de Broglie wavelength of the probing particle and d is the diameter of the target.

For a small scattering angle (which we now call θ), it is clear from figure 18.4 that

$$\theta \approx q/p, \tag{18.5}$$

where p is the momentum of the incident electron and q is the momentum transfer.

If q_{max} is the maximum momentum transfer for which there is significant scattering, then we can write

$$q_{max}/p = \theta_{max} \approx \lambda/d, \tag{18.6}$$

where the factor of 2 in the denominator on the right side has been dropped since this is an approximate analysis.

However, since $\lambda = h/p$, we find that

$$q_{max} \approx \frac{h}{d}. \tag{18.7}$$

Thus, the momentum transfer for which the measured form factor becomes small compared to one gives us an immediate estimate of the diameter of an atomic nucleus: $d \approx h / q_{max}$.

The results obtained by Hofstadter show that nuclear diameters are typically a few times 10^{-15} m.

More than just size information can be extracted from the form factor.

Hofstadter's experiments also led to a great deal of information about the internal structure of atomic nuclei.

18.4.4 Deep Inelastic Scattering of Electrons from Protons

TheThe construction of the Stanford Linear Accelerator Center (SLAC), which accelerates electrons up to 40 GeV, allowed experiments like Hofstadter's to be carried out at much higher energies.

At these energies, many of the collisions between electrons and protons and neutrons are inelastic — generally a great mess of short-lived particles is spewed out, and are very difficult to interpret.

However, the so-called *deep inelastic* collisions, where the electron scatters through a large angle and therefore transfers a large momentum, q, to the proton, yield very interesting results.

In particular, these collisions occur essentially with a probability proportional to q^{-4} — just as in the Geiger-Marsden experiment!

The electron is a point particle as far as we know.

However, previous experiments showed the proton to have a finite size, of order 10^{-15} m.

Therefore, the scattering probability should drop off more rapidly with increasing momentum transfer q than with q-4, as in the earlier Hofstadter experiments.

James Bjorken and Richard Feynman showed a way out of this dilemma.

They proposed that the proton actually consists of a small number of point particles bound together by weakly attractive forces.

A sufficiently energetic photon is able to knock a single one of these particles out of the proton, as illustrated in the right panel of figure 18.6.

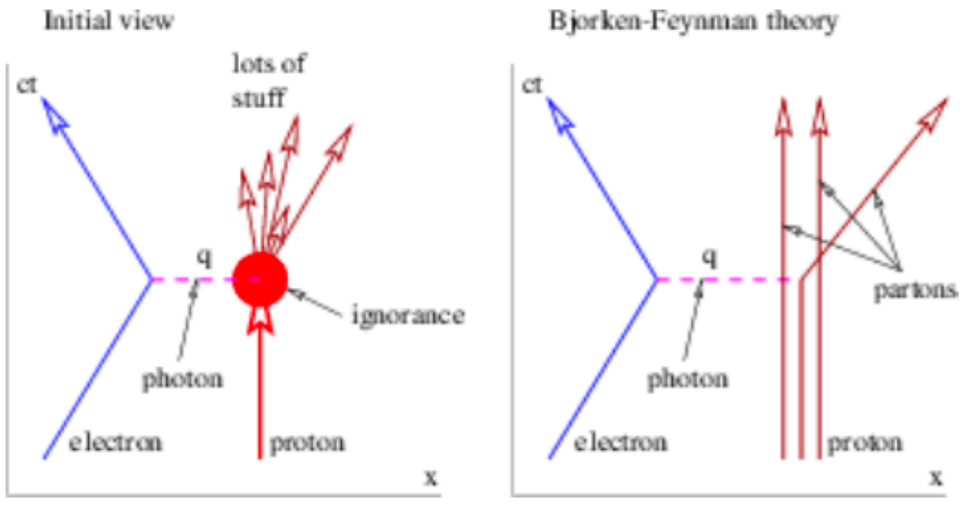


Figure 18.6: Deep inelastic scattering of a high energy electron by a proton occurs when the momentum transfer q is large and many particles are produced. According to the Bjorken- Feynman theory of this process, the proton consists of a number of partons flying in "loose formation" . A sufficiently energetic photon, i. e., with large momentum transfer q, kicks out just one of these partons, leaving the others undisturbed.

This leads to a subsequent set of reactions that produce the profusion of particles seen in the left panel of this figure.

Feynman called the particles that make up the proton *partons*.

However, we now know that they are actually *quarks*, spin 1/2 particles with fractional electronic charge that are thought to be the fundamental building blocks of matter, and *gluons*, the massless spin 1 intermediary particles that carry the strong force.

18.4.5 Storage Rings and Colliders

An alternate way to create interesting collisions is to crash particles and antiparticles of the same energy into each other.

This is done via a storage ring, as shown in figure 18.7.

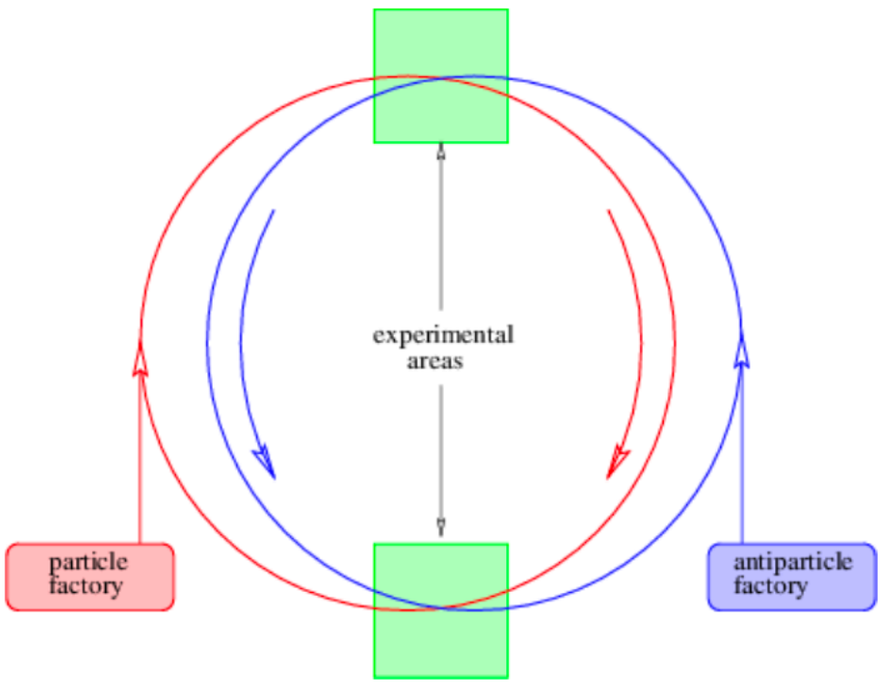


Figure 18.7: Schematic model of a particle-antiparticle collider. The particles and antiparticles are injected into the storage rings shown and are made to go in a circle by magnetic fields. The beams cross at two points and equipment is set up around these points to observe the products of collisions.

A set of magnets forces particles and antiparticles (which have opposite charges) to move in opposing circles within a high vacuum.

The circles are slightly offset so that the beams cross at only two points.

Collisions occur at these points and are observed by various types of experimental equipment.

An alternate type of collider has two storage rings that intersect at only one point.

This type of system can be used to collide particles of the same type together, e. g., protons colliding with protons.

18.4.6 Proton-Antiproton Collisions

If collisions occur by the exchange of a single intermediary particle of zero mass between point particles, the q^{-4} dependence of the collision probability on momentum transfer will occur in proton-antiproton collisions as in the Geiger-Marsden experiment.

However, if the colliding particles are not point particles, a form factor that decreases for increasing momentum transfer will occur as with the Hofstadter experiments.

When collisions between protons and antiprotons of a few hundred GeV are arranged, certain types of events called *two-jet events* are recorded.

In these events, two jets, each containing many particles, are emitted in opposite directions at wide angles (i. e., with large momentum transfer) from the colliding beams.

Furthermore, these jets show a probability distribution as a function of momentum transfer very close to q^{-4}.

This indicates that the colliding particles are point-like, at least down to the minimum spatial resolutions available to today's accelerators.

According to the Bjorken-Feynman parton model of the proton, the collision between highly energetic protons and antiprotons should operate as shown in figure 18.8.

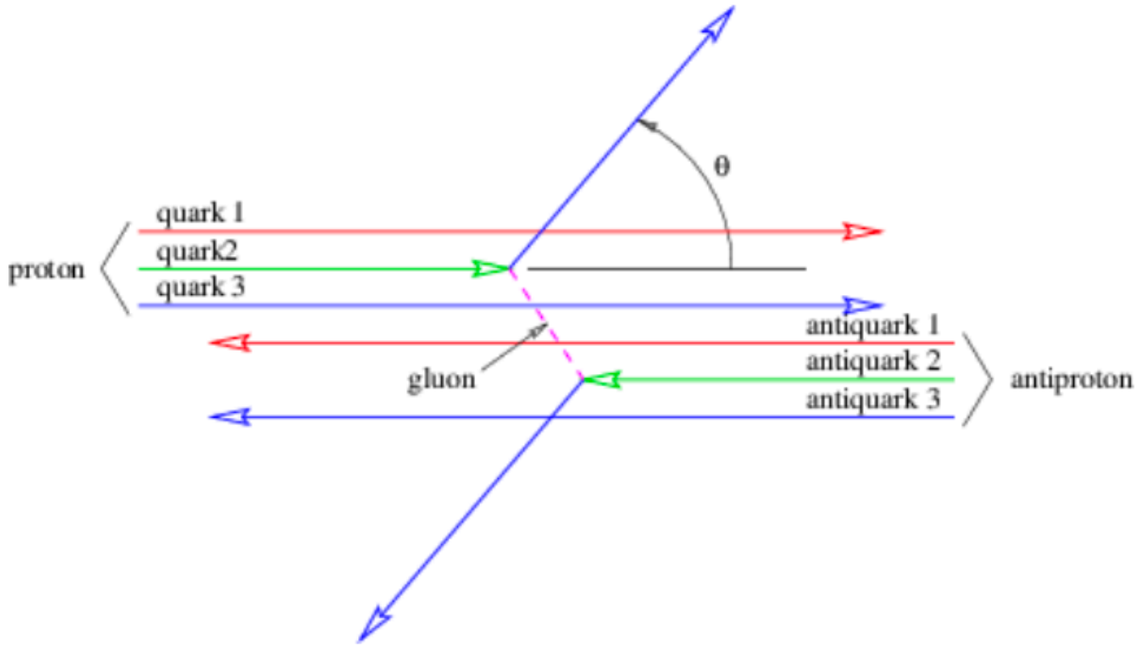


Figure 18.8: Illustration of what happens in a high energy collision between a proton and an antiproton according to the Bjorken-Feynman parton model.

Figure 18.8 illustrates the collision between a quark in the proton and an antiquark in the antiproton.

The result of this interaction is the scattering of these particles out of the incident particles, resulting ultimately in a two jet event as described above.

18.4.7 Electron-Positron Collisions

Two-jet events can also be created by the collision of high energy electrons and positrons.

Figure 18.9 shows how this process is thought to work.

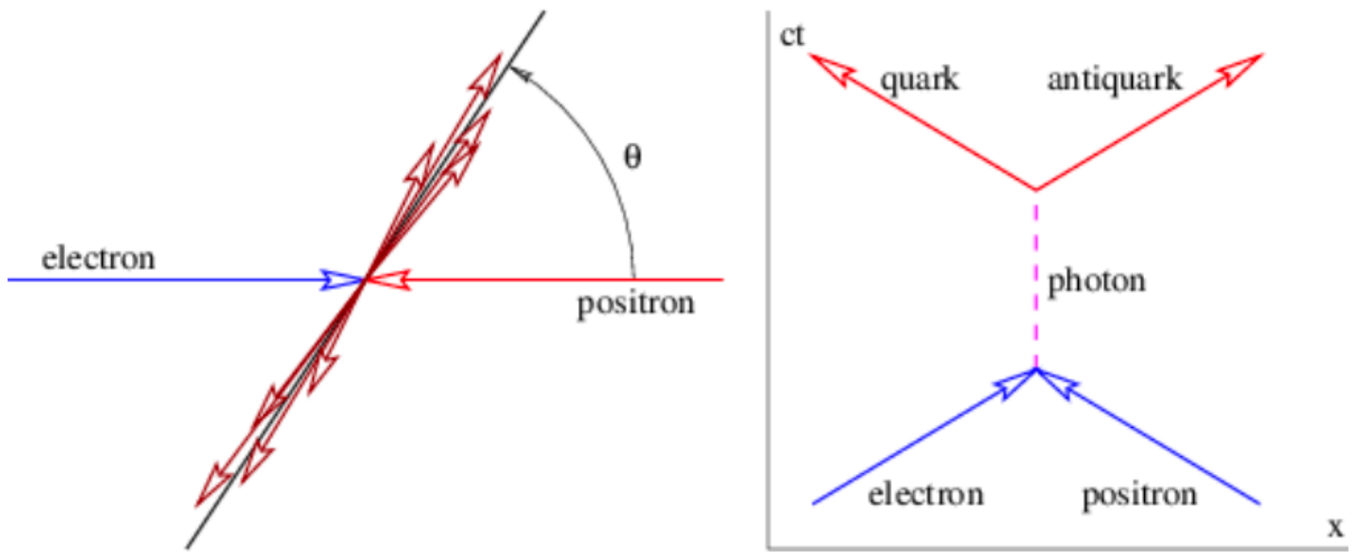


Figure 18.9: Two-jet events resulting from the annihilation of high energy electrons and positrons. The virtual photon decays into a quark-antiquark pair that in turn generates the oppositely pointing jets of particles.

The annihilation of the electron and positron results in a virtual photon, which in turn decays into a quark-antiquark pair.

The quarks then produce the jets.

These results suggest that quarks can indeed occur outside of protons, at least if they occur in quark-antiquark pairs.

18.5 Commentary

We have examined a selected set of experiments performed over the last 100 years.

Though complicated in detail, we have seen that they can be understood in their essence using one idea, namely the uncertainty principle.

This principle underlies the diffraction angle formula and also turns out (in an argument that we have not made) to be central to the q^{-4} dependence of scattering probability for point particles.

For momentum transfers of order 1000 GeV/c, we are able to probe spatial scales of order 10^{-17} m, or a factor of 100-500 less than the scale of the atomic nucleus.

Even on this scale it appears that both the electron and the quark act like point particles.

They thus appear to be the ultimate “atoms” of matter in the original sense of the word.

However, it is possible that experiments at even higher momentum transfers would show the electron or the quark to have some kind of internal structure.

Perhaps this heirarchy of structure, of which we have noted the atom, the atomic nucleus, nucleons, and quarks, goes on forever.

That is all happening now with experiments using the Large Hadron Collider (LHC) up to energies of 10^{14} TeV.

I will not discuss the LHC at this point - it is a vast subject and part of other classes.

Chapter 19

Atoms

In this chapter we investigate the structure of atoms.

However, before we can understand these, we first need to review some facts about angular momentum in quantum mechanics.

19.1 Fermions and Bosons

19.1.1 Review of Angular Momentum in Quantum Mechanics

As we learned earlier, angular momentum is quantized in quantum mechanics.

We can simultaneously measure only the magnitude of the angular momentum vector and one component, usually taken to be the z component.

Measurement of the other two components simultaneously with the z component is forbidden by the uncertainty principle.

The magnitude of the orbital angular momentum of an object can take on the values $|\mathbf{L}| = [l(l + 1)]^{1/2}\hbar$ where $l = 0, 1, 2,\ldots$.

The z component can likewise equal $L_z = m\hbar$ where $m = -l,-l + 1,\ldots,l$.

Particles can have an intrinsic spin angular momentum as well as an orbital angular momentum.

The possible values for the magnitude of the spin angular momentum are $|\mathbf{S}| = [s(s + 1)]^{1/2}\hbar$ and the z component of the spin angular momentum $S_z = m_s\hbar$ where $m_s = -s,-s + 1,\ldots,s$.

Spin differs from orbital angular momentum in that the spin can take on half-integer as well as integer values: $s = 0, 1/2, 1, 3/2,\ldots$ are possible spin quantum numbers.

Spin is an intrinsic, unchangeable quantity for an elementary particle.

Particles with half-integer spins, $s = 1/2, 3/2, 5/2,\ldots$, are called fermions, while particles with integer spins, $s = 0, 1, 2,\ldots$ are called bosons.

Fermions can only be created or destroyed in particle-antiparticle pairs, whereas bosons can be created or destroyed singly.

19.1.2 Two-Particle Wave Functions

We learned in quantum mechanics that a particle is represented by a wave, $\psi(x, y, z, t)$, the absolute square of which gives the relative probability of finding the particle at some point in spacetime.

If we have two particles, then we must ask a more complicated question: What is the relative probability of finding particle 1 at point x_1 *and* particle 2 at point x_2?

This probability can be represented as the absolute square of a *joint wave function* $\psi(x_1, x_2)$, i. e., a single wave function that represents both particles.

If the particles are not identical (say, one is a proton and the other is a neutron) and if they are not interacting with each other via some force, then the above wave function can be broken into the product of the wave functions for the individual particles:

$$\psi(x_1, x_2) = \psi_1(x_1)\psi_2(x_2) \quad \text{(non-interacting dissimilar particles).} \tag{19.1}$$

In this case the probability of finding particle 1 at x_1 and particle 2 at x_2 is just the absolute square of the joint wave amplitude: $P(x_1,x_2) = P_1(x_1)P_2(x_2)$.

This is consistent with classical probability theory.

The situation in quantum mechanics when the two particles are identical is quite different.

If $P(x_1,x2)$ is, say, the probability of finding one electron at x_1 and another electron at x_2, then since we can't tell the difference between one electron and another, the probability distribution cannot change if we switch the electrons.

In other words, we must have $P(x_1,x_2) = P(x_2, x_1)$.

There are two obvious ways to make this happen: Either $\psi(x_1,x_2) = \psi(x_2, x_1)$ or $\psi(x_1,x_2) = -\psi(x_2, x_1)$.

It turns out that the wave function for two identical fermions is *antisymmetric* to the exchange of particles whereas for two identical bosons it is *symmetric*.

In the special case of two non-interacting particles, we can construct the joint wave function with the correct symmetry from the wave functions for the individual particles as follows:

$$\psi(x_1, x_2) = \psi_1(x_1)\psi_2(x_2) - \psi_1(x_2)\psi_2(x_1) \quad \text{(non-interacting fermions)} \tag{19.2}$$

for fermions and

$$\psi(x_1, x_2) = \psi_1(x_1)\psi_2(x_2) + \psi_1(x_2)\psi_2(x_1) \quad \text{(non-interacting bosons)} \tag{19.3}$$

for bosons.

Figure 19.1 shows the joint probability distribution for two particles in different energy states in an infinite square well: $P(x_1, x_2) = |\psi(x_1, x_2)|^2$

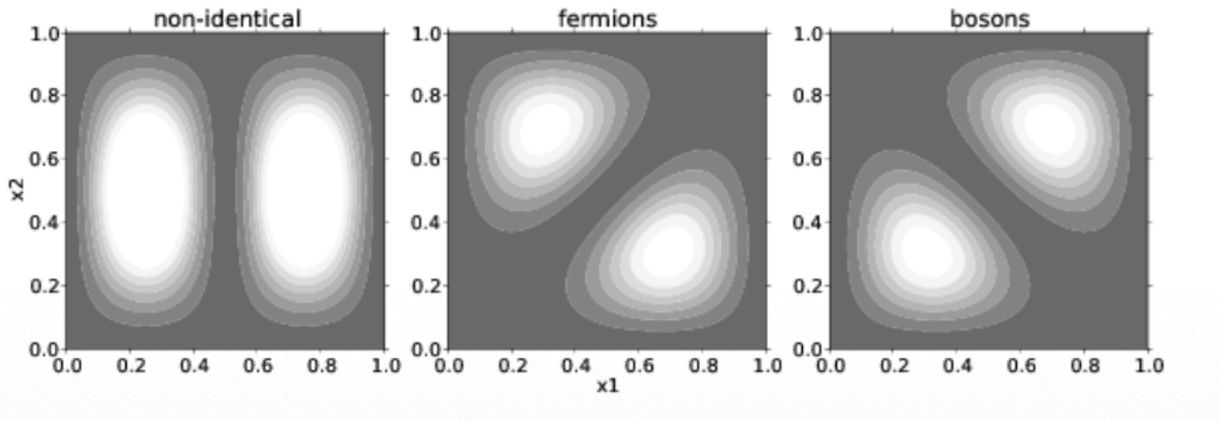


Figure 19.1: Joint probability distributions for two particles, one in the ground state and one in the first excited state of a one-dimensional box. Left panel: non-identical particles. Middle panel: identical fermions. Right panel: identical bosons. The curved lines are contours of constant probability. The lighter shading shows where the probability is large.

Three different cases are shown: non-identical particles, identical fermions, and identical bosons.

Notice that the probability of finding two fermions at the same point in space, i. e., along the diagonal dotted line in the center panel of figure 19.1, is zero.

This follows immediately from equation (19.2), which shows that $\psi(x_1, x_2) = 0$ for fermions if $x_1 = x_2$.

Notice also that if two fermions are in the same energy level (say, the ground state of the one-dimensional box) so that $\psi_1(x) = \psi_2(x)$, then $\psi(x_1, x_2) = 0$ everywhere.

This demonstrates that the two fermions cannot occupy the same state.

This result is called the *Pauli exclusion principle*.

On the other hand, bosons tend to cluster together.

Figure 19.1 shows that the highest probability in the joint distribution occurs along the line $x_1 = x_2$, i. e., when the particles are colocated.

This tendency is accentuated when more particles are added to the system.

When there are a large number of bosons, this tendency creates what is called a *Bose-Einstein condensate* in which most or all of the particles are in the ground state.

Bose-Einstein condensation is responsible for such phenomena as superconductivity in metals and superfluidity in liquid helium at low temperatures.

19.2 The Hydrogen Atom

The hydrogen atom consists of an electron and a proton bound together by the attractive electrostatic force between the negative and positive charges of these particles.

Our experience with the one-dimensional particle in a box shows that a spatially restricted particle takes on only discrete values of the total energy.

This conclusion carries over to arbitrary attractive potentials and three dimensions.

The energy of the ground state can be qualitatively understood in terms of the uncertainty principle.

A particle restricted to a region of size a by an attractive force will have a momentum equal at least to the uncertainty in the momentum predicted by the uncertainty principle: $p \approx \hbar / a$.

This corresponds to a kinetic energy $K = mv^2/2 = p^2/(2m) \approx \hbar^2/(2ma^2)$.

For the particle in a box there is no potential energy, so the kinetic energy equals the total energy.

Comparison of this estimate with the computed ground state energy of a particle in a box of length a, $E_1 = \hbar^2\pi^2/(2ma^2)$, shows that the estimate differs from the exact value by only a numerical factor π^2.

We can make an estimate of the ground state energy of the hydrogen atom using the same technique if we can somehow take into account the potential energy of this atom.

Classically, an electron with charge -*e* moving in a circular orbit of radius *a* around a proton with charge *e* at speed *v* must have the centripetal acceleration multiplied by the mass equal to the attractive electrostatic force, $mv^2/a = e^2/(4\pi\epsilon_0 a^2)$, where *m* is the electron mass.

The proton is so much more massive than the electron that we can assume it to be stationary.

Multiplication of this equation by *a*/2 results in

$$K = \frac{mv^2}{2} = \frac{e^2}{8\pi\epsilon_0 a} = -\frac{U}{2}, \tag{19.4}$$

where *U* is the (negative) potential energy of the electron and *K* is its kinetic energy.

Solving for *U*, we find that *U* = -2*K*.

The total energy *E* is therefore related to the kinetic energy by

$$E = K + U = K - 2K = -K \quad \text{(hydrogen atom)}. \tag{19.5}$$

Since the total energy is negative in this case, and since *U* = 0 when the electron is infinitely far from the proton, we can define a *binding energy* that is equal to minus the total energy:

$$E_B \equiv -E = K = -U/2 \quad \text{(virial theorem)}. \tag{19.6}$$

The binding energy is the minimum additional energy that needs to be added to the electron to make the total energy zero, and thus to remove it to infinity.

Equation (19.6) is called the *virial theorem*, and it is even true for non-circular orbits if the energies are properly averaged over the entire trajectory.

Proceeding as before, we assume that the momentum of the electron is $p \approx \hbar / a$ and substitute this into equation (19.4).

Solving this for $a \equiv a_0$ yields an estimate of the radius of the hydrogen atom:

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{e^2 m} = \left(\frac{4\pi\epsilon_0\hbar c}{e^2}\right)\left(\frac{\hbar}{mc}\right) \tag{19.7}$$

This result was first obtained by the Danish physicist Niels Bohr, using another method, in an early attempt to understand the quantum nature of matter.

The grouping of terms by the large parentheses in equation (19.7) is significant.

The dimensionless quantity

$$\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} \approx \frac{1}{137} \quad \text{(fine structure constant)} \tag{19.8}$$

is called the *fine structure constant* for historical reasons.

However, it is actually a fundamental measure of the strength of the electromagnetic interaction.

The Bohr radius can be written in terms of the fine structure constant as

$$a_0 = \frac{\hbar}{\alpha mc} = 5.29 \times 10^{-11}\ \text{m} \quad \text{(Bohr radius)}. \tag{19.9}$$

The binding energy predicted by equations (19.4) and (19.6) is

$$E_B = -\frac{U}{2} = \frac{e^2}{8\pi\epsilon_0 a_0} = \alpha\frac{\hbar c}{2a_0} = \frac{\alpha^2 mc^2}{2} = 13.6\ \text{eV}. \tag{19.10}$$

The binding energy between the electron and the proton is thus proportional to the electron rest energy multiplied by the square of the fine structure constant.

The above estimated binding energy turns out to be precisely the ground state binding energy of the hydrogen atom.

The energy levels of the hydrogen atom turn out to be

$$E_n = -\frac{E_B}{n^2} = -\frac{\alpha^2 mc^2}{2n^2}, \quad n = 1, 2, 3, \ldots \quad \text{(hydrogen energy levels)}, \tag{19.11}$$

where n is called the *principal quantum number* of the hydrogen atom.

19.3 Complex Atoms

The energy levels of the hydrogen atom whose energies are given by equation (19.11) are actually *degenerate*, in that each energy has more than one state associated with it.

Three extra degrees of freedom are associated with angular momentum, expressed by the quantum numbers l, m, and m_s.

For energy level n, the orbital angular momentum quantum number can take on the values $l = 0, 1, 2,\ldots, n - 1$.

Thus, for the ground state, $n = 1$, the only possible value of l is zero.

For a given value of l, there are $2l + 1$ possible values of the orbital z component quantum number, $m = -l,-l + 1,\ldots, l$.

Finally, there are two possible values of the spin orientation quantum number, m_s.

Thus, for the nth energy level there are

$$N_n = 2\sum_{l=0}^{n-1}(2l+1) \qquad (19.12)$$

states.

In particular, for $n = 1, 2, 3,\ldots$, we have $N_n = 2, 8, 18,\ldots$.

This is summarized in figure 19.2.

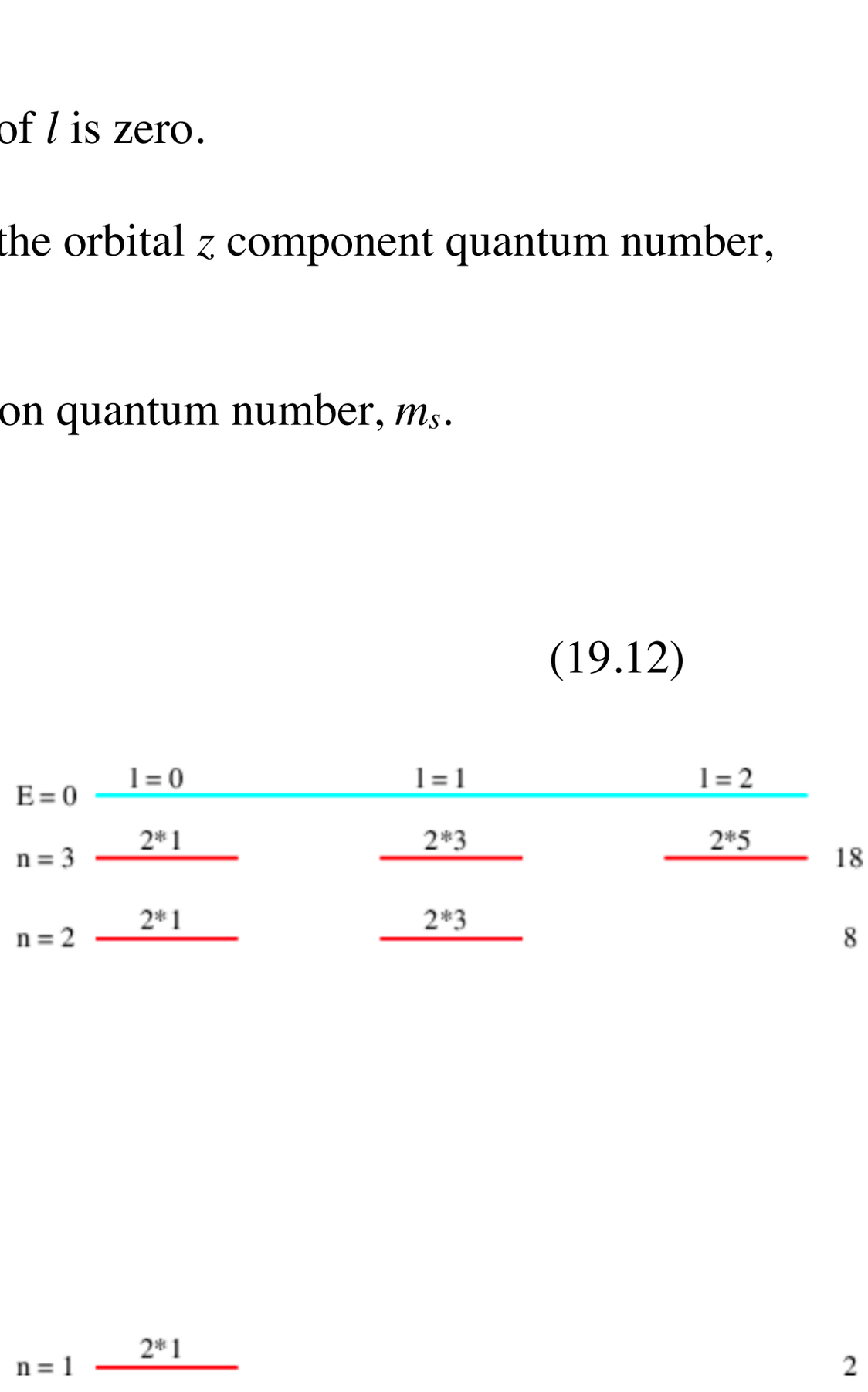


Figure 19.2: Energy levels of the hydrogen atom. Energy increases upward and angular momentum increases to the right. The numbers above each level indicate the spin orientation multiplied by the orbital orientation degeneracy for each level. The numbers at the right show the total degeneracy for each value of n. Only the first three values of n are shown.

These results have implications for the character of atoms with more than one proton in the nucleus.

Let us imagine how such atoms might be built.

The binding energy of a single electron in the ground state of a nucleus with Z protons is Z^2 multiplied by the binding energy of the electron in the ground state of a hydrogen atom.

If the force between electrons can be ignored compared to the force between an electron and the nucleus (a very poor but initially useful assumption that we will discuss below), then we could construct an atom by dropping Z electrons one by one into the potential well of the nucleus.

The Pauli exclusion principle prevents all of these electrons from falling into the ground state.

Instead, the available states will fill in order of lowest energy first until all Z electrons are added and the atom becomes electrically neutral.

From figure 19.2 we see that $Z = 2$ fills the $n = 1$ levels with two electrons, one spin up and one spin down, both with zero orbital angular momentum.

For $Z = 10$ the $n = 2$ levels fill such that two electrons have $l = 0$ and six have $l = 1$.

As electrons are added to an atom, previous electrons tend to shield subsequent electrons from the nucleus, since their negative charge partially compensates for the nuclear positive charge.

Thus, binding energies are considerably less than would be expected on the basis of the non-interacting electron model.

Furthermore, the binding energies for states with higher orbital angular momentum are smaller than those with lower values, since electrons in these states tend to be more effectively shielded from the nucleus by other electrons.

This effect becomes sufficiently important at higher Z to disrupt the sequence in which states are filled by electrons — sometimes level $n + 1$ states with low l start to fill before all the level n states with large l are full.

Accurate calculations of atomic properties in which electron-electron interactions are taken into account are possible, but are computationally expensive.

19.4 Atomic Spectra

The best evidence for atomic energy levels comes from the emission of light by atoms in a gas at low pressure.

If the atoms are put in an excited state by some mechanism, say, collisions with energetic electrons accelerated by a potential difference between electrodes, then light is emitted at particular frequencies called *spectral lines*.

These frequencies can be separated by a device called a *spectroscope*.

Spectroscopes use either a prism or a diffraction grating and ancillary optics to make the separation visible to the eye.

The frequency of a spectral line is equal to the energy difference between two states divided by Planck's constant.

This is a consequence of the conservation of energy — the energy released when an atom undergoes a transition from a state with energy E_2 to a state with energy E_1 is just the difference between these energies.

The frequency of the emitted photon is then derived from the Planck formula.

In terms of the angular frequency,

$$\omega_{21} = \frac{E_2 - E_1}{\hbar}. \tag{19.13}$$

Figure 19.3 shows the possible transitions between the lowest four energy levels of hydrogen plus the ionized state in which the electron is initially a large distance from the hydrogen nucleus.

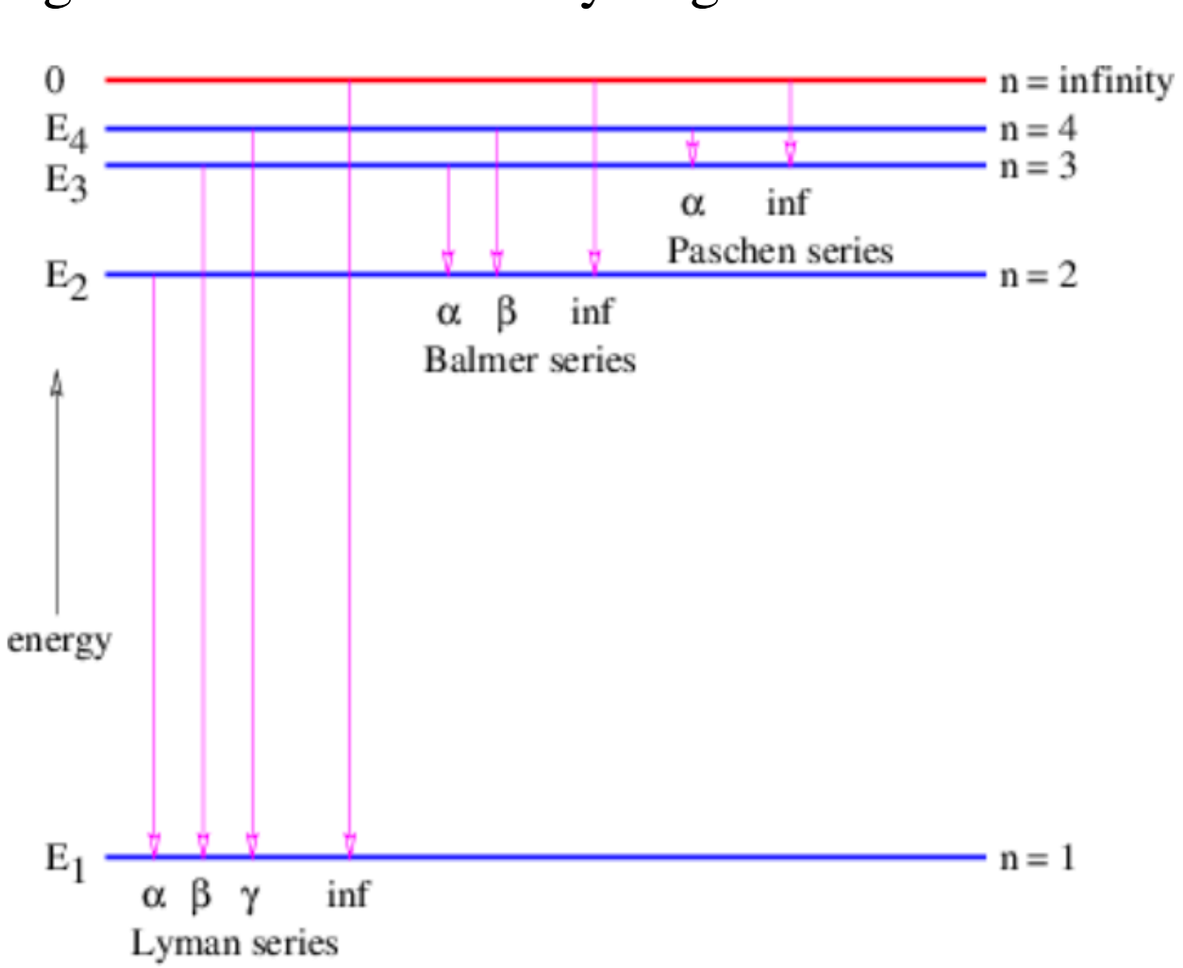


Figure 19.3: Spectral lines from transitions between electron energy states in hydrogen.

Transitions from any state to the ground state form a *series* called the *Lyman series*, while transitions to the first excited state are called the *Balmer series*, transitions to the second excited state are called the *Paschen series*, and so on.

Within each series, increasing frequencies are labeled using the Greek alphabet, so the transition from $n = 2$ to $n = 1$ is called the Lyman-α spectral line, etc.

Atoms can absorb as well as emit radiation.

For instance, if hydrogen atoms in the ground state are bombarded with photons of energy equal to the energy difference between the ground state and some excited state, some of the atoms will absorb these photons and undergo transitions to the excited state.

If white light (i. e., many photons with a continuous distribution of frequencies) irradiates such atoms, just those photons with the right energies will be absorbed.

Examination of the light with a spectroscope after it passes through a gas of atoms will show *absorption lines* where the photons with the critical energies have been removed.

This is one of the main ways in which astrophysicists learn about the elemental constitution of stars and interstellar gases.

Atoms in excited states emit photons spontaneously.

However, a process called *stimulated emission* is also possible.

This occurs when a photon with energy equal to the difference between two atomic energy levels interacts with an atom in the *higher* energy state.

The amplitude for this process is equal to the spontaneous emission amplitude times $n + 1$, where n is the number of incident photons with energy equal to the energy of the photon which would be spontaneously emitted.

If a beam of photons with the right energy shines on atoms in an excited state, the beam will gain energy at a rate that is proportional to the initial intensity of the beam.

For intense beams, this stimulated emission process overwhelms spontaneous emission and a large amount of energy can be rapidly extracted from the excited atoms.

This is how a *laser* works.

Chapter 20

The Standard Model

In this chapter we learn about the most fundamental known particles of the universe, and how they act as building blocks for everything that we know.

The theory describing this scheme is called the *standard model*.

Speculations exist about possible, more fundamental structures in the universe, such as the constructs of string theory - No predictions are made!!!!!.

However, with the standard model we have reached the frontier of what isknown with any degree of certainty.

20.1 Hadrons and Leptons

The standard model of hadrons and leptons is a united set of quantum mechanical theories encompassing electromagnetism; the weak force, which is responsible for beta decay; and the strong force, which holds atomic nuclei together.

Before investigating the standard model, we need to describe the state of affairs previous to its development.

The creation of high energy particle accelerators led to the discovery of a plethora of particles in addition to those already known.

These particles fall into the following categories:

- *Leptons* are spin 1/2 particles that do not interact via the strong force. The electron, muon, and the electron and muon neutrinos are examples.
- *Hadrons* are particles that interact via the strong force. They are divided into two sub-categories depending on their spin:
    - *Baryons* are hadrons with half-integral spin, mainly 1/2 and 3/2. The proton and neutron are well known examples. The neutral lambda particle is another.
    - *Mesons* are hadrons with integral spin, mainly 0 and 1. Examples are the pions and kaons.
- *Strange particles* are baryons and mesons that are unstable, but have much longer half-lives than other particles of similar mass and spin. This is interpreted to mean that such particles possess a property called *strangeness* that is conserved by strong processes, thus making strange particles stable against strong decay into non-strange particles. However, strangeness is not conserved by weak processes, allowing strange particles to decay via the weak interaction, which indeed is much weaker than the strong interaction at low energies. This explains their anomalously long half-lives. Strange particles are always created in pairs by strong processes in such a way that the total strangeness remains zero. For instance, if one particle has strangeness +1 then the other must have strangeness -1. An example of strange particle production is when a negative pion collides with proton, giving rise to a neutral lambda particle and a neutral kaon.

- *Intermediary particles* are those that transfer energy, momentum, charge, and other properties from one particle to another in association with one of the four fundamental forces.
    - *Photons* transmit the electromagnetic force and have zero mass and spin 1.
    - *Gravitons* are thought to transmit the gravitational force, though they have not been directly observed. The graviton is postulated to have zero mass and spin 2.

We will discover additional intermediary particles in our discussion of the standard model.

- *Antiparticles* exist for all particles. These have the same mass and spin but opposite values of the electric charge and various other quantum numbers such as lepton number or baryon number. The lepton number is the number of leptons minus the number of antileptons, with a similar definition for baryon number. Thus, a lepton has lepton number 1 and a baryon has baryon number 1. Their antiparticles have lepton number -1 and baryon number -1. As far as we know, baryon number and lepton number are absolutely conserved, which means that baryons and leptons can only be created or destroyed in particle-antiparticle pairs. Antiparticles are represented by the symbol of the particle with an overbar.

20.2 Quantum Chromodynamics

The standard model postulates that all known particles are either fundamental point particles or are composed of fundamental point particles according to a remarkably small set of rules.

Just as atoms are bound states of atomic nuclei and electrons, atomic nuclei are bound states of protons and neutrons.

Atomic nuclei are discussed in the next section.

In this section we delve one step deeper in the heirarchy of the universe.

We now believe that all hadrons are actually bound states of fundamental spin 1/2 particles called *quarks*.

Whereas all other known charged particles have an electric charge equal to an integer multiple of $\pm e$ where e is the proton charge, quarks have electric charges equal to either $-e/3$ or $+2e/3$.

Leptons themselves are considered to be fundamental, so the leptons and the quarks form the basic building blocks of all matter in the universe.

Quarks are subject to electromagnetic forces via their charge, but interact most strongly via the so-called *strong force*.

The strong force is carried by massless, uncharged, spin 1 bosons called *gluons*.

When Murray Gell-Mann and George Zweig first proposed the quark model in 1963, they needed to postulate only three types or *flavors* of quarks: *up*, *down*, and *strange*.

These were sufficient to explain the constitution of all hadrons known at the time.

We currently know of six different flavors of quarks.

Their properties are listed in table 20.1.

Type	Charge	Rest energy	*s*	*c*	*b*	*t*
down (d)	-1/3	0.333	0	0	0	0
up (u)	+2/3	0.330	0	0	0	0
strange (s)	-1/3	0.486	-1	0	0	0
charm (c)	+2/3	1.65	0	+1	0	0
bottom (b)	-1/3	4.5	0	0	-1	0
top (t)	+2/3	176	0	0	0	+1

Table 20.1: Table of quark types, charge (as a fraction of the proton charge), rest energy (in GeV), and the four "exotic" flavor quantum numbers.

The properties *charm*, *topness*, and *bottomness* are analogous to strangeness — these properties are conserved in strong interactions.

Weak interactions, discussed in the next section, can turn quarks of one flavor into another flavor.

However, the strong and electromagnetic forces cannot do this.

Just as the proton and the neutron have antiparticles, so do quarks.

Antiquarks of a particular type have strong and electromagnetic charges of the sign opposite to the corresponding quarks.

Quarks have baryon number equal to 1/3, while antiquarks have -1/3.

Thus combining three quarks results in a baryon number equal to 1, while together a quark plus an antiquark have baryon number zero.

All baryons are thus combinations of three quarks, while all mesons are combinations of a quark and an antiquark.

Table 20.2 lists a sampling of hadrons and some of their properties.

Notice that the same combination of quarks can make up more than one particle, e. g., the positive pion and the positive rho.

The positive rho may be considered as an excited state of the ud system, while the positive pion is the ground state of this system.

Type	Charge	Rest energy	Spin	Composition	Mean life
proton	+1	938.280	1/2	uud	stable
neutron	0	939.573	1/2	udd	898
lambda	0	1116	1/2	uds	3.8×10^{-9}
delta++	+2	1232	3/2	uuu	5.6×10^{-24}
positive pion	+1	140	0	$u\bar{d}$	2.6×10^{-8}
negative pion	-1	140	0	$\bar{u}d$	2.6×10^{-8}
neutral pion	0	135	0	$u\bar{u} - d\bar{d}$	8.7×10^{-17}
positive rho	+1	770	1	$u\bar{d}$	4×10^{-24}
positive kaon	+1	494	0	$u\bar{s}$	1.24×10^{-8}
neutral kaon	0	498	0	$d\bar{s}$	8.6×10^{-11}
J/psi	0	3097	1	$c\bar{c}$	1.5×10^{-20}

Table 20.2: Sample hadrons. The charge is specified as a fraction of the proton charge, the rest energy is in MeV, and the mean life (1.44 multiplied by the half life) is in seconds. The composition is presented in terms of the flavors of quarks and antiquarks that make up the particle.

Yet to be mentioned is the quantum number *color*, which has nothing to do with real colors, but has analogous properties.

Each flavor of quark can take on three possible color values, conventionally called red, green, and blue.

This is illustrated in figure 20.1.

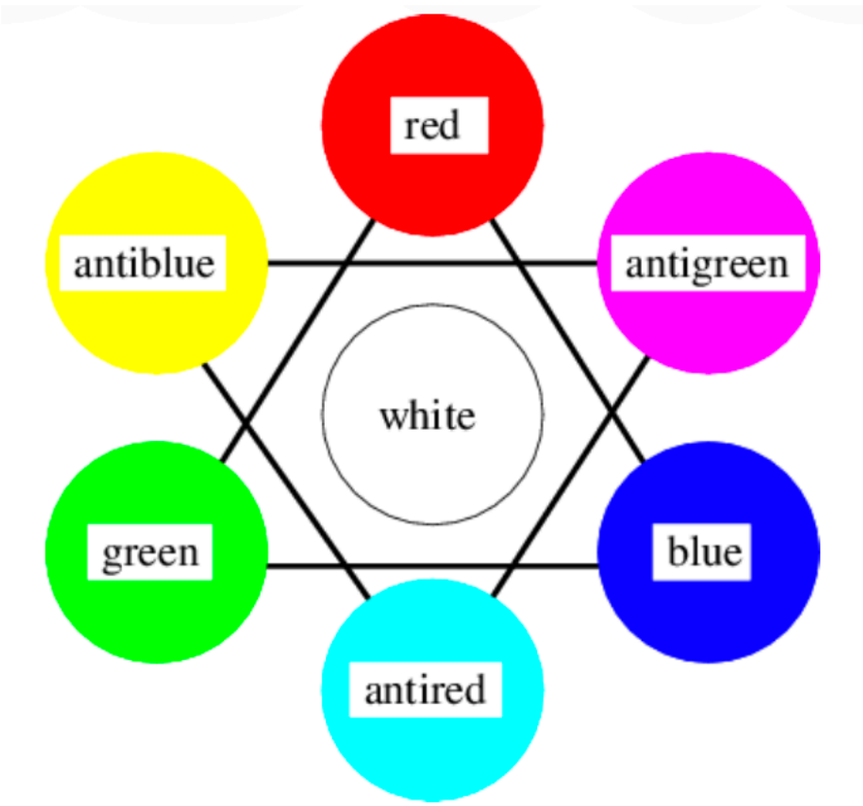


Figure 20.1: The color table of quantum chromodynamics. Quarks have colors red, green, and blue, while antiquarks have colors antired (cyan), antigreen (magenta), and antiblue (yellow). The combination of a quark and its corresponding antiquark is colorless or white, as is the combination of three quarks (or antiquarks) of three different colors.

TheAntiquarks can be thought of as having the colors antired, antigreen, and antiblue, also known as cyan, magenta, and yellow.

Because of this, the theory of quarks and gluons is called *quantum chromodynamics*.

Counting all color and flavor combinations, there are $6 \times 3 = 18$ known varieties of quarks.

As in electromagnetism, the strong force has associated with it a "strong charge" , g_s.

However, this charge is somewhat more complicated than electromagnetic charge in that there are three kinds of strong charge, one for each of the strong force colors.

Each color of charge can take on positive and negative values equal to $\pm g_s$.

As with electromagnetism, positive and negative charges (of the same color) cancel each other.

However, in quantum chromodynamics there is an additional way in which charges can cancel.

A combination of equal amounts of red, green, and blue charges results in zero net strong charge as well.

The delta++ particle (see table 20.2) is good evidence for the existence of the color quantum number.

Since all three u-quarks in the delta++ have spin orientation up, the Pauli exclusion principle would only allow one of these quarks to exist in the ground state if color did not exist, resulting in a much larger mass.

As it is, each of the quarks in the delta++ takes on a different value of the color quantum number (red, green, or blue), which means that the Pauli exclusion principle does not prevent them from all from residing in the ground state.

Gluons, the intermediary particles of the strong interaction come in eight different varieties, associated with differing color-anticolor combinations.

Since gluons don't interact via the weak force, there is no flavor quantum number for gluons — quarks of all flavors interact equally with all gluons.

The quark model of matter has led to extensive searches for free quark particles.

However, these searches for free quarks have proven unsuccessful.

The current interpretation of this result is that quarks cannot exist in a free state, basically because the attractive potential energy between quarks increases linearly with separation.

This appears to be related to the fact that gluons, the intermediary particles for the strong force, can interact with each other as well as with quarks.

This leads to a series of increasingly complex processes as quarks move farther and farther apart.

The result is called *quark confinement* — apparently, individual quarks can never be observed outside of the confines of the observable particles that contain them.

Confinement works not only on single quarks, but on any "colored" combinations of quarks and gluons, e. g., a red up quark combined with a green down quark.

It appears that long range inter-quark forces only vanish for interactions between "white" or "color-neutral" combinations of quarks.

This is why only color-neutral combinations of quarks — three quarks of three different colors or a quark-antiquark pair of the same color — are actually seen as observable particles.

The strong equivalent of the fine structure constant is the *coupling constant* for the strong force:

$$\alpha_s = \frac{g_s^2}{4\pi\hbar c}. \tag{20.1}$$

Note that α_s is dimensionless.

The binding energy between quarks is comparable to the rest energies of the quarks themselves.

In other words, $\alpha_s \approx 1$.

Furthermore, as we have noted, the potential energy between two quarks appears to increase indefinitely with separation.

Though forces exist between color-neutral particles, they are weak and of short range compared to the forces between quarks or colored combinations of quarks.

However, they are still relatively strong compared to, say, electromagnetic forces.

As we shall see later, these residual strong forces are responsible for nuclear processes.

Interactions between hadrons can be thought of as resulting from interactions between the individual quarks making up the hadrons.

Two sample strong interactions are shown in figure 20.2.

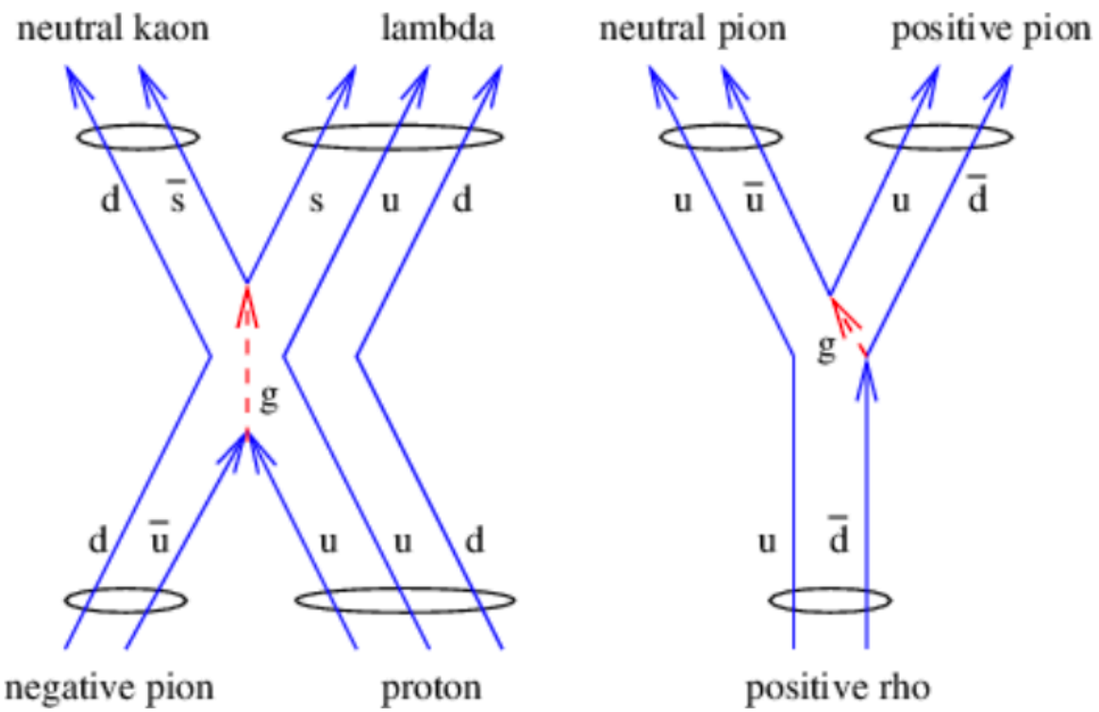


Figure 20.2: Some sample strong interactions illustrated in terms of gluon emission and absorption. The process on the left shows the reaction $\pi^- + p \rightarrow K^0 + \Lambda^0$, while the one on the right shows the decay $\varrho^+ \rightarrow \pi^+ + \pi^0$. Quarks are labeled with solid lines while gluons are shown by dashed lines.

Virtual gluons can be emitted and absorbed by quarks much as virtual photons can be emitted and absorbed by electrically charged particles.

Particles unstable to strong decay processes (such as the positive rho particle) typically live only about 10^{-23} s, whereas particles stable to strong decay but unstable to weak decay live of order 10^{-10} s or longer, depending strongly on how much energy is liberated in the decay.

Particles subject to electromagnetic decay processes, such as the neutral pion, take on mean lifetimes intermediate between strong and weak values, typically of order 10^{-18} s.

20.3 The Electroweak Theory

The strong force acts only on quarks and the strong force carrier, the gluon.

It does not act on leptons, e. g., electrons, muons, or neutrinos.

Table 20.3 shows all of the known leptons.

The so-called *weak* force acts on leptons as well as on quarks.

Type	Charge	Rest energy	Mean life
electron (e^-)	-1	0.000511	stable
electron neutrino (ν_e)	0	≈ 0	stable
muon (μ^-)	-1	0.106	2.2×10^{-6}
mu neutrino (ν_μ)	0	≈ 0	stable
tau (τ)	-1	≈ 1.7	3.0×10^{-13}
tau neutrino (ν_τ)	0	≈ 0	stable

Table 20.3: Table of lepton types, charge (as a fraction of the proton charge), rest energy (in GeV), and mean life (in seconds).

In 1979 Sheldon Glashow, Abdus Salam, and Steven Weinberg won the Nobel Prize for their *electroweak theory*, which unites the electromagnetic and weak interactions.

Unlike the strong and electromagnetic forces, the intermediary particles of the weak interaction, the W^+, the W^-, and the Z^0 have rather large masses.

In particular, the rest energy of the $W^\pm$ is 81 GeV while that of the Z^0 is 92 GeV.

Electroweak theory considers electromagnetism and the weak interactions to be different aspects of the same force.

A key aspect of the theory is the explanation of why three out of four of the intermediary particles of the electroweak force are massive.

The photon is the massless one.

Unfortunately, the details of why this is so are highly technical, so we cannot delve into this subject here.

We only note that the explanation requires the existence of a massive spin zero boson called the Higgs particle.

It has now been determined that the Higgs particle exists - it was found in 2012!

It properties are: It is a boson which interacts through the electroweak force, has zero charge and zero spin, has a mass of 125 GeV and has a mean lifetime of 1.56×10^{-22} sec.

All properties were predicted by the nobel-prize winning theory.

The weak force has certain bizarre properties not shared by the other forces of nature:

- The weak interaction can change quark flavors. For instance, the beta decay of a neutron converts a down quark into an up quark. On the other hand, the weak interaction is "colorblind", i. e., it is insensitive to quark colors.
- The weak interaction is not left-right symmetric. In other words, the physical laws governing the weak interaction look different when seen in a mirror.
- The weak interaction is slightly asymmetric to the interchange of particles and antiparticles in certain situations.

The prototypical weak interaction is the decay of the neutron into a proton, an electron, and an antineutrino.

This decay is energetically possible because the neutron is slightly more massive than the proton and is illustrated in the left panel of figure 20.3.

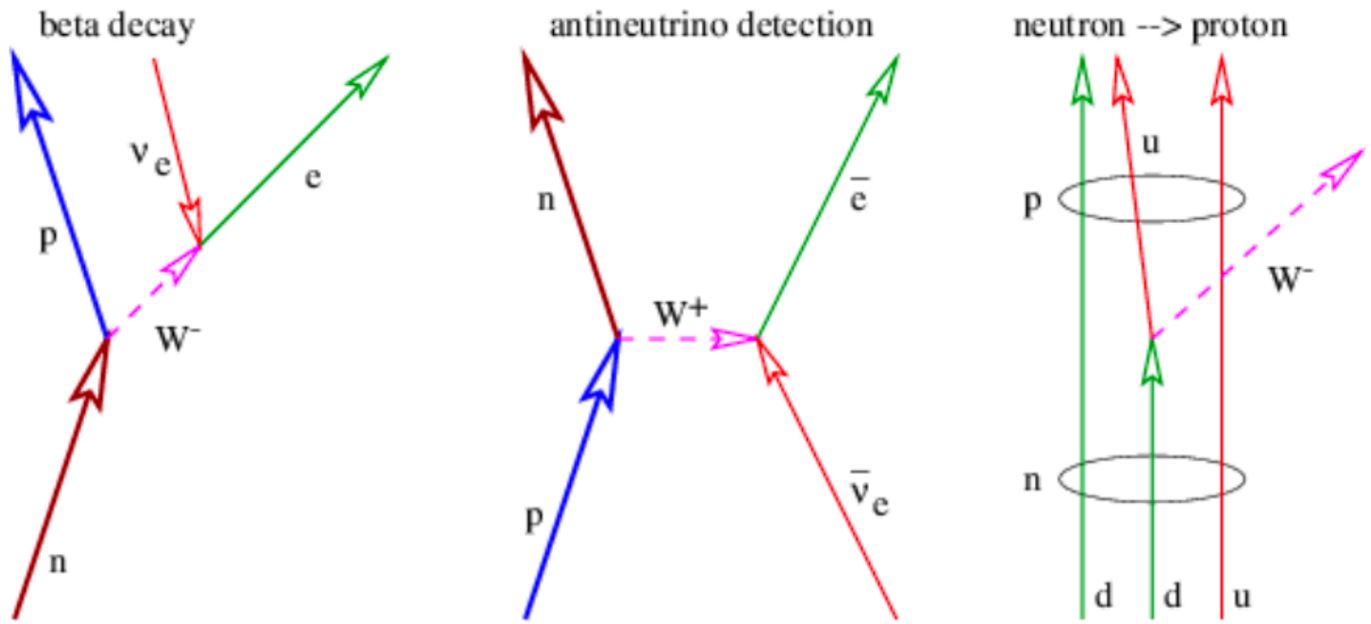


Figure 20.3: Illustration of two weak reactions. The left panel shows beta decay while the middle panel shows how electron antineutrinos can be detected by conversion to a positron. The right panel shows how W^- emission works according to the quark model, resulting in the conversion of a down quark to an up quark and the resulting transformation of a neutron into a proton.

Note that this figure is drawn as if a neutrino moving backward in time absorbs a W^- particle, with a resulting electron exiting the reaction forward in time.

However, we know that this is equivalent to an electron and an antineutrino both exiting the reaction forward in time according to the Feynman interpretation of negative energy states.

The weak interaction is called “weak” because it appears to be so in commonly observed processes.

For instance, the range of a relativistic electron in ordinary matter is of order centimeters to meters.

This is because the electromagnetic force between the charge of the electron and the charges on atomic nuclei are strong enough to rapidly cause the energy of the electron to be dissipated.

However, the range in matter of a neutrino produced by beta decay is many orders of magnitude greater than that of an electron.

This is *not* because the weak force is intrinsically weak — the value of the "fine structure constant" for the weak force is

$$\alpha_w \approx 10^{-2} \tag{20.2}$$

according to the standard model, and is actually larger than α for electromagnetism.

The real reason for the apparent weakness of the weak force is the large mass of the intermediary particles.

As we have seen, large mass translates into short range for a virtual particle at low momentum transfers.

This short range is what causes the weak force to appear weak for momentum transfers much less than the masses of the W and Z particles, i. e., for $q \ll 100$ GeV.

For leptons and quarks with energies $E \gg 100$ GeV, the weak force acts with much the same strength as the electromagnetic force.

20.4 Grand Unification?

Table 20.4: Generations of leptons and quarks. Members of each generation tend to fit together.

Generation	Leptons	Quarks
1	electron	down
	electron neutrino	up
2	muon	strange
	mu neutrino	charm
3	tau	bottom
	tau neutrino	top

The standard model is a great achievement, but it leaves a number of questions unanswered.

As table 20.4 shows, nature seems to have produced more particles than are needed to construct the universe.

Virtually everything we know of is composed of electrons, electron neutrinos, up quarks, and down quarks.

These four particles seem to fall naturally together in a *family* or *generation*.

Why then are there apparently unneeded additional generations?

What role do muons, taus, and the exotic quark forms play in the universe?

Another question concerns the dichotomy between leptons and quarks.

Electrons and electron neutrinos can be converted into each other by weak interactions, as can up and down quarks.

Why then can't quarks be converted into leptons and vice versa?

In the standard model, electromagnetic and weak forces are truly united as aspects of a single phenomenon.

However, quantum chromodynamics stands more on its own.

One could imagine further advances that would show that the electroweak and strong forces were in fact different aspects of the same phenomenon.

This could be characterized as a *grand unification* of the forces of nature.

As previously noted, the strong force coupling constant, α_s, gets smaller with increasing momentum transfer.

It turns out that the weak coupling constant, α_w, exhibits similar behavior, while the electromagnetic coupling constant, the fine structure constant α, becomes stronger at higher energies.

This behavior is illustrated in figure 20.4, though it is based on data only up to about 10_3 GeV/c.

Figure 20.4 is thus largely speculative.

However, if the observed trends do continue to very high momentum transfers, this would be evidence in favor of grand unification.

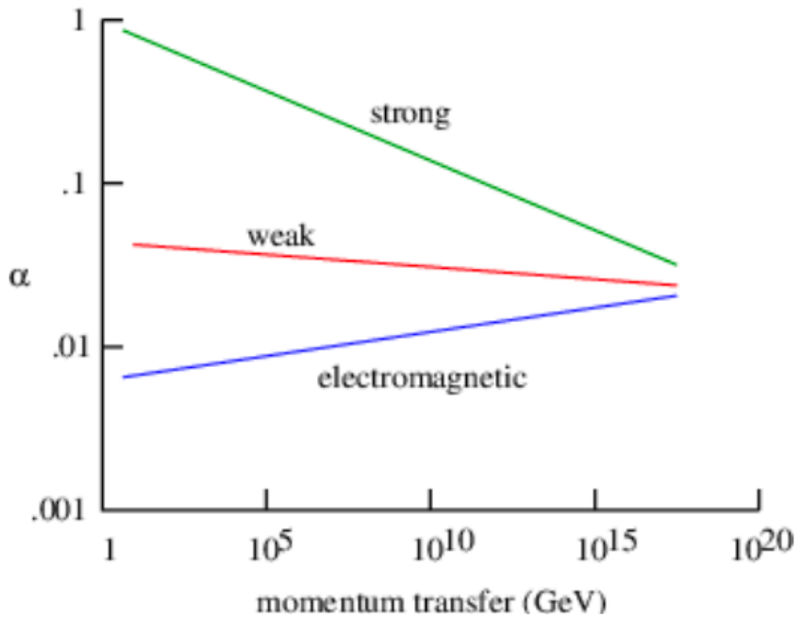


Figure 20.4: Speculated behavior of the dependence of the coupling constant α on momentum transfer for each of the forces. Extrapolation based on current measurements suggests that these constants come together to a common value at very high momentum transfer.

A number of speculative grand unification theories have been proposed.

Most such theories view leptons and quarks as being different states of the same particle and also predict that leptons can turn into quarks and vice versa, albeit at very low rates.

One of the consequences of such theories is that the proton would be an unstable particle, but with a very long lifetime, of order 10^{33} yr.

Experiments have been done to detect the decay of the proton, but so far without success.

These experiments are sufficient to rule out some but not all of the proposed grand unification theories.

One task that would not be accomplished by grand unification is the incorporation of gravity into a common framework with the strong, weak and electromagnetic forces.

Creation of a satisfactory quantum theory of gravity has been a very difficult problem and is unsolved to this day, though many people are working on it.

Chapter 21

Atomic Nuclei

Atomic nuclei are composite particles made up of protons and neutrons.

These two particles are collectively known as *nucleons*.

In order to better understand atomic nuclei, we first make an analogy with molecules.

We then investigate the binding energies of atomic nuclei.

This information is central to the subjects of radioactive decay as well as nuclear fission and fusion.

21.1 Molecules — an Analogy

Molecules are bound states of two or more atoms.

In chemistry we identify several modes of molecular binding, e. g., covalent and ionic bonds, the hydrogen bond, and binding at low temperatures due to the van der Waal's force.

All of these bonds involve electromagnetic forces, but all (except arguably the ionic bond) are relatively subtle residual forces between atoms that are electrically neutral.

The ways in which atoms form molecules are therefore complex and resistant to accurate calculation.

Atomic nuclei are the nuclear equivalent of molecules, in that they are bound states of nucleons, which are themselves "uncharged" composite particles.

The charge we refer to here is not the electric charge (nuclei do of course possess this!), but the strong or color charge.

As we discovered in the previous chapter, nucleons are color-neutral combinations of quarks.

Thus, the "strong" forces between nucleons are subtle residuals of inter-quark forces.

This is reflected in the binding energies; quark-quark binding energies are on the order of the rest energies of the quarks themselves.

However, nuclear binding energies are typically of order 10 MeV per nucleon, or about 1% of the rest energy of a nucleon.

The residual nature of nuclear forces makes them complex and difficult to calculate from our basic knowledge of quantum chromodynamics for the same reasons that intermolecular forces are difficult to calculate.

An empirical approach is thus needed in order to understand their effects.

In contrast to molecules and atomic nuclei, atoms are relatively easy to understand.

This is true for two reasons: (1) Electrons appear to be truly fundamental point particles. (2) Though the atomic nucleus itself is a very complex system, little of this complexity spills over into atomic calculations, because on the atomic scale the nucleus is very nearly a point particle.

Thus, both main ingredients in atoms are "simple" from the point of view of atomic calculations.

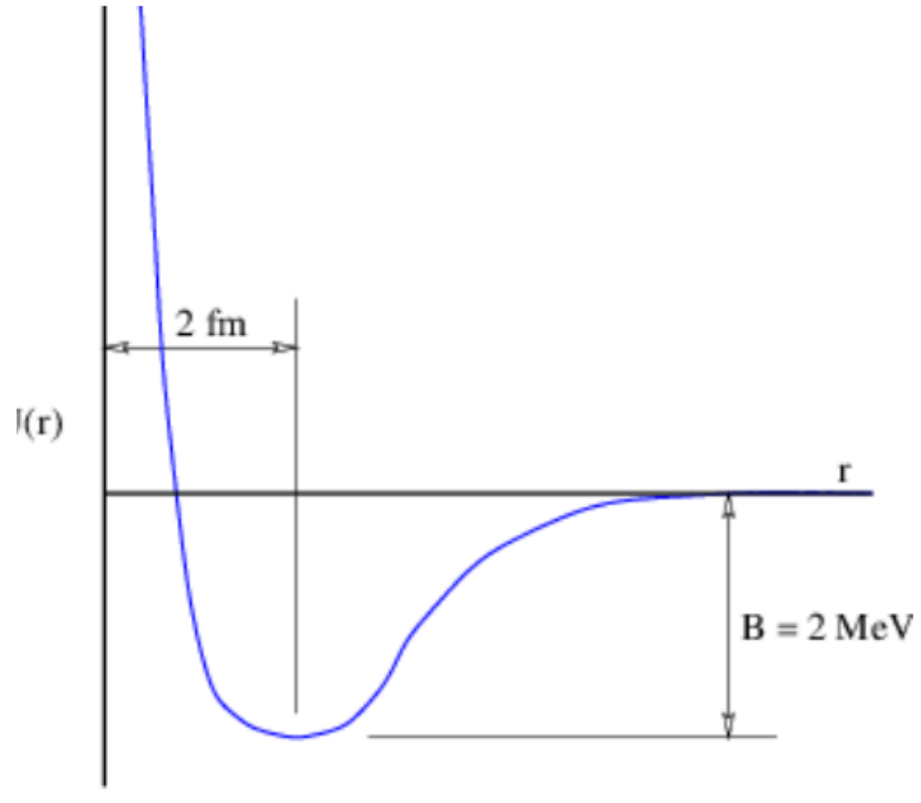
The above result is true because by some accident of nature, the mass of the electron is so much less than the masses of quarks.

It would be interesting to speculate what atomic theory would be like if this weren't true — there would be no scale separation between the atomic and nuclear scales, and the world would be a very different place!

21.2 Nuclear Binding Energies

It is impossible to specify an accurate inter-nucleon force valid under all circumstances, but figure 21.1 gives an approximate representation of the potential energy associated with the strong force as the function of nucleon separation.

Figure 21.1: Approximate sketch of the strong force potential energy between two nucleons. 1 fm = 10^{-15} m. The binding energy B is the energy required to separate the two nucleons. If the nucleons are bound together, the rest energy of the resulting combination, $M_{combo}c^2$ is less than the sum of the rest energies of the two nucleons, M_1c^2 , M_2c^2 , by the amount B: $M_{combo}c^2 = M_1c^2 + M_2c^2 - B$.



The binding energy is of order 2 MeV, with an attractive force for separations greater than about 2×10^{-15} m and an intense repulsive force for smaller separations.

At large distances the potential energy decays exponentially with distance rather than according to the r^{-1} law of the Coulomb potential.

The short range of the inter-nuclear force means that atomic nuclei can be thought of as conglomerations of “sticky billiard balls”.

The nuclear force is essentially a contact force and each nucleon simply binds to all its nearest neighbors.

When nucleons are close-packed, the binding energy per nucleon due to the strong force is simply the number of nearest neighbors for each nucleon, multiplied by the binding energy per nucleon pair, divided by 2. he cube root of the number of nucleons.

The factor of 1/2 accounts for the fact that each nuclear bond is shared by two nucleons.

Several other effects need to be accounted for in the nucleus.

The nucleons on the surface of the nucleus do not have as many bonds as nucleons in the interior.

Thus, to compute the nuclear binding energy of a nucleus with a finite number of nucleons, a correction must be made for this effect.

This contributes negatively to the nuclear binding energy in proportion to the surface area of the nucleus, which scales as the number of nucleons to the two-thirds power.

In addition to the nuclear force, the repulsive electrostatic force between protons needs to be accounted for.

Since the electrostatic force is a long range force, the (negative) contribution to the binding energy of the nucleus goes as the square of the number of protons divided by the radius of the nucleus.

The latter goes as the cube root of the number of nucleons.

The Pauli exclusion principle operates in nuclei so as to favor equal numbers of protons and neutrons.

This effect is illustrated in figure 21.2.

The If a proton is converted into a neutron in a nucleus in which equal numbers of the two particles occur, then the exclusion principle forces these nucleons to move to a higher energy level than they previously occupied.

The binding energy of the nucleus is correspondingly decreased.

This effect opposes the weaker, repulsive Coulomb potential that occurs when there are more neutrons and fewer protons.

The net result of all these effects is a nuclear binding energy equation with four terms representing the four above-mentioned effects:

$$B(Z,A) = a_v A - a_s A^{2/3} - a_c Z^2/A^{1/3} - a_a (2Z - A)^2/A \qquad (21.1)$$

where Z is the *atomic number* or the number of protons, N is the number of neutrons, and $A = Z + N$ is the *atomic mass number*, or number of nucleons.

Equation (21.1) represents the binding energy of the entire nucleus.

The binding energy per nucleon is just B/A.

Fitting equation 21.1 to observed binding energies in nuclei yields the following values for the coefficients of the above equation: $a_v \approx 16$ MeV, $a_s \approx 17$ MeV, $a_c \approx 0.70$ MeV, and $a_a \approx 23$ MeV.

A contour plot of binding energy per nucleon, *B/A*, is shown in figure 21.3.

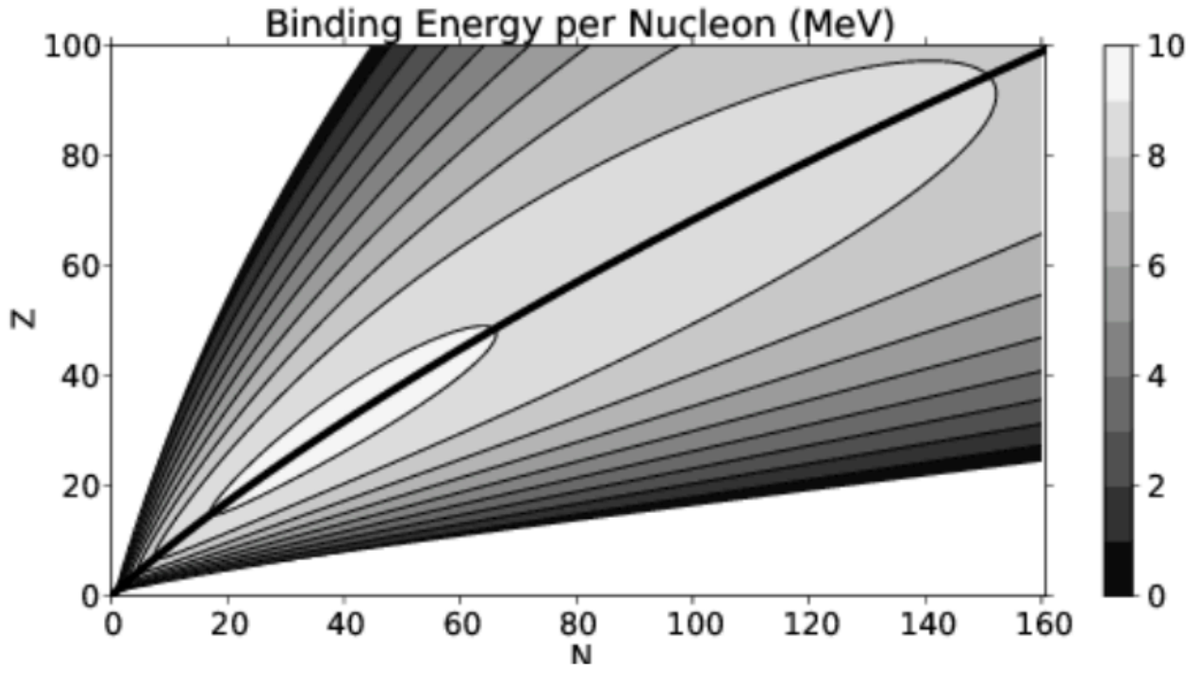


Figure 21.3: Nuclear binding energy per nucleon $B(Z,A)/A$, calculated from equation (21.1). The thick curved line starting near the origin gives the line of stability for atomic nuclei. The white areas near the horizontal and vertical axes indicate negative binding energy.

We note that this equation doesn't work well for nuclei with only a few nucleons.

For instance, the helium nucleus with $A = 4$ is more stable than the lithium nucleus with $A = 6$, and there is no stable nucleus at all with $A = 5$.

Part of the reason for the problem at small A is that even numbers of protons and neutrons tend to bind more strongly together than nuclei containing odd numbers of either.

This is because pairs of protons or neutrons with opposite spins fully occupy nuclear states while an odd nucleon occupies a state by itself with energy greater than that of all the other occupied states.

This behavior can be approximately accounted for by adding the term $a_p/A^{1/2}$ to equation (21.1), where $a_p = 12$ MeV if N and Z are both even, $a_p = 0$ if either N or Z is odd, and $a_p = -12$ MeV if both are odd.

We leave this term off even though it is sometimes quite important, in order to make equation (21.1) a smooth function of Z and A and thus representative of the general trend of binding energy.

For a given value of A, it is easy to demonstrate that the maximum nuclear binding energy in equation (21.1) occurs when

$$Z = \frac{A}{2(1 + a_c A^{2/3}/4a_a)}. \tag{21,2}$$

This formula confirms the trend seen in figure 21.3 that the most stable nuclear configuration contains an increasing fraction of neutrons as A increases.

The function $Z(N)$ given by equation (21.2) and illustrated by the curve starting near the origin in figure 21.3 defines the *line of stability* for atomic nuclei.

Figure 21.4 shows the binding energy per nucleon as a function of nucleon number A along the line of stability.

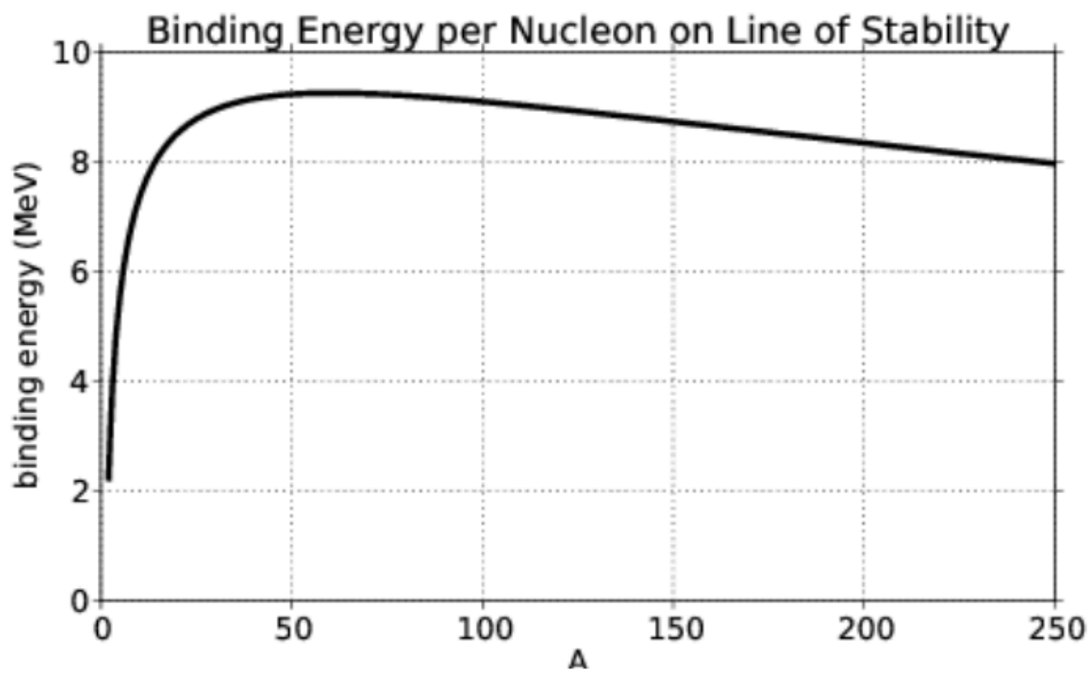


Figure 21.4: Binding energy per nucleon along line of stability according to equations (21.3) and (21.2).

The rapid increase in binding energy for small A reflects the decreasing surface effect as the number of nucleons increases.

The subsequent decrease is a result of the combined effects of Coulomb repulsion of protons and the Pauli exclusion principle.

Notice that the maximum binding energy per nucleon occurs near $A = 60$.

The chemical properties of the atom associated with an atomic nucleus are determined by the number of protons, Z, in the nucleus.

In many cases there exists more than one stable or long-lived nucleus with a given value of Z.

These nuclei differ in their neutron number, N.

Nuclei with the same Z and differing N are called *isotopes* of the *element* defined by the specified value of Z.

For instance, there are three isotopes of the element hydrogen, normal hydrogen, deuterium, and tritium, with zero, one, and two neutrons respectively.

21.3 Radioactivity

Radioactive decay is the emission of some particle from an atomic nucleus, accompanied by a change of state or type of the nucleus, depending on the type of radioactivity.

Gamma rays or photons are emitted when a nucleus decays from an excited state to its ground state.

No transformation of the nuclear type occurs.

Photons are often emitted when some other form of radioactive decay leaves the resulting nucleus in an excited state.

Beta minus decay is the conversion of a neutron into a proton, an electron, and an electron antineutrino.

This and the inverse reaction, beta plus decay, or conversion of a proton into a neutron, a positron, and an electron neutrino, were described in chapter 18.

These processes occur in the nucleons contained in nuclei when they are energetically possible.

Alpha particle emission occurs in heavy elements where it is energetically possible.

Since an alpha particle is just a helium 4 nucleus containing two protons and two neutrons, the values of Z and N of the emitting nucleus are each reduced by two.

The rest energy of a nucleus (ignoring atomic effects) is just the sum of the rest energies of all the nucleons minus the total binding energy for the nucleus:

$$M(Z,A)c^2 = ZM_pc^2 + NM_nc^2 - B(Z,A), \tag{21.3}$$

where $M_pc^2 = 938.280$ MeV is the rest energy of the proton and $M_nc^2 = 939.573$ MeV is the rest energy of the neutron.

Energy conservation requires that

$$M(Z,A)c^2 = M(Z-2,A-4)c^2 + M(2,4)c^2 + Q \tag{21.4}$$

for the alpha decay of a nucleus.

If $Q > 0$, then the decay is energetically possible.

The excess energy, Q, goes into kinetic energy of the new nucleus and the alpha particle, mainly the latter.

Substitution of equation (21.3) into equation (21.4) yields

$$Q = B(Z-2,A-4) + B(2,4) - B(Z,A) \quad \text{(alpha decay)}. \tag{21.5}$$

The binding energy of the alpha particle is not accurately represented by equation (21.1), but is known to be about $B(2, 4) = 28.3$ MeV.

On the other hand, the heavy elements are generally well represented by equation (21.1).

The curve of Q versus A is plotted in figure 21.5, and it shows that alpha decay for nuclei along the line of stability is energetically impossible (i. e., $Q < 0$) for nuclei with A less than about 175.

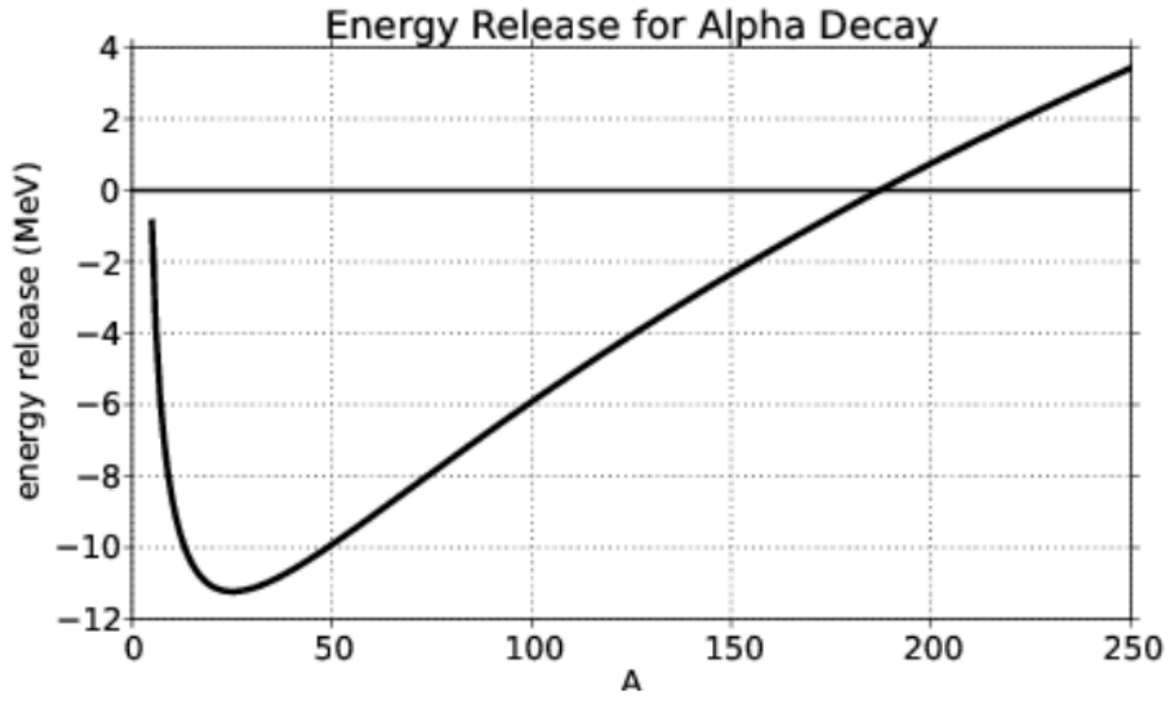


Figure 21.5: Approximate curve for the energy released in alpha decay of a nucleus on the line of stability. Decay is only possible if $Q > 0$.

Figure 21.6 shows schematically how alpha and beta decay transform atomic nuclei in the *N*-*Z* plane.

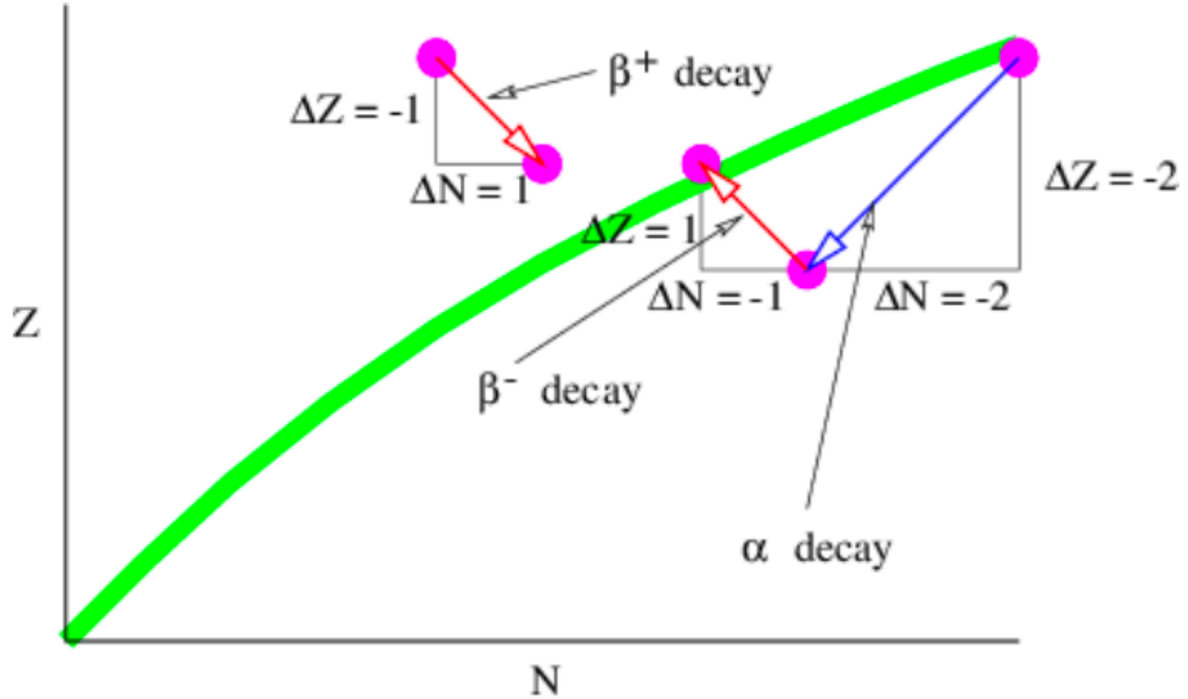


Figure 21.6: Schematic illustration of the paths of nuclear transformations in the *N*-*Z* plane due to alpha and beta decay. The thick line represents the line of stability. β^+ is the decay of a proton into a neutron, positron, and electron neutrino, while β^- is the decay of a neutron into a proton, electron, and electron antineutrino.

As previously indicated, alpha decay decreases both *Z* and *N* by two.

Ordinary beta decay (i. e., $n \rightarrow p^+ + e^- + \bar{\nu}_e$) decreases *N* by one and increases *Z* by one.

This is sometimes called β- decay since it produces an electron with negative charge.

Though the proton in isolation is stable, the energetics of atomic nuclei are such that a nucleus with a higher proton-neutron ratio than specified by the line of stability can sometimes release energy by the reaction $p^+ \longrightarrow n + e^+ + \nu_e$.

This is called β^+ decay since it produces a positively charged positron.

Certain isotopes of very heavy elements are at the head of a chain of radioactive decays.

This chain consists of a combination of alpha decays interspersed with β^- decays.

The latter are needed because the alpha decays create nuclei with too low a ratio of protons to neutrons relative to the line of stability, as illustrated in figure 21.6.

The beta decays push the chain back toward this line.

An example of a chain is one that starts with the element thorium ($Z = 90, A = 232$) and ends with lead ($Z = 82, A = 208$).

Radioactive decay thus accomplishes what medieval alchemists tried, but failed, to do: transmute elements from one type into another.

Unfortunately, no radioactive chain ends at the element gold!

Radioactive decay is governed by a simple law, namely that the rate at which nuclei decay is proportional to the number of remaining nuclei.

In mathematical terms, this is expressed as follows:

$$\frac{dN}{dt} = -\lambda N, \tag{21.6}$$

where $N(t)$ is the number of remaining nuclei at time t and λ is called the *decay rate*.

This differential equation has the solution

$$N(t) = N(0)\exp(-\lambda t) \quad \text{(radioactive decay)}, \tag{21.7}$$

which shows that the number of nuclei decreases exponentially with time.

The *half-life*, $t_{1/2}$ of a certain nuclear type, is the time required for half the nuclei to decay.

Setting $N(t_{1/2}) = N(0)/2$, we find that

$$t_{1/2} = \frac{\ln(2)}{\lambda} \quad \text{(half-life)}. \tag{21.8}$$

The nature of exponential decay means that half the particles are left after one half-life, a quarter after two half-lives, an eighth after three half-lives, etc.

The actual value of λ, and hence $t_{1/2}$, depends on the character of the nucleus in question, with half-lives ranging from a small fraction of a second to many billions of years.

21.4 Nuclear Fusion and Fission

From figure 21.4 it is clear that atomic nuclei with $A < 60$ can combine to form more tightly bound nuclei and in so doing release energy.

This is called *nuclear fusion* and it is the process that powers stars.

It is not easy to fuse two nuclei.

As figure 21.7 shows, the nuclear force, which is attractive but short in range, and the Coulomb force, which is repulsive, combine to create a potential barrier that must be surmounted in order to release energy from fusion.

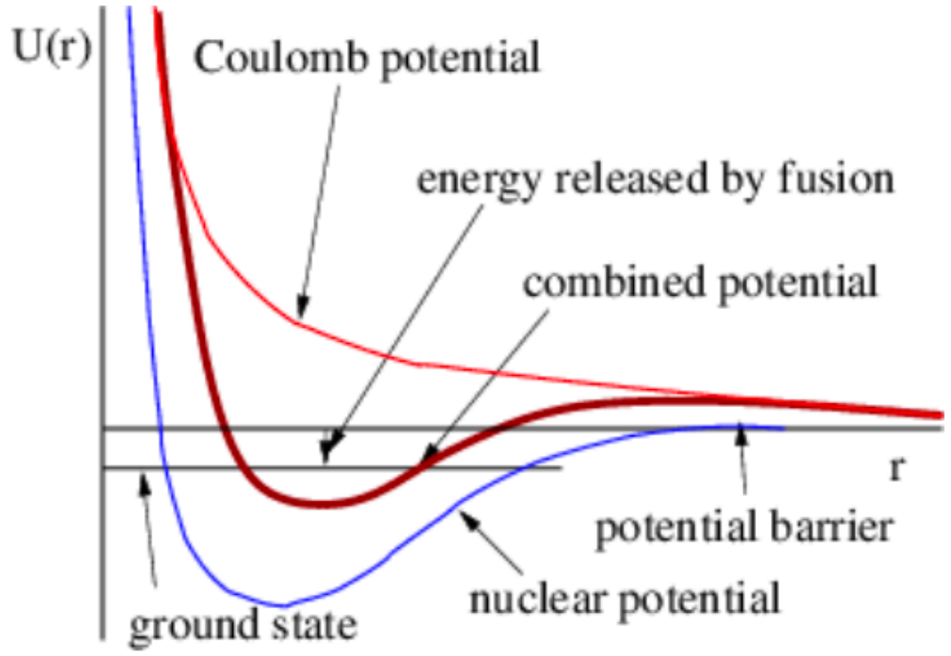


Figure 21.7: Combined nuclear and Coulomb potentials between two light nuclei. The resulting potential barrier repels the two nuclei unless their kinetic energy is very large. However, if the nuclei are able to overcome this barrier, substantial energy is released.

Nuclei must therefore somehow attain large kinetic energy for fusion to take place.

We shall discover later that temperature is a measure of the translational kinetic energy of atoms and nuclei.

Therefore, one way to create fusion is to heat the appropriate material to a very high temperature.

The interiors of ordinary stars are hot enough to fuse hydrogen into helium.

Somewhat hotter stars can create slightly heavier elements.

However, we believe that only the interior of a type of exploding star called a *supernova* is hot enough to create the heavy elements we find in the universe.

Thus, the iron in your automobile engine and the copper in your electrical wiring were created in some of the most spectacular explosions in the universe!

In computing energy balances for light nuclei, it is important to use exact values of binding energies, not the approximate values obtained from the binding energy formula given by equation (21.1), as the values given by this equation for small A can be off by a large amount.

Sample values for such nuclei are given in table 21.1.

Table 21.1: Binding energies of light nuclei.

Nucleus	Z	A	B (MeV)
deuterium	1	2	2.22
tritium	1	3	8.48
helium-3	2	3	7.72
helium-4	2	4	28.30
lithium-6	3	6	32.00
lithium-7	3	7	39.25

It is possible for a heavy nucleus such as uranium, with atomic number and atomic mass number (Z, A) to spontaneously fission or split into two lighter nuclei with (Z', A') and (Z - Z', A - A') if there is a net energy release from this process:

$$Q \equiv B(Z - Z', A - A') + B(Z', A') - B(Z, A) > 0 \quad \text{(fission possible)}. \qquad (21.9)$$

An energy of order 160 MeV per nucleus can be released by causing uranium ($Z = 92$) or plutonium ($Z = 94$) to fission.

Even if $Q > 0$, spontaneous fission generally occurs at a very slow rate.

This is because a potential energy barrier of order 5 MeV typically must be overcome for this split to occur.

Barrier penetration allows fission to occur spontaneously in the absence of the energy needed to overcome this barrier, as illustrated in figure 21.8, but is generally a slow process

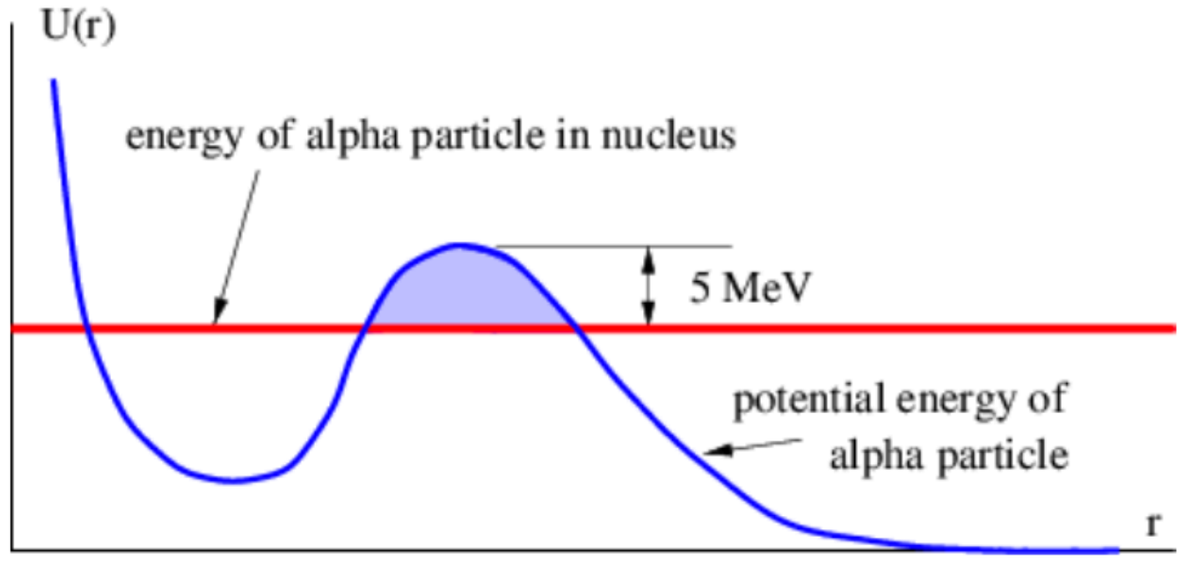


Figure 21.8: Spontaneous fission of a heavy nucleus into a slightly lighter nucleus and an alpha particle occurs when the alpha particle penetrates the potential barrier illustrated by the shading and leaves the nucleus. Other types of spontaneous fission occur in a similar manner. Compared to the case of two light nuclei in figure 21.7, the Coulomb potential is more important here, which makes the resultant force more repulsive.

Alpha decay is an example of spontaneous fission of a heavy nucleus by barrier penetration in which $Z' = 2$ and $A' = 4$.

If a heavy nucleus collides with an energetic particle such as a neutron, photon, or alpha particle, it can be induced to fission if the energy transferred to the nucleus exceeds the approximate 5 MeV needed to breach the potential barrier.

If the heavy nucleus has an odd number of neutrons, another way for fission to occur is for the nucleus to capture a slow neutron, i. e., one with energy much less than the 5 MeV needed to directly overcome the potential barrier.

In this case neutron capture actually converts the nucleus from atomic number and mass (Z, A) to atomic number and mass (Z, $A + 1$).

The binding energy per nucleon of a nucleus with an even number of neutrons is greater than the binding energy per nucleon of one with an odd number, since in the former case all neutron spins are paired.

Thus, if the initial nucleus has an odd number of neutrons, the capture of a slow neutron makes it more tightly bound than if the initial nucleus has an even number of neutrons.

If the difference in binding energy between the initial nucleus and the nucleus modified by neutron capture exceeds the 5 MeV needed to overcome the potential barrier for spontaneous fission, then energy conservation leaves the new nucleus in a sufficiently high excited state that it instantly fissions.

Examples of nuclei subject to fission by slow neutron absorption are uranium 235 and plutonium 239.

Note that both have odd numbers of neutrons.

In contrast, uranium 238 has an even number of neutrons and slow neutron bombardment does not cause fission.

Chapter 22

Heat, Temperature, and Friction

Human beings have long had an intuitive understanding of heat and temperature from personal experience.

We sense that different things often have different temperatures and we know that objects tend to acquire the same temperature after being placed in physical contact for some time.

We view this equilibration process as a flow of “heat” (whatever that is) from the warmer body to the cooler body.

A need for a more precise understanding of the behavior of heat and temperature was felt with the development of the steam engine.

The science of *thermodynamics* arose out of this need.

Thermodynamics was developed before we understood the atomic nature of matter.

More recently the ideas of thermodynamics were related to mechanical processes happening on the atomic scale.

Today we understand the phenomena of heat and temperature to be aspects of the collective mechanical behavior of large numbers of atoms and molecules.

22.1 Temperature

We measure temperature by a variety of means.

The most primitive measurement is direct sensing by the human body.

We immediately discern whether something we touch is hot or cold relative to our own body.

Furthermore, we can detect a hot stove from a distance by the feeling of warmth on our skin.

In the case of direct contact, heat is transferred to our hand by *conduction*, whereas in the latter case the transfer takes place by *thermal radiation*.

Our body considers something to be hot if heat is transferred from the object to our body, whereas it is perceived as being cold if the transfer of heat is from our body to the object.

A more objective measure of temperature is obtained by using the fact that ordinary material objects expand when they become warmer and contract when they cool.

Empirically it is found that the fractional change in the length of a solid body, $\Delta L/L$, is related to the change in temperature ΔT, as illustrated in figure 22.1:

$$\frac{\Delta L}{L} = \alpha \Delta T, \tag{22.1}$$

where α is called the *linear coefficient of thermal expansion*.

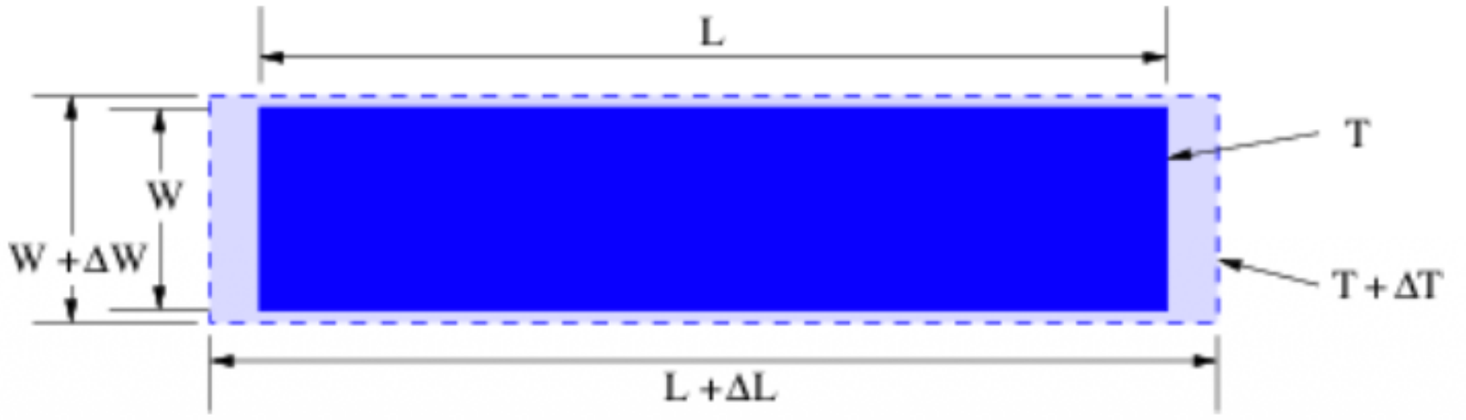


Figure 22.1: Most solid bodies expand by the same fractional amount in all directions when their temperature increases, so that $\Delta L/L = \Delta W/W$. Thus, the ratio $\alpha = \Delta L/(L dT)$ is the same for all objects constructed of the same material, generally over a considerable range of temperature.

For liquids the fractional change in volume, $\Delta V/V$, is easier to relate to the change in temperature than the fractional change in linear dimension:

$$\frac{\Delta V}{V} = \beta \Delta T, \qquad (22.2)$$

where β is the volume coefficient of thermal expansion.

The quantities α and β depend on the material properties and on the temperature scale being used.

The ordinary thermometer is based on the thermal expansion of a liquid such as mercury.

The most commonly used temperature scales in science are the Celsius and Kelvin scales.

Roughly speaking, water freezes at 0^o C and it boils (at sea level) at 100^o C.

More precise definition of the Celsius scale depends on a detailed understanding of the phase changes of water, which we won't develop here.

There is a limit to how cold something can be.

The Kelvin scale is designed to go to zero at this minimum temperature.

The relationship between the Kelvin temperature T and the Celsius temperature T_C is

$$T = T_C + 273.15. \tag{22.3}$$

Thus, water freezes at about 273 K and boils at about 373 K.

Notice that the little circle or degree sign is used for Celsius temperatures but not Kelvin temperatures.

Unless otherwise noted, we will use the Kelvin scale.

Table 22.1 gives values of α and β for some common materials.

Table 22.1: Values of the linear coefficient of thermal expansion for common solids and the volume coefficient of expansion for common liquids. Invar is an alloy that is specifically formulated to have a low coefficient of thermal expansion.

Material	α (K^{-1})	β (K^{-1})
steel	12×10^{-6}	—
copper	16×10^{-6}	—
aluminum	23×10^{-6}	—
invar	0.7×10^{-6}	—
glass	9×10^{-6}	—
lead	29×10^{-6}	—
methyl alcohol	—	1.22×10^{-3}
glycerine	—	0.53×10^{-3}
mercury	—	0.182×10^{-3}
water (15° C)	—	0.15×10^{-3}
water (35° C)	—	0.35×10^{-3}
water (90° C)	—	0.70×10^{-3}

Accurate temperature measurements depend in practice on a knowledge of the properties of materials under temperature changes.

However, we shall find later that the concept of temperature can be defined in a way that is completely independent of material properties.

22.2 Heat

Two types of experiments suggest that heating is a form of energy transfer.

First of all, on the macroscopic or everyday scale of things, there are forces that are apparently nonconservative.

This is in marked contrast to the microscopic world, where forces are either conservative (gravity, electrostatics), don't change a particle's energy (magnetic force), or convert energy from one known form to another (non-static electric forces).

With these fundamental forces all energy is accounted for — it is neither created or destroyed.

In contrast, macroscopic energy routinely disappears in the everyday world.

Cars once set in motion don't continue in motion forever on a level road once the engine is stopped; a soccer ball once kicked eventually comes to rest; electrical energy powering a light bulb appears to be lost.

Careful measurements show that whenever this type of energy loss is found, heating occurs.

Since we believe that macroscopic forces are really just large scale manifestations of fundamental microscopic forces, we do not believe that energy really disappears as a result of these forces — it must simply be converted from a form visible to us into an invisible form.

We now know that such forces convert macroscopic energy to internal energy, a form of energy that is just the kinetic and potential energy of atomic and molecular motions.

Thus, the apparent disappearance of macroscopic energy is just a consequence of the conversion of this energy into microscopic form.

The second type of experiment that suggests that heating converts macroscopic energy to internal energy is one in which this energy is converted back to macroscopic form.

An example of this process is illustrated in figure 22.2.

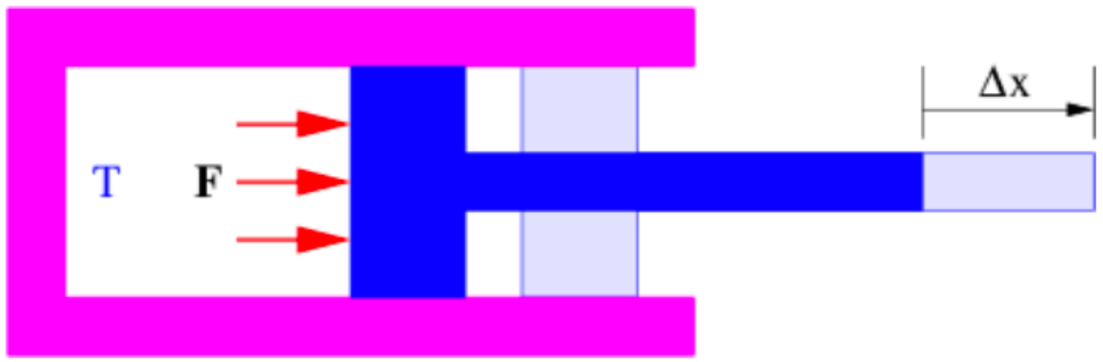


Figure 22.2: Conversion of internal energy of gas in the cylinder to macroscopic energy. The work done by the force of the gas on the piston as it moves outward results in a decrease in temperature of the gas.

As the piston moves out of the cylinder under the force exerted on it by the gas, work is done that can be stored or used by, say, compressing a spring or running an electric generator.

As the piston moves out, the gas in the cylinder decreases in temperature, which indicates that the gas is losing microscopic energy.

22.2.1 Specific Heat

Conversion of macroscopic energy to microscopic kinetic energy thus tends to raise the temperature, while the reverse conversion lowers it.

It is easy to show experimentally that the amount of heating needed to change the temperature of a body by some amount is proportional to the amount of matter in the body.

Thus, it is natural to write

$$\Delta Q = MC\Delta T \tag{22.4}$$

where M is the mass of material, ΔQ is the amount of energy transferred to the material, and ΔT is the change of the material's temperature.

The quantity C is called the *specific heat* of the material in question and is the amount of heating needed to raise the temperature of a unit mass of material by one degree.

C varies with the type of material.

Values for common materials are given in table 22.2.

Table 22.2: Specific heats of common materials.

Material	C (J kg^{-1} K^{-1})
brass	385
glass	669
ice	2092
steel	448
methyl alcohol	2510
glycerine	2427
water	4184

22.2.2 First Law of Thermodynamics

We now address some questions of terminology.

The use of the terms “heat” and “quantity of heat” to indicate the amount of microscopic kinetic energy inhabiting a body has long been out of favor due to their association with the discredited “caloric” theory of heat.

Instead, we use the term *internal energy* to describe the amount of microscopic energy in a body.

The word *heat* is most correctly used only as a verb, e. g., “to heat the house”.

Heat thus represents the transfer of internal energy from one body to another or conversion of some other form of energy to internal energy.

Taking into account these definitions, we can express the idea of energy conservation in some material body by the equation

$$\Delta E = \Delta Q - \Delta W \quad \text{(first law of thermodynamics)}, \tag{22.5}$$

where ΔE is the change in internal energy resulting from the *addition* of heat ΔQ to the body and the work ΔW done *by* the body *on* the outside world.

This equation expresses the *first law of thermodynamics*.

Note that the sign conventions are inconsistent as to the direction of energy flow.

However, these conventions result from thinking about *heat engines*, i. e., machines that take in heat and put out macroscopic work.

Examples of heat engines are steam engines, coal and nuclear power plants, the engine in your automobile, and the engines on jet aircraft.

22.2.3 Heat Conduction

As noted earlier, internal energy may be transferred through a material from higher to lower temperature by a process known as *heat conduction*.

The rate at which internal energy is transferred through a material body is known empirically to be proportional to the temperature difference across the body.

For a rectangular body, the rate of transfer is also known to scale in proportion to the cross sectional area of the body perpendicular to the temperature gradient and to scale inversely with the distance over which the temperature difference exists.

This is known as the law of heat conduction and is expressed in the following mathematical form:

$$F_{heat} = \frac{\kappa A \Delta T}{L}, \tag{22.6}$$

where F_{heat} is the internal energy per unit time flowing down the temperature gradient, A is the cross sectional area of the body normal to the internal energy flow direction, L is the length of the body in the direction of heat flow, ΔT is the temperature difference along its length, and κ is a constant characteristic of the material known as the thermal conductivity.

The geometry is illustrated in figure 22.3 and the thermal conductivities of common materials are shown in table 22.3.

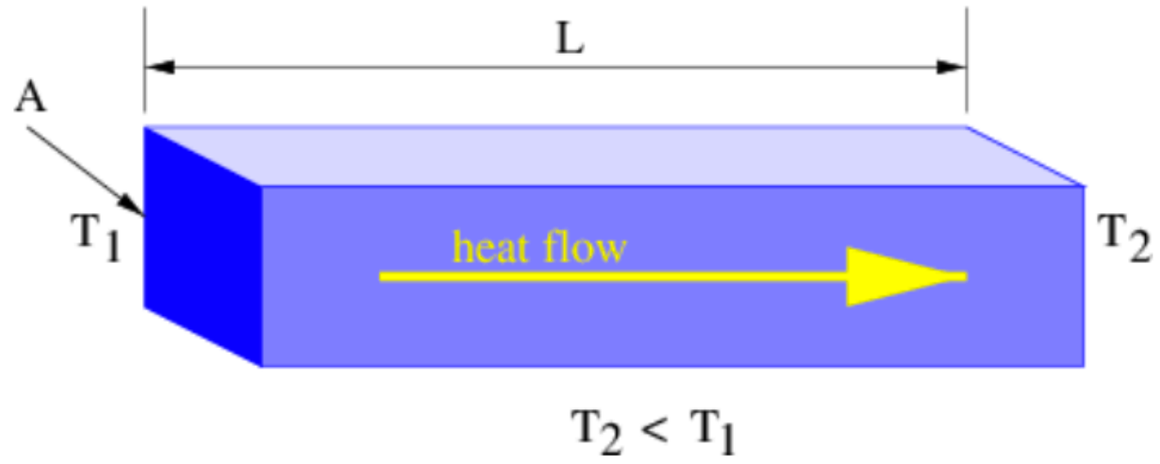


Figure 22.3: Geometry of heat flow problem. Heat flows from higher to lower temperature.

Material	$\varkappa$ (W m^{-1} K^{-1})
brass	109
brick	0.50
concrete	1.05
ice	2.2
paper	0.050
steel	46

Table 22.3: Values of thermal conductivity for common materials.

22.2.4 Thermal Radiation

Energy can also be transmitted though empty space by *thermal radiation*.

This is nothing more than photons with a mixture of frequencies near a frequency $\omega_{thermal}$ that is a function only of the temperature T of the body that is emitting them:

$$\omega_{thermal} = KT, \tag{22.7}$$

where the constant $K = 3.67 \times 10^{11}$ s^{-1} K^{-1}.

The amount of thermal energy per unit area per unit time emitted by a material surface is called the *flux* of radiation and is given by *Stefan's law*:

$$J_E = \varepsilon\sigma T^4 \quad \text{(Stefan's law)}, \tag{22.8}$$

where $\sigma = 5.67 \times 10^{-8}$ W m^{-2} K^{-4} is the Stefan-Boltzmann constant and ε is the emissivity of the material surface.

The emissivity lies in the range $0 \leq \varepsilon \leq 1$ and depends on the type of material and the temperature of the surface.

Surfaces that emit thermal radiation at a particular frequency can also reflect radiation at that frequency.

If J_I is the flux of radiation incident on the surface, then the reflected radiation is just

$$J_R = (1 - \varepsilon) J_I \quad \text{(reflected radiation)} \tag{22.9}$$

and the balance of the radiation is absorbed by the surface:

$$J_A = \varepsilon J_I \quad \text{(absorbed radiation)}. \tag{22.10}$$

Thus, high thermal emissivity goes along with high absorbed fraction and vice versa.

A little thought indicates why this has to be so.

If the emissivity were high and the absorption were low, then the object would spontaneously cool relative to its environment.

If the reverse were true, it would spontaneously warm up.

Thus, the universally observed behavior that internal energy flows from higher to lower temperatures would be violated.

Imagine two surfaces of equal temperature T facing each other.

The radiation emitted by one surface is partially absorbed and partially reflected from the other surface, as illustrated in figure 22.4.

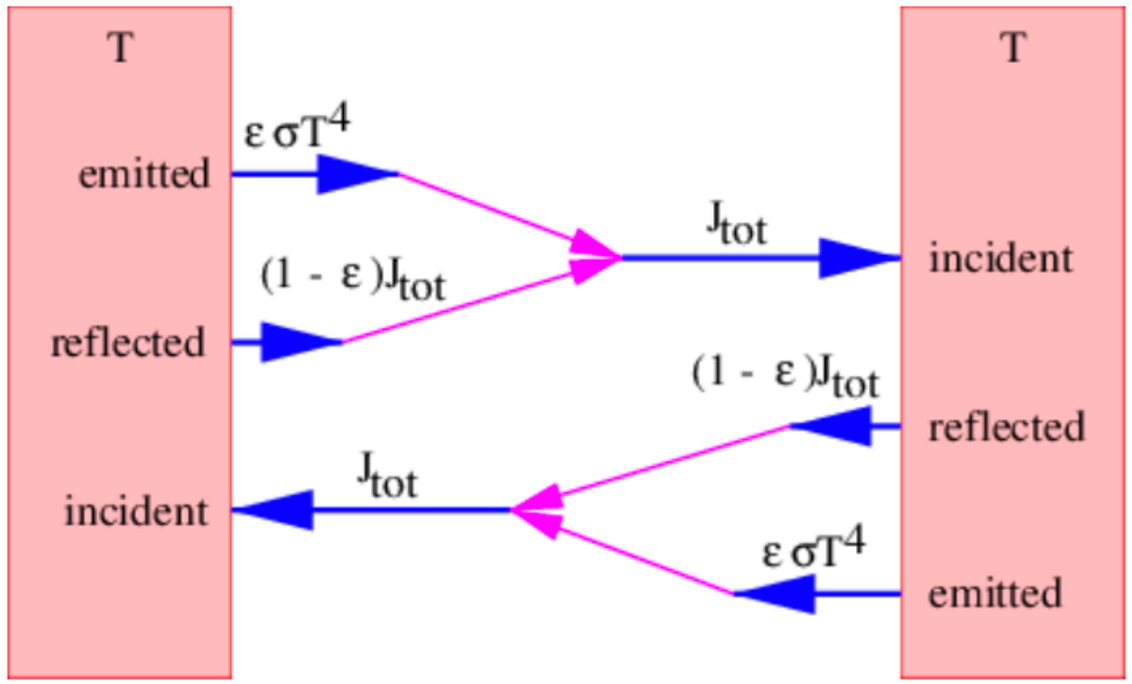


Figure 22.4: Two surfaces facing each other, each with emissivity ε and temperature T.

The total radiative flux, J_{tot}, coming from each surface is the sum of the reflected radiation originating from the other surface, $(1 - \varepsilon)J_{tot}$, and the emitted thermal radiation, $\varepsilon\sigma T^4$.

Thus,

$$J_{tot} = (1 - \varepsilon)J_{tot} + \varepsilon\sigma T^4. \tag{22.11}$$

Solving for J_{tot}, we find that

$$J_{tot} \equiv J_{BB} = \sigma T^4. \tag{22.12}$$

Note that the total radiation originating from each surface, J_{tot}, is independent of the emissivity of the surfaces and depends only on the temperature.

This radiative flux is called the *black body flux*.

We give it the special name J_{BB}.

Because it no longer depends on ε, it is independent of the character of the material making up the emitting surfaces.

Different materials result in different fractions of thermal and reflected radiation, but the sum is always equal to the black body flux if both surfaces are at the same temperature.

Planck's arguments that led to the energy-frequency relationship of quantum mechanics, $E = \hbar\omega$, came from his attempt to explain black body radiation.

The laws of black body radiation presented here can be derived from quantum mechanics.

22.3 Friction

In this section we consider the quantitative forms of non-conservative forces on the macroscopic level.

We first examine the frictional force between two solid bodies and then consider viscosity in liquids.

22.3.1 Frictional Force Between Solids

The frictional force F_k between two solid objects in contact obeys an empirical law.

If the two objects are sliding over each other, the frictional force on each object acts so as to oppose the relative motion of the two objects. (See figure 22.5.)

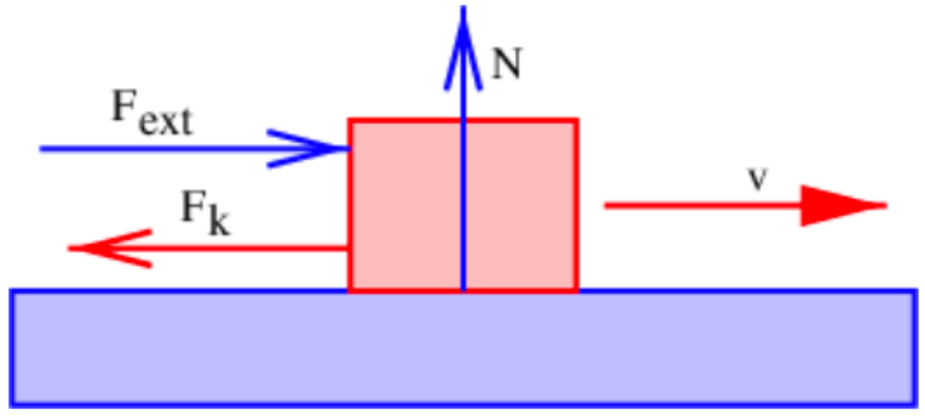


Figure 22.5: The kinetic frictional force F_k is exerted on the upper body by the stationary lower body. The upper body is moving with velocity v and is pressed together with the lower body by a normal force N. It may also be acted upon by an additional external force F_{ext}.

The frictional force is proportional to the normal force N pressing the objects together:

$$F_k = \mu_k N \quad \text{(kinetic friction)}. \tag{22.13}$$

The dimensionless quantity μ_k is called the *coefficient of kinetic friction*.

This quantity is different for different pairs of materials rubbing together.

It is typically of order one, but may be much less for particularly slippery materials.

Equation (22.13) is only valid if the two objects are moving relative to each other.

If they are not in relative motion, but if some other force is being exerted on one of them, a static frictional force F_s will precisely counteract this force so as to result in zero net force on the object.

However, the static frictional force will keep the bodies from slipping only up to some limit defined by

$$|F_s| \leq \mu_s N \quad \text{(static friction)}, \tag{22.14}$$

where μ_s is the coefficient of static friction.

Generally we find that $\mu_s > \mu_k$, so gradually increasing the external force on an object in static frictional contact with another object will cause it to suddenly break loose and accelerate when the maximum sustainable static frictional force is exceeded.

Once the object is in motion, a lesser external force is needed to keep it moving at a constant velocity.

22.3.2 Viscosity

If two objects are not in physical contact but are separated by a thin layer of fluid (i. e., a liquid or a gas), there is still a frictional or viscous drag force between the two objects, but its behavior is different.

Figure 22.6 tells the story.

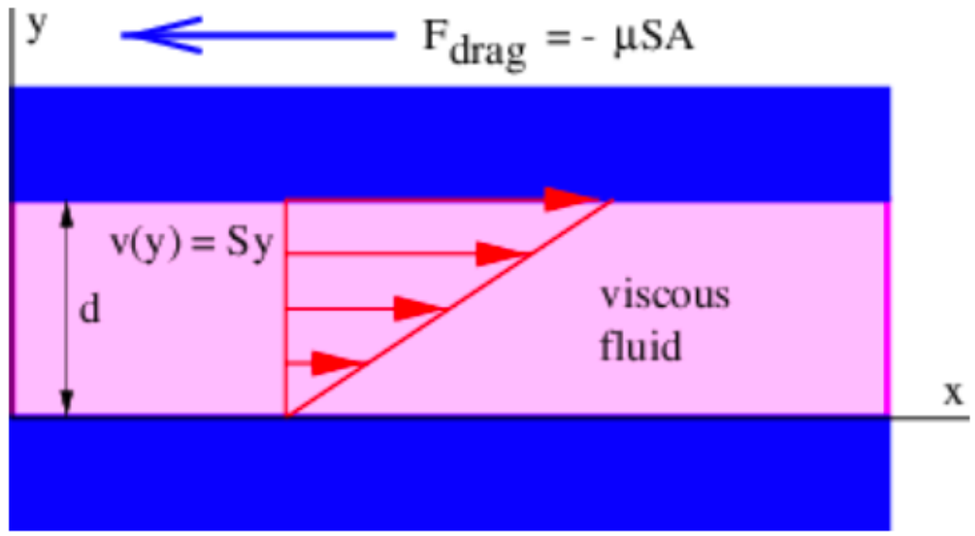


Figure 22.6: Two solid plates separated by a distance d, the gap being filled by a viscous fluid. The lower plate is stationary and the upper plate is moving to the right at speed $v_p = v(d) = Sd$. The fluid is sheared, moving according to $v(y) = Sy$. The fluid velocity matches that of the plates where the fluid touches the plates. The upper plate experiences a drag force $F_{drag} = -\mu SA$ where μ is the viscosity of the fluid and A is the area of the plate.

The viscous drag force in this case is

$$F_{drag} = -\mu SA \tag{22.15}$$

where $S = v_p / d$ is the *shear* in the fluid, A is the area of the plates, and μ is the *viscosity* of the fluid.

Don't confuse this parameter with the static and dynamic coefficients of friction!

The parameter v_p is the velocity of the top plate with respect to the bottom plate and d is the separation between the plates.

Viscosity has the dimensions mass per length per time.

The most common unit of viscosity is the *Poise*: 1 Poise = 1 g cm^{-1} s^{-1}.

The viscosity of water varies from 0.0179 Poise at 0^oC to 0.0100 Poise at 20^o C to 0.0028 Poise at 100^o C.

The viscosity of water thus decreases with increasing temperature, which is typical of liquids.

In contrast, the viscosity of a gas is independent of the density of the gas and is proportional to the square root of its absolute temperature.

The viscosity of a gas thus increases with temperature, in contrast to the viscosity of a liquid.

For air at 20^o C, the viscosity is 1.81×10^{-4} Poise.

Thin layers of oil between moving parts are commonly used in machinery to reduce friction, since the resulting viscous drag is generally much less than the corresponding kinetic friction that would occur if the parts were in direct contact.

The ways in which the layer of oil is maintained between moving parts are fascinating, but beyond the scope of this course.

Chapter 23

Entropy

So far we have taken a purely empirical view of the properties of systems composed of many atoms.

However, as previously noted, it is possible to understand such systems using the underlying principles of mechanics.

The resulting branch of physics is called *statistical mechanics*. J. Willard Gibbs, a late 19th century American physicist from Yale University, almost single-handedly laid the groundwork for the modern form of this subject.

Interestingly, the quantum mechanical version of statistical mechanics is much easier to understand than the version that Gibbs developed, which is based on classical mechanics.

It also gives correct answers where the Gibbs version fails.

A system of many atoms has many quantum mechanical states in which it can exist.

Think of, say, a brick.

The atoms in a brick are not stationary; they are in a continual flurry of vibration at ordinary temperatures.

The kinetic and potential energies associated with these vibrations constitute the internal energy of the brick.

Though the details of each state are unimportant, the *number* of states turns out to be a crucial piece of information.

To understand why this is so, let us imagine two bricks identical in composition and mass.

Brick A has internal energy between E and $E + \Delta E$ and brick B has energy between 0 and ΔE.

Think of ΔE as the uncertainty in the energy of the bricks; we can only observe a brick for a finite amount of time Δt, so the uncertainty principle asserts that the uncertainty in the energy is $\Delta E \approx \hbar/\Delta t$.

The brick is a complex system consisting of many atoms, so in general there are many possible quantum mechanical states available to brick A in the energy range E to $E + \Delta E$.

It turns out, for reasons that we will see later, that significantly fewer states are available to brick B in the energy range 0 to ΔE than are available to brick A.

Roughly speaking, the larger the internal energy of an object per unit mass, the higher is its temperature.

Thus, we infer that brick A has a much higher temperature than brick B.

What happens when we bring the two bricks into thermal contact?

Our experience tells us that heat (i. e., internal energy) immediately starts to flow from one brick to the other, ultimately resulting in an equilibrium state in which the temperature is the same in the two bricks.

We explain this process as follows.

Statistical mechanics hypothesizes that any system of atoms (such as a brick) is free to roam through all quantum mechanical states that are energetically available to it.

In fact, this roaming is assumed to be continually taking place.

Given this picture and the assumption that the roaming between states is completely random, one would expect equal probabilities for finding the system in any particular state.

Of course, this probability argument assumes that we don't know anything about the initial state of the system.

If the system is known to be in some particular state at time $t = 0$, then it will take some time for the system to evolve in such a way that it has "forgotten" the initial state.

During this interval our knowledge of the initial state and the quantum mechanical dynamics of the system can be used (in principle) to follow the evolution of the system.

Eventually the uncertainty in our initial knowledge of the system catches up with us and we cannot predict the future evolution of the system beyond this point.

The brick develops "amnesia" and its probability of being in any of the energetically allowed states is then uniform.

Something like this happens to the two bricks if they are brought into thermal contact. Initially brick A has virtually all of the energy and brick B has only a tiny amount.

When the bricks are brought into contact, they eventually can be treated as a single brick of twice the size.

However, it takes time for the new, larger brick to evolve to the point where it has forgotten the fact that it started out as two separate bricks at different temperatures.

In this interval the temperature of brick A is decreasing while the temperature of brick B is increasing as a result of internal energy flowing from one to the other.

This evolution continues until equilibrium is reached.

Even though the combined brick has forgotten its initial state, there is a small chance that it will return to this state, since the probability of finding the brick in any state, including the original one, is non-zero.

Thus, according to the postulates of statistical mechanics, one might suddenly find the brick again in a state in which virtually all of the internal energy is concentrated in former brick A.

Actually, the issue is slightly more complicated than this.

Brick A actually had many states available to it before being brought together with brick B.

Thus, a more interesting problem is to find the probability of the system suddenly finding itself in *any* of the states in which (virtually) all of the energy is concentrated in former brick A.

Given the randomness assumption of statistical mechanics, this probability is simply the number of states that correspond to all of the energy being in brick A, divided by the total number of states available to the combined brick.

Computing this number is the task we set for ourselves.

23.1 States of a Brick

In this section we demonstrate the above assertions by making a crude model of the quantum mechanical states of a brick.

We approximate the atoms of the brick as a collection of harmonic oscillators, three oscillators per atom, since each atom can oscillate in three dimensions under the influence of interatomic forces (see figure 23.1).

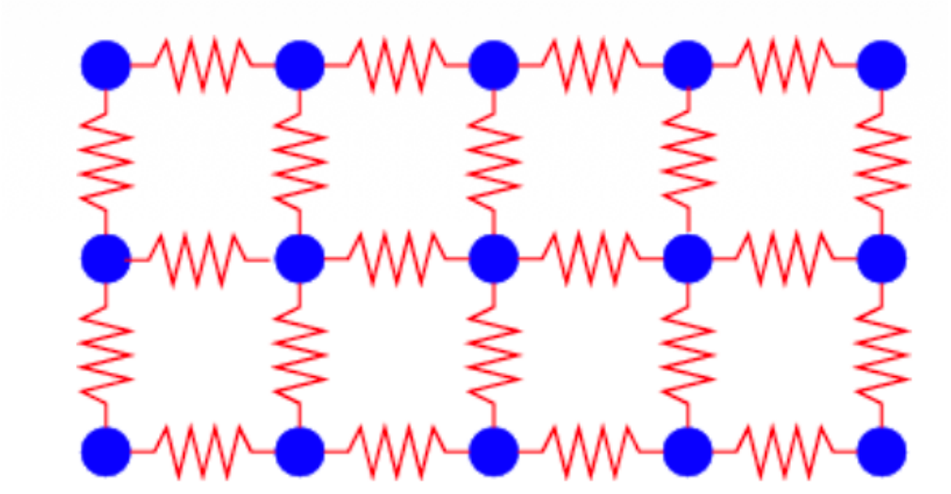


Figure 23.1: "Inner-spring mattress" model of the atoms in a solid body. Interatomic forces act like miniature springs conn

For simplicity we assume that all of the oscillators have the same classical oscillation frequency, ω_0, so that the energy of each oscillator is given by

$$E_n = (n + 1/2)\hbar\omega_0 \equiv (n + 1/2)E_0, \quad n = 0, 1, 2, \ldots, \tag{23.1}$$

as reported in chapter 12.

This assumption is a rather poor approximation to the behavior of a solid body when the total amount of internal energy is so small that many of the harmonic oscillators are in their ground state.

However, it is adequate for situations in which the energy per oscillator is several times the ground state oscillator energy.

We further assume that each oscillator is weakly coupled to its neighbor.

This allows a slow transfer of energy between oscillators without appreciably affecting the energy levels of each oscillator.

The next step is to calculate the number of states of a system of harmonic oscillators for which the total energy is less than some maximum value E.

This calculation is easy for a system consisting of a single oscillator.

From equation (23.1) we infer that the number of states, , of one oscillator with energy less than E is

$$\mathcal{N} = E/E_0 \quad \text{(one oscillator)}, \tag{23.2}$$

since the states are evenly spaced in energy with spacing E_0.

The calculation for a system of two oscillators is slightly more complicated.

The dots in the left panel of figure 23.2 show the states available to a two oscillator system.

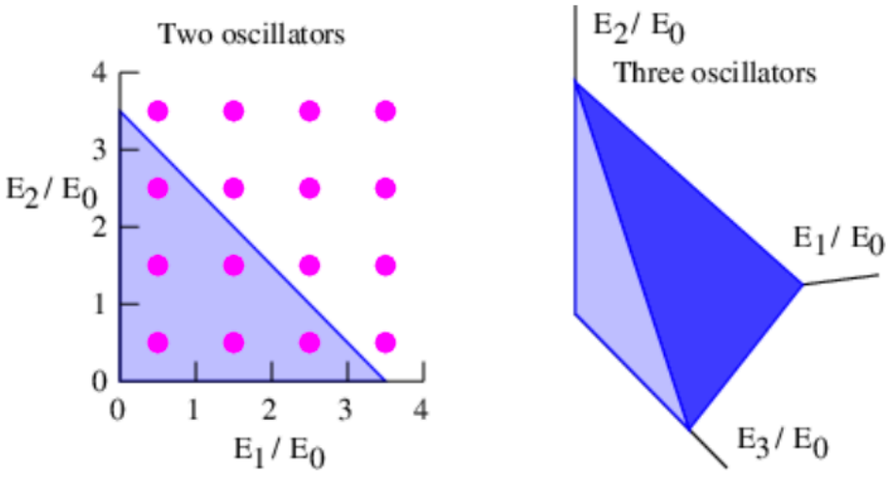


Figure 23.2: Diagrams for counting states of systems of two (left panel) and three (right panel) harmonic oscillators with the same classical oscillation frequency.

Each dot corresponds to a unique pair of values of the quantum numbers n_1 and n_2 for the two oscillators.

The total energy of the two oscillators together is $E_{total} = E_1 + E_2 = (n_1 + n_2 + 1)E_0$.

The line defined by the equation $E/E_0 = E_1/E_0 + E_2/E_0$ is illustrated by the hypotenuse of the shaded triangle in the left panel of figure 23.2.

The number of states with total energy less than E is obtained by simply counting the dots inside this triangle.

An easy way to do this "counting" is to note that there is one dot per unit area in the plot, so that the number of dots approximately equals the area of the triangle:

$$\mathcal{N} = \frac{1}{2}\left(\frac{E}{E_0}\right)^2 \quad \text{(two oscillators)}. \tag{23.3}$$

For a system of three oscillators the possible states of the system form a cubical grid in a three-dimensional space with axes E_1/E_0, E_2/E_0, and E_3/E_0, as shown in the right panel of figure 23.2.

The dots representing the states are omitted for clarity, but one state per unit volume exists in this space.

The dark-shaded oblique triangle is the surface of constant total energy E defined by the equation $E_1/E_0 + E_2/E_0 + E_3/E_0 = E/E_0$, so the volume of the tetrahedron formed by this surface and the coordinate axis planes equals the number of states with energy less than E.

This volume is computed as the area of the base of the tetrahedron, $(E/E_0)^2/2$, times its height, E/E_0, times $1/3$.

We get

$$\mathcal{N} = \frac{1}{2 \cdot 3} \left(\frac{E}{E_0} \right)^3 \quad \text{(three oscillators)}. \tag{23.4}$$

There is a pattern here.

We infer that there are

$$\mathcal{N}(E) = \frac{1}{1 \cdot 2 \cdot 3 \ldots N} \left(\frac{E}{E_0} \right)^N = \frac{1}{N!} \left(\frac{E}{E_0} \right)^N \quad (N \text{ oscillators}) \tag{23.5}$$

states available to N oscillators with total energy less than E.

The notation $N!$ is shorthand for $1 \cdot 2 \cdot 3 \ldots \cdot N$ and is pronounced "*N factorial*".

Let us summarize what we have accomplished.

$\mathcal{N}(E)$ is the number of states of a system of harmonic oscillators, taken together, with total energy less than E.

What we need is an estimate of the number of states between two energy limits, say E and $E + \Delta E$.

This is easily obtained from $\mathcal{N}(E)$ as follows: $\mathcal{N}(E)$ is the number of states with energy less than E, while $\mathcal{N}(E + \Delta E)$ is the number of states with energy less than $E + \Delta E$.

We can obtain the number of states with energies between E and $E + \Delta E$ by subtracting these two quantities:

$$\Delta\mathcal{N} = \mathcal{N}(E + \Delta E) - \mathcal{N}(E) = \frac{\mathcal{N}(E + \Delta E) - \mathcal{N}(E)}{\Delta E}\Delta E \approx \frac{\partial \mathcal{N}}{\partial E}\Delta E. \tag{23.6}$$

For N harmonic oscillators we find that

$$\Delta\mathcal{N} = \frac{N}{N!}\frac{E^{N-1}}{E_0^N}\Delta E = \frac{1}{(N-1)!}\left(\frac{E}{E_0}\right)^{N-1}\frac{\Delta E}{E_0}. \tag{23.7}$$

Table 23.1 shows the number of states of a system of a small number of harmonic oscillators with energy between E and $E + \Delta E$ where we have chosen $\Delta E = E_0$.

Table 23.1: Number of states $\Delta\mathcal{N}$ available to N identical harmonic oscillators between energies E and $E + \Delta E$, where $E = rNE_0$ and where we have chosen $\Delta E = E_0$. Results are shown for two different values of r.

N	$\Delta\mathcal{N}$ $(r = 5)$	$\Delta\mathcal{N}$ $(r = 10)$
1	1	1
2	5	10
3	50	200
4	563	4500
5	6667	106667
6	81381	2604167
7	1012500	64800000
8	12765734	1634013889
9	162539683	41610158730
10	2085209002	1067627008928
11	26911444555	27557319223986
12	349006782021	714765889577822

Results are shown for systems up to $N = 12$ (i. e., "microbricks" with up to 4 atoms, each with 3 modes of oscillation).

The quantity r is defined to be the average value of the quantum number n of all the harmonic oscillators in the system; $r = E/(NE_0)$.

Thus, rE_0 is the average energy per oscillator.

Recall that our calculation is only valid if r is appreciably greater than one.

The number of available states is computed for $r = 5$ and 10.

We see that a few atoms considered jointly have an astonishingly large number of possible states.

For instance, a system of 4 atoms (i. e., 12 oscillators) with $r = 5$ has about 3.5×10^{11} states.

Suppose we now confine this energy to only 2 of the atoms or 6 oscillators.

In this case r doubles to a value of 10 since the same amount of internal energy is now spread among half the number of oscillators.

Table 23.1 shows that this reduced system has only about 2.6×10^{6} states.

The probability of having all of the energy of the 4 atom system in these 2 atoms is the ratio of the number of states in the 2 atom case to the total number of possible states of the 4 atom system, or $2.6 \times 10^{6} / 3.5 \times 10^{11} = 7.4 \times 10^{-6}$.

This is a rather small number, which means that it is rare to find the system with all internal energy concentrated in two atoms.

We now determine how the number of states available to a system of harmonic oscillators behaves for a very large number of oscillators such as might be found in a real brick.

Values of $\Delta\mathcal{N}$ become so large in this case that it is useful to work in terms of the natural logarithm of $\Delta\mathcal{N}$.

For large N we can safely approximate N -1 by N.

Using the properties of logarithms, we get

$$\begin{aligned}\ln(\Delta\mathcal{N}) &= \ln\left(\frac{(E/E_0)^{N-1}}{(N-1)!}\frac{\Delta E}{E_0}\right)\\ &\approx \ln\left(\frac{(E/E_0)^{N}}{N!}\frac{\Delta E}{E_0}\right)\\ &= N\ln(E/E_0) - \ln(N!) + \ln(\Delta E/E_0). \qquad (23.8)\end{aligned}$$

A useful mathematical result for large N is the *Stirling approximation*

$$\ln(N!) \approx N\ln(N) - N \quad \text{(Stirling approximation)}. \qquad (23.9)$$

Substituting this into equation (23.8), using the fact that $N\ln(E/E_0)$ - $N\ln N = N\ln[E/(NE_0)]$, and rearranging results in

$$\ln(\Delta\mathcal{N}) = N\left[\ln\left(\frac{E}{NE_0}\right)+1\right] + \ln\left(\frac{\Delta E}{E_0}\right) \quad (N \text{ oscillators}). \qquad (23.10)$$

We now return to the original question, which we state in this form: What fraction of the states of a brick corresponds to the special situation with all of the internal energy in half of the brick?

A real brick has of order 3×10^{25} atoms or about $N = 10^{26}$ oscillators.

Half of the brick thus has $N' = 5 \times 10^{25}$ oscillators.

If, as before, we assume that $r = 5$ when the internal energy is distributed throughout the brick, then we have $r' = 10$ when all the energy is in half of the brick.

Therefore the logarithm of the total number of available states is $\ln(\Delta\mathcal{N}) = N[\ln(r) + 1] + \ln(\Delta E/E_0)$, while the logarithm of the number of states available when all the energy is in half of the brick is $\ln(\Delta\mathcal{N}') = N'[\ln(r') + 1] + \ln(\Delta E/E_0)$.

Putting in the numbers, we find that the probability of finding all the energy in half of the brick is

$$\begin{aligned}
\Delta\mathcal{N}'/\Delta\mathcal{N} &= \exp[\ln(\Delta\mathcal{N}') - \ln(\Delta\mathcal{N})] \\
&= \exp[N' \ln(r') + N' + \ln(\Delta E/E_0) \\
&\qquad - N \ln(r) - N - \ln(\Delta E/E_0)] \\
&= \exp(-0.96N) = \exp(-9.6 \times 10^{25}) = 10^{-4.2\times 10^{25}}
\end{aligned} \tag{23.11}$$

This probability is *extremely* small, and is zero for all practical purposes.

Notice that ΔE, which we haven't specified, cancels out.

This typically happens in the theory when measurable quantities are calculated, and it shows that the actual value of ΔE isn't important.

Furthermore, for very large values of N typical of normal bricks, the term in equation (23.10) containing ΔE is always negligible for any reasonable values of ΔE.

We therefore drop it in future calculations.

The variable $\ln(\Delta\mathcal{N})$ is proportional to a quantity that we call the *entropy*, S. The actual relationship is

$$S = k_B \ln(\Delta\mathcal{N}) \quad \text{(definition of entropy)} \tag{23.12}$$

where $k_B = 1.38 \times 10^{-23}$ J K^{-1} is called *Boltzmann's constant*.

Ludwig Boltzmann was a 19th century Austrian physicist who played a pivotal role in the development of the concept of entropy.

The entropy of a brick containing N oscillators is therefore

$$S = Nk_B \left[\ln\left(\frac{E}{NE_0}\right) + 1\right] \quad \text{(entropy of } N \text{ oscillators).} \tag{23.13}$$

As with the speed of light and Planck's constant, Boltzmann's constant is not really needed for a complete development of statistical mechanics. Its only role is to convert entropy and related quantities to everyday units.

The conventional dimensions of entropy are thus the same as those of Boltzmann's constant, or energy divided by temperature.

However, more fundamentally, we consider entropy (without Boltzmann's constant) to be a dimensionless quantity since it is just the logarithm of the number of available states.

23.2 Second Law of Thermodynamics

What use is entropy?

In our example we found that the number of states for the situation in which all of the internal energy of a brick is restricted to half of the brick is much less than the number of states available when no restrictions are put upon the distribution of the same amount of internal energy through the entire brick.

Thus, the entropy, which is just proportional to the logarithm of the number of available states, is less in the restricted case than it is in the unrestricted case.

This turns out to be generally true.

Any measurable restriction we place on the distribution of internal energy in the brick turns out to result in a *much* smaller number of available quantum mechanical states and hence a smaller value for the entropy.

Once such a restriction is lifted, all possible states become available, and according to the postulates of statistical mechanics the brick eventually evolves to the point where it is roaming randomly through these states.

The probability of the brick revisiting the original restricted set of states is so small as to be completely ignorable once it forgets its initial state, because these states form only a miniscule fraction of the states available to the brick.

Thus, with a very high degree of certainty, one can say that the entropy of the brick increases when the restriction is lifted.

Strictly speaking, our definition of entropy is only valid after the brick has reached equilibrium, i. e., when the initial state has been forgotten.

The entropy during the equilibration period according to our definition is technically undefined.

Our inferences about a brick can be extended to any isolated system, i. e., any system that doesn't exchange mass or energy with the outside world: *The entropy of any isolated system consisting of a large number of atoms will not spontaneously decrease with time*.

This principle is called the *second law of thermodynamics*.

23.3 Two Bricks in Thermal Contact

Where does the idea of temperature fit into the picture?

This concept has come up informally, but we need to give it a precise definition.

If two objects at different temperatures are placed in contact with each other, we observe that internal energy flows from the warmer object to the cooler object, as illustrated in figure 23.3.

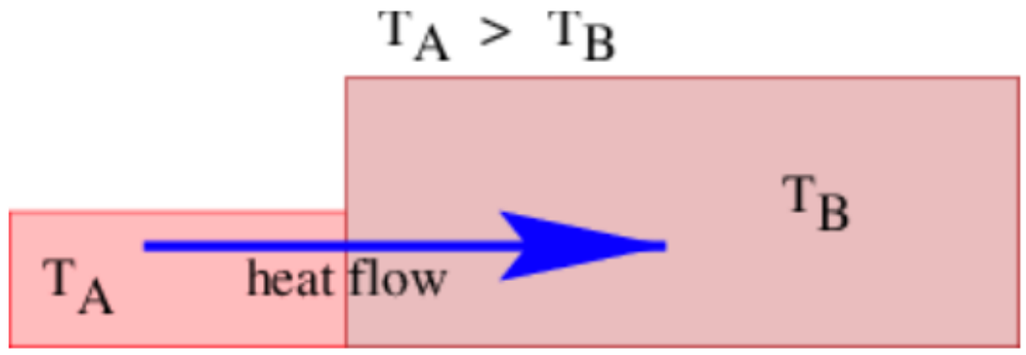


Figure 23.3: Two bricks in thermal contact, one at temperature T_A, the other at temperature T_B. If $T_A > T_B$, internal energy flows from brick A to brick B

We wish to see if the role of temperature differences in the flow of internal energy can be related to the ideas developed in the previous section.

Let us consider two bricks as before, but possibly of different size, and therefore containing different numbers of harmonic oscillators.

Suppose brick A has N_A oscillators and energy E_A while brick B has N_B oscillators and energy E_B.

The two bricks have entropies

$$S_A = k_B N_A \left[\ln \left(\frac{E_A}{N_A E_0} \right) + 1 \right] \tag{23.14}$$

and

$$S_B = k_B N_B \left[\ln \left(\frac{E_B}{N_B E_0} \right) + 1 \right]. \tag{23.15}$$

If the two bricks are thermally isolated from each other but are nevertheless considered together as one system, then the total number of states available to this combined system is just the product of the numbers of states available to each brick separately:

$$\Delta \mathcal{N} = \Delta \mathcal{N}_A \Delta \mathcal{N}_B. \tag{23.16}$$

To make an analogy, the total number of ways of arranging two coins, each of which may either be heads up or tails up, is $4 = 2 \times 2$, or heads-heads, heads-tails, tails-heads, and tails-tails.

We compute the states of the combined system just as we compute the total number of ways of arranging the coins, i. e., by taking the product of the numbers of states of the individual systems.

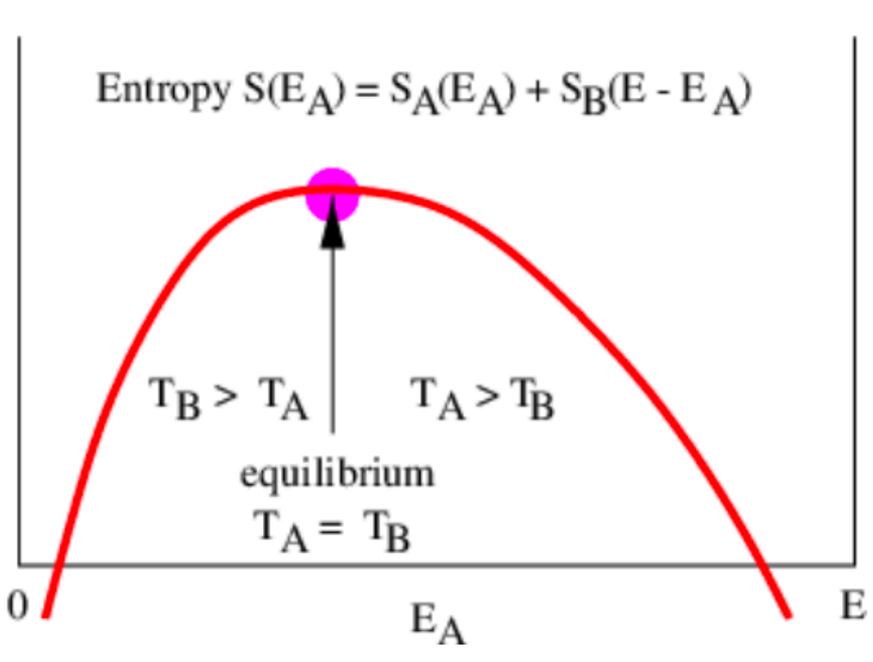
Taking the logarithm of $\mathcal{N}$ and multiplying by Boltzmann's constant results in an equation for the combined entropy S of the two bricks:

$$S = S_A + S_B. \tag{23.17}$$

In other words, the combined entropy of two (or more) isolated systems is just the sum of their individual entropies.

We can determine how the total entropy of the two bricks depends on the distribution of energy between them by using equations (23.14) and (23.15).

Plotting the sum of the entropies of the two bricks $S_A(E_A) + S_B(E_B)$ versus the energy E_A of brick A under the constraint that the total energy $E = E_A + E_B$ is constant yields a curve that typically looks something like figure 23.4.



The**Figure 23.4:** Total entropy of two systems for fixed total energy $E = E_A+E_B$ as a function of E_A, the energy of system A.

Notice that the total entropy reaches a maximum for some critical value of E_A.

Since the slope of $S(E_A)$ is zero at this point, we can determine the corresponding value of E_A by setting the derivative to zero of the total entropy with respect to E_A, subject to the condition that the total energy is constant.

Under the constraint of constant total energy E, we have $dE_B/dE_A = d(E - E_A)/dE_A = -1$, so

$$\frac{\partial S}{\partial E_A} = \frac{\partial S_A}{\partial E_A} + \frac{\partial S_B}{\partial E_A} = \frac{\partial S_A}{\partial E_A} + \frac{\partial S_B}{\partial E_B}\frac{dE_B}{dE_A} = \frac{\partial S_A}{\partial E_A} - \frac{\partial S_B}{\partial E_B} = 0. \tag{23.18}$$

The partial derivatives indicate that parameters besides the energy are held constant while taking the derivative of entropy.

Thus,

$$\frac{\partial S_A}{\partial E_A} = \frac{\partial S_B}{\partial E_B} \quad \text{(equilibrium condition)} \tag{23.19}$$

at the point of maximum entropy.

Once the equilibrium values of E_A and E_B are found, we can calculate the total entropy $S = S_A + S_B$ of two thermally isolated bricks.

We now assert that this entropy doesn't change when two bricks in equilibrium are brought into thermal contact.

Why is this so?

The derivative of the entropy of a system with respect to energy turns out to be one over the temperature of the system.

Thus, the temperatures of the bricks can be found from

$$\frac{1}{T} \equiv \frac{\partial S}{\partial E} \quad \text{(definition of temperature)}. \tag{23.20}$$

The condition for equilibrium (23.19) therefore reduces to $1/T_A = 1/T_B$, or $T_A = T_B$.

This is consistent with observations of the behavior of real systems.

Thus, at the equilibrium point the temperatures of the two bricks are the same and bringing them together causes no heat flow to occur.

The process of bringing two bricks at the same temperature into thermal contact is thus completely reversible, since separating them leaves each with the same amount of energy it started with.

The temperature of a brick is easily calculated using equation (23.20):

$$T = \frac{E}{k_B N} \quad \text{(temperature of } N \text{ harmonic oscillators)}. \tag{23.21}$$

We see that the temperature of a brick is just the average energy per harmonic oscillator in the brick divided by Boltzmann's constant.

23.4 Thermodynamic Temperature

Equation (23.20) provides us with a physical definition of temperature that is independent of specific material properties such as the thermal expansion coefficient of some particular metal.

Though different materials have different dependences of entropy on internal energy, the *derivative of entropy with respect to energy* will be the same for any two materials in thermal equilibrium with each other

Note that the unit of temperature is the Kelvin degree according to this theory.

If we had left off Boltzmann's constant in the definition of entropy, the dimensions of temperature would be that of energy.

Boltzmann's constant is thus simply a scaling factor that changes temperature to energy just as multiplication by the speed of light converts time to distance.

23.5 Specific Heat

How can we compute the specific heat of a collection of harmonic oscillators?

Starting from the temperature of a brick, as given by equation (23.21), we solve for the brick's internal energy:

$$E = Nk_BT \quad \text{(internal energy of } N \text{ oscillators)}. \tag{23.22}$$

Recall that the specific heat is the heat required per unit mass to increase the temperature of the brick by one degree.

For a solid body, essentially all the heat added to the body goes into increasing its internal energy.

Thus, if the mass of the brick is $M = Nm$ where m is the mass per oscillator, then the predicted specific heat of the brick is

$$C \equiv \frac{1}{M}\frac{dQ}{dT} \approx \frac{1}{M}\frac{dE}{dT} = \frac{k_B}{m} \quad \text{(specific heat of harmonic oscillators)}. \tag{23.23}$$

This formula is in reasonable agreement with measurements when the temperature is high enough so that all the harmonic oscillators are in excited states, i. e., with $r > 1$.

We equate $dQ = dE$ using the first law of thermodynamics, since no work is being done by the brick.)

23.6 Entropy and Heat Conduction

Though entropy is formally not defined in a system that is not in thermodynamic equilibrium, one can imagine situations in which elements of a system interact only weakly with other elements.

Each element is therefore very close to internal equilibrium, so that the entropy of each element can be defined.

However, the elements are not in equilibrium with each other.

Figure 23.5 shows an example of such a situation.

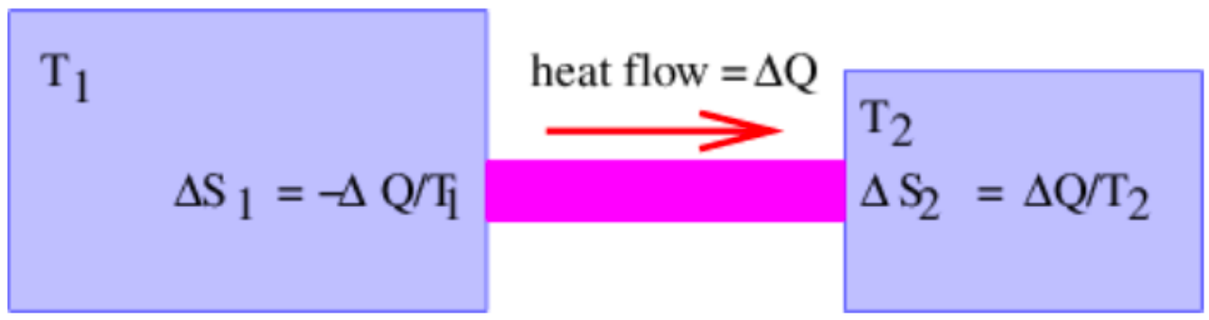


Figure 23.5: The two regions at temperatures T_1 and $T_2 < T_1$ are connected by a thin bar that conducts heat slowly from the first to the second region. For heat ΔQ transferred, the entropy of region 1 decreases according to $\Delta S_1 = -\Delta Q/T_1$, while the entropy of region 2 increases by $\Delta S_2 = \Delta Q/T_2$.

Since $1/T = \partial S/\partial E$, one can write

$$\Delta S_1 = -\Delta Q/T_1, \tag{23.24}$$

since heat flowing out of region 1 results in a decrease in internal energy $\Delta E_1 = -\Delta Q$.

Likewise, we find that

$$\Delta S_2 = \Delta Q / T_2, \tag{23.25}$$

since the internal energy of region 2 increases by $\Delta E_2 = \Delta Q$.

The total change of entropy of the system is therefore

$$\Delta S = \Delta S_1 + \Delta S_2 = \Delta Q \left(\frac{1}{T_2} - \frac{1}{T_1} \right). \tag{23.26}$$

From our experience, we know that heat will only flow from region 1 to region 2 if $T_1 > T_2$.

However, equation (23.26) shows that the net entropy change is positive when this is true.

Conversely, if $T_1 < T_2$, then the net entropy change would be negative and heat would be flowing spontaneously from lower to higher temperatures.

Thus, the statement that heat cannot spontaneously flow from lower to higher temperatures is equivalent to the statement that the entropy of an isolated system must not decrease.

An alternative statement of the second law of thermodynamics is therefore *heat cannot spontaneously flow from lower to higher temperatures*.

If entropy increases in some process, we call it *irreversible*.

Spontaneous heat flow is always irreversible.

However, in the limit in which the temperature difference is very small, the entropy increase due to heat flow is also small.

Of course, the rate of flow of heat is also quite slow in this case.

Nevertheless, this situation forms a useful idealization.

In the idealized limit of very small, but nonzero temperature difference, the flow of heat is said to be *reversible* because the generation of entropy is negligible.

Chapter 24

The Ideal Gas and Heat Engines

All heat engines have the common property of turning internal energy into useful macroscopic energy. They extract internal energy from a high temperature reservoir, convert part of this energy to useful work, and transfer the rest to a low temperature reservoir.

The second law of thermodynamics imposes a firm limit on the fraction of the initial internal energy that can be converted to macroscopic energy.

Almost all heat engines work by means of expansions and contractions of a gas.

A simple theoretical model called the *ideal gas* model quite accurately predicts the behavior of the gases in most heat engines of this type

Our first task is to build the ideal gas model using the techniques learned in the previous chapter.

We then use this model to understand the operation of heat engines.

We are particularly interested in determining the maximum theoretical efficiency at which these devices can convert heat to useful work.

24.1 Ideal Gas

An ideal gas is an assembly of atoms or molecules that interact with each other only via occasional collisions.

The distance between molecules is much greater than the range of inter-molecular forces, so gas molecules behave as free particles most of the time.

We assume here that the density of molecules is also low enough to make the probability of finding more than one molecule in a given quantum mechanical state very small.

For this reason it doesn't matter whether the molecules are bosons or fermions for our calculations.

J. Willard Gibbs tried computing the entropy of an ideal gas using his version of statistical mechanics, which was based on classical mechanics.

The result was wrong in a very fundamental way — the calculated amount of entropy was not proportional to the amount of gas.

In fact, the amount of entropy of an ideal gas at fixed temperature and pressure is calculated to have a non-linear dependence on the number of gas molecules.

In particular, doubling the amount of gas more than doubles the entropy according to the Gibbs formula.

The significance of this error is illustrated in figure 24.1.

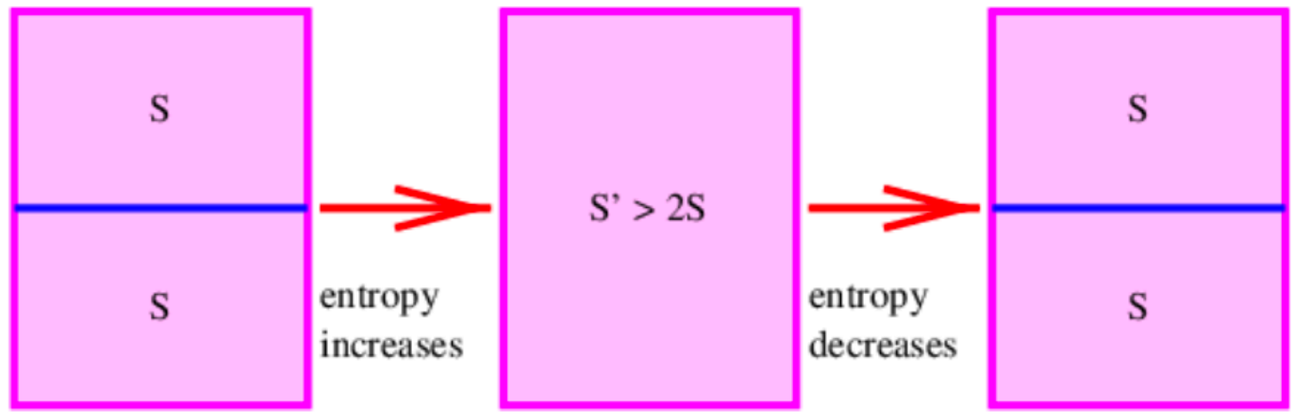


Figure 24.1: Consequence of the incorrect classical calculation of entropy of an ideal gas by Gibbs.Two parts of a container separated by a divider each contain the sam type of gas at the same temperature and pressure. The tota entropy is $2S$ where S is the classically calculated entropy of each half. If the divider is removed, the classical calculation yields an entropy for the entire body of gas S' $> 2S$. Reinserting the divider returns the container to the initial state in which the total entropy is $2S$.

Imagine a container of gas of a certain type, temperature, and pressure that is divided into two equal parts by a sheet of material.

The total entropy of this state is $2S$, where S is the entropy calculated separately for each half of the body of gas.

This follows because the two halves are completely separate systems.

If the divider is now removed, a calculation of the entropy of the full body of gas yields $S' > 2S$ according to the Gibbs formula, since the calculated entropy doesn't scale with the amount of gas.

Furthermore, replacing the divider restores the system to the initial state in which the total entropy is $2S$.

Thus, simply inserting or removing the divider, an operation that transfers no heat and does no work on the system, is able to increase or decrease the entropy of the gas at will according to Gibbs.

This is at variance with the second law of thermodynamics and is known not to occur.

Its prediction by the formula of Gibbs is called the *Gibbs paradox*.

Gibbs was well aware of the serious nature of this problem but was unable to come up with a satisfying solution.

The resolution of the paradox is simply that the Gibbs formula for the entropy of an ideal gas is incorrect.

The correct formula is only obtained when the quantum mechanical version of statistical mechanics is used.

The failure of Gibbs to obtain the proper entropy was an early indication that classical mechanics had problems on the atomic scale.

We will now calculate the entropy of a body of ideal gas using quantum statistical mechanics.

In order to reduce the difficulty of the calculation, we will take a shortcut and *assume* that the amount of entropy is proportional to the amount of gas.

However, the more rigorous calculation confirms that this actually is true.

24.1.1 Particle in a Three-Dimensional Box

The quantum mechanical calculation of the states of a particle in a three-dimensional box forms the basis for our treatment of an ideal gas.

Recall that a non-relativistic particle of mass M in a one-dimensional box of width a can only support wavenumbers $k_l = \pm\pi l/a$ where $l = 1, 2, 3,\ldots$ is the quantum number for the particle.

Thus, the possible momenta are $p_l = \pm\hbar\pi l/a$ and the possible energies are

$$E_l = p_l^2/(2M) = \hbar^2\pi^2 l^2/(2Ma^2) \quad \text{(one-dimensional box)}. \tag{24.1}$$

If the box has three dimensions, is cubical with edges of length a, and has one corner at $(x,y,z) = (0, 0,0)$, the quantum mechanical wave function for a single particle that satisfies $\psi = 0$ on all the walls of the box is a three-dimensional standing wave,

$$\psi(x, y, z) = \sin(lx\pi/a)\sin(my\pi/a)\sin(nz\pi/a), \tag{24.2}$$

where the quantum numbers l,m,n are positive integers.

You can verify this by examining ψ for $x = 0$, $x = a$, etc.

Equation (24.2) is a solution in which the x, y, and z wavenumbers are respectively $k_x=\pm l\pi/a$, $k_y=\pm m\pi/a$, and $k_z=\pm n\pi/a$.

The corresponding components of the kinetic momentum are therefore p_x etc.

The possible energy values of the particle are

$$
\begin{aligned}
E_{lmn} &= \frac{p^2}{2M} = \frac{p_x^2 + p_y^2 + p_z^2}{2M} = \frac{\hbar^2\pi^2(l^2 + m^2 + n^2)}{2Ma^2} \\
&= \frac{\hbar^2\pi^2 L^2}{2MV^{2/3}} \equiv E_0 L^2 \quad \text{(three-dimensional box)}.
\end{aligned}
\tag{24.3}
$$

In the last line of the above equation we have eliminated the linear dimension a of the box in favor of its volume $V = a^3$ and have adopted the shorthand notation $L^2 = l^2 + m^2 + n^2$ and $E_0 = \hbar^2\pi^2/(2Ma^2) = \hbar^2\pi^2/(2MV^{2/3})$.

The quantity E_0 is the ground state energy for a particle in a one-dimensional box of size a.

Figure 24.2 shows the energy levels of a particle in a one-dimensional and a three-dimensional box.

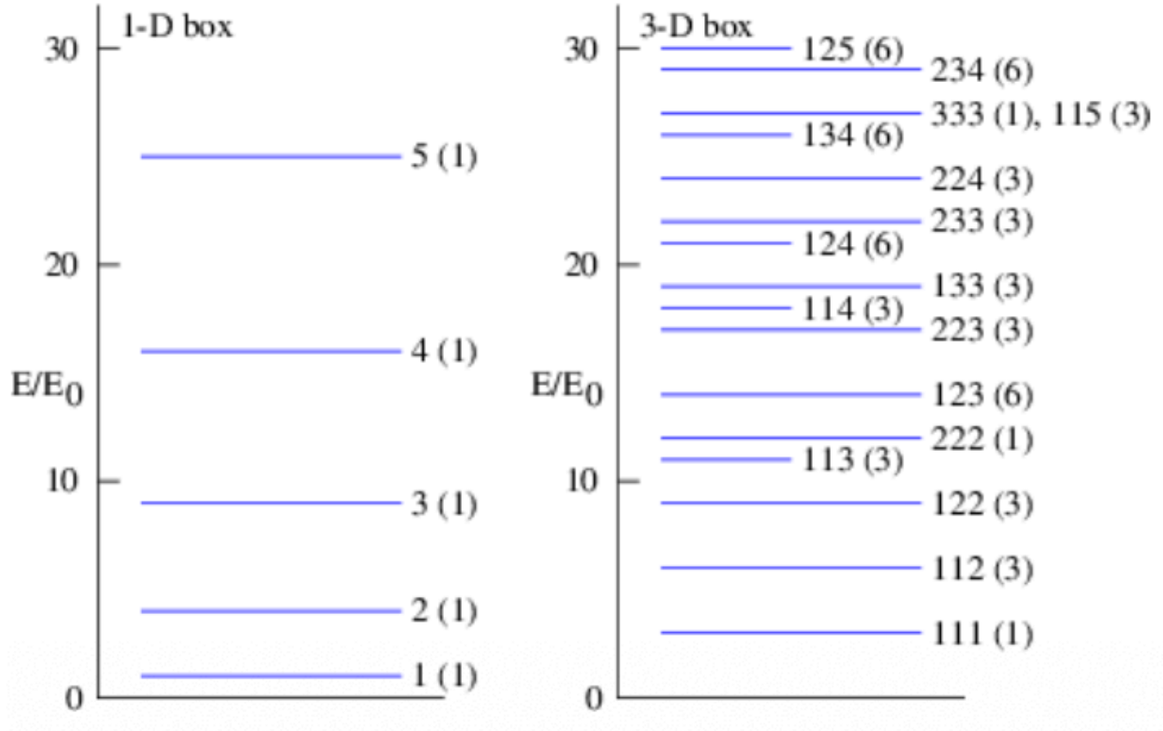


Figure 24.2: Energy levels for a non-relativistic particle in a one-dimensional and a three-dimensional box, each of side length a. The value E_0 is the ground state energy of the one-dimensional particle in a box of length a. The numbers to the right of the levels respectively give the values of l for the 1-D oscillator and the values of l, m, and n for the 3-D oscillator. The numbers in parentheses give the degeneracy of each energy level (see text).

Different values of l, m, and n can result in the same energy in the three-dimensional case.

For instance, $(l,m,n) = (1, 1, 2), (1, 2, 1), (2, 1, 1)$ all yield $L^2 = 6$ and hence energy $6E_0$.

This energy level is thus said to have a *degeneracy* of 3.

Similarly, the states (1, 2, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1), (2, 1, 3), (1, 3, 2) all have the same value of L^2, so this level has a degeneracy of 6.

However, the state (1, 1, 1) is unique and thus has a degeneracy of 1.

From this we see that the degeneracy of an energy level is the number of different physically distinguishable states that have the same energy.

Counting the effects of degeneracy, the particle in a three-dimensional box has 60 distinct states for $E \leq 30E_0$, while the one-dimensional box has 5.

As the limiting value of E/E_0 increases, this ratio becomes even larger.

24.1.2 Counting States

In order to compute the entropy of a system, we need to count the number of states available to the system in a particular band of energies.

Figure 24.3 shows how to count the states with energy less than some limiting value for a particle in a two-dimensional box.

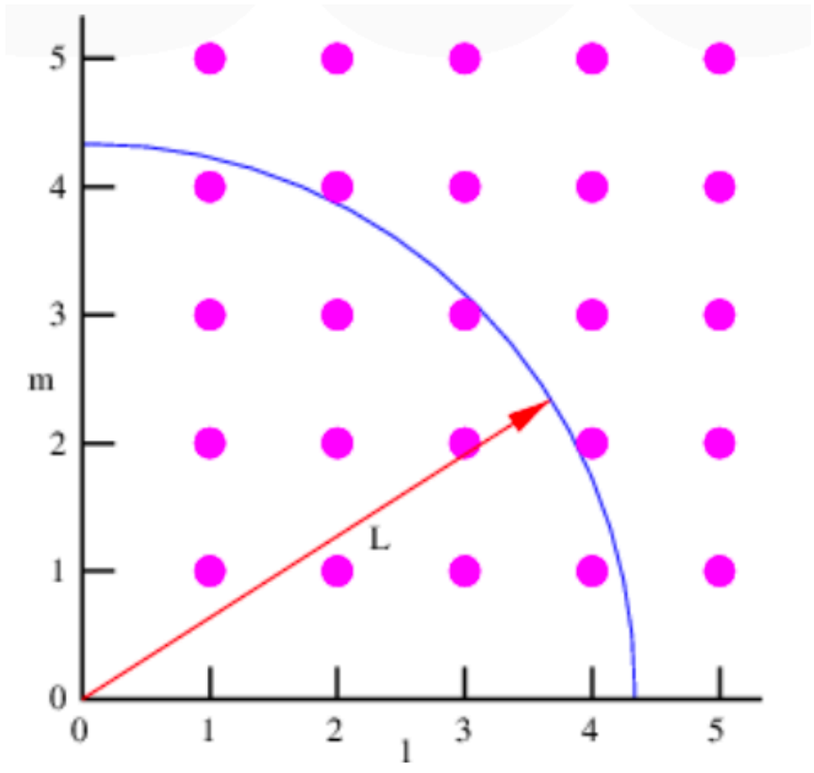


Figure 24.3: The states of a particle in a two-dimensional box. The dots indicate particular states associated with allowed values of the x and y direction quantum numbers, l and m. The pie-shaped segment bounded by the arc of radius L and the l and m axes encompasses all of the states with $l^2 + m^2 \leq L^2$.

The pie-shaped segment bounded by the arc of radius L and the l and m axes has an area equal to one fourth the area of a circle of radius L, or $\pi L^2/4$.

The dots represent allowed values of the l and m quantum numbers.

One dot, and hence one state, exists per unit area in this graph, so the above expression tells us how many states $\mathcal{N}$ exist with $l^2 + m^2 \leq L^2$.

In two dimensions the particle energy is $E = (l^2 + m^2)E_0$.

Thus, the number of states with energy less than or equal to some maximum energy E is

$$\mathcal{N} = \frac{\pi L^2}{4} = \frac{\pi}{4}\left(\frac{E}{E_0}\right) \quad \text{(two-dimensional box)}. \tag{24.4}$$

Similar arguments can be made to calculate the number of states of a particle in a three-dimensional box.

The equivalent of figure 24.3 would be a plot with three axes, l, m, and n representing the x, y, and z quantum numbers.

The volume of a sphere with radius L is then $4\pi L^3/3$ and the region of the sphere with $l,m,n > 0$, i. e., an eighth of the sphere, contains real physical states.

The result is that

$$\mathcal{N} = \frac{4\pi L^3}{3 \cdot 8} = \frac{\pi}{6}\left(\frac{E}{E_0}\right)^{3/2} \quad \text{(three-dimensional box)} \tag{24.5}$$

states exist with energy less than E.

24.1.3 Multiple Particles

An ideal gas of only one molecule isn't very interesting.

Calculating the number of states available to many particles in a box is a bit complex.

However, by analogy with the case of multiple harmonic oscillators, we guess that the number of states of an N-particle gas is the number of states available to a single particle to the Nth power multiplied by some as yet unknown function of N, $F(N)$:

$$\mathcal{N} = F(N)\left(\frac{E}{E_0}\right)^{3N/2} \qquad (N \text{ particles in 3-D box}). \tag{24.6}$$

(Note that the $(\pi/6)^N$ from equation (24.5) has been absorbed into $F(N)$.)

Substituting $E_0 = \hbar^2\pi^2/(2MV^{2/3})$ results in

$$\mathcal{N} = F(N)\left(\frac{2MEV^{2/3}}{\hbar^2\pi^2}\right)^{3N/2}. \tag{24.7}$$

Now, $\pi^2\hbar^2/(2M)$ has the units of energy multiplied by volume to the two-thirds power, so we write this combination of constants in terms of constant reference values of E and V :

$$\pi^2\hbar^2/(2M) = E_{ref}V_{ref}^{2/3}. \tag{24.8}$$

TheGiven the above assumption, we can rewrite the number of states with energy less than E as

$$\mathcal{N} = F(N)\left(\frac{E}{E_{ref}}\right)^{3N/2}\left(\frac{V}{V_{ref}}\right)^{N}. \tag{24.9}$$

We now argue that the combination $F(N)$ *must* take the form $KN^{-5N/2}$ where K is a dimensionless constant independent of N.

Substituting this assumption into equation (24.9) results in

$$\mathcal{N} = K\left(\frac{E}{NE_{ref}}\right)^{3N/2}\left(\frac{V}{NV_{ref}}\right)^{N}. \tag{24.10}$$

It turns out that we will not need the actual values of any of the three constants K, E_{ref}, or V_{ref}.

The effect of the above hypothesis is that the energy and volume occur only in the combinations $E/(NE_{ref})$ and $V/(NV_{ref})$.

First of all, these combinations are dimensionless, which is important because they will become the arguments of logarithms.

Second, because of the N in the denominator in both cases, they are in the form of energy per particle and volume per particle.

If the energy per particle and the volume per particle stay fixed, then the only dependence of $\mathcal{N}$ on N is via the exponents $3N/2$ and N in the above equation.

Why is this important? Read on!

24.1.4 Entropy and Temperature

Recall now that we need to compute the number of states in some small energy interval ΔE in order to get the entropy.

Proceeding as for the case of a collection of harmonic oscillators, we find that

$$\Delta\mathcal{N} = \frac{\partial\mathcal{N}}{\partial E}\Delta E = \frac{3KN\Delta E}{2E}\left(\frac{E}{NE_{ref}}\right)^{3N/2}\left(\frac{V}{NV_{ref}}\right)^{N}. \tag{24.11}$$

The entropy is therefore

$$S = k_B \ln(\Delta\mathcal{N}) = Nk_B\left[\frac{3}{2}\ln\left(\frac{E}{NE_{ref}}\right) + \ln\left(\frac{V}{NV_{ref}}\right)\right] \quad \text{(ideal gas)}, \tag{24.12}$$

where we have dropped the term $k_B \ln[3KN\Delta E/(2E)]$.

Since this term is not multiplied by the number of particles N, it is unimportant for systems made up of lots of particles.

Notice that this equation has a very important property, namely, that the entropy is proportional to the number of particles for fixed E/N and V/N.

It thus satisfies the criterion that Gibbs was unable to satisfy in his computation of the entropy of an ideal gas.

However, we cannot claim that our calculation is superior to his, because we *cheated*!

The reason we assumed that $F(N) = KN^{-5N/2}$ is precisely so that we would obtain this result.

The temperature is the inverse of the E-derivative of entropy:

$$\frac{1}{T} = \frac{\partial S}{\partial E} = \frac{3Nk_B}{2E} \quad \Longrightarrow \quad T = \frac{2E}{3Nk_B} \quad \text{or} \quad E = \frac{3Nk_BT}{2}. \tag{24.13}$$

24.1.5 Slow and Fast Expansions

How does the entropy of a particle in a box change if the volume of the box is changed?

The answer to this question depends on how rapidly the volume change takes place.

If an expansion or compression takes place slowly enough and no heat is added or removed, the quantum numbers of the particle don't change.

This fact may be demonstrated by the tuning of a guitar.

A guitar string is tuned in frequency by adjusting the tension on the string with the tuning peg.

If the first harmonic mode (corresponding to quantum number $n = 2$ for particle in a one-dimensional box) is excited on a guitar string as illustrated in figure 24.4, changing the tension changes the frequency of the vibration but it does not change the mode of vibration of the string — for instance, if the first harmonic is initially excited, it remains the primary mode of oscillation.

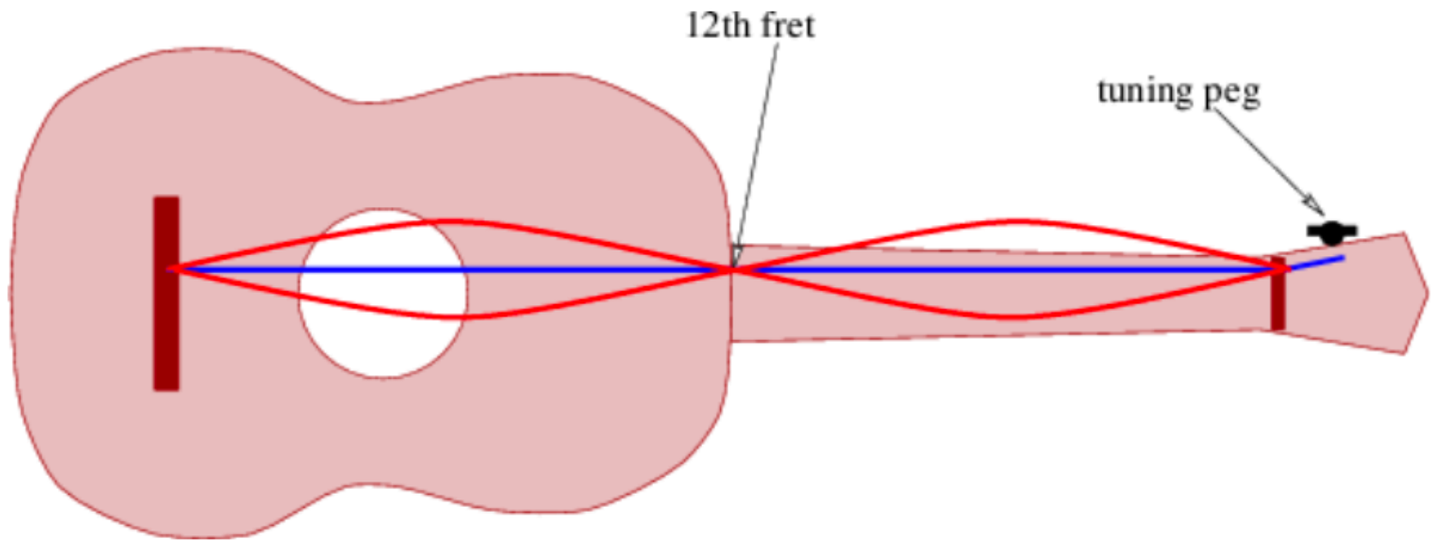


Figure 24.4: Harmonics on a guitar. Plucking a string while a finger rests lightly on the string at the 12th fret results in excitation of the first harmonic on the guitar string. Only one string is shown for clarity

Slowly changing the volume of a gas consisting of many particles, each with its own set of quantum numbers, results in the same behavior — changing the dimensions of the box results in no switching of quantum numbers beyond that which would normally take place as a result of particle collisions.

As a consequence, the number of states available to the system, $\Delta\mathcal{N}$, and hence the entropy, doesn't change either.

A process that changes the macroscopic condition of a system but that doesn't change the entropy is called *isentropic* or *reversible adiabatic*.

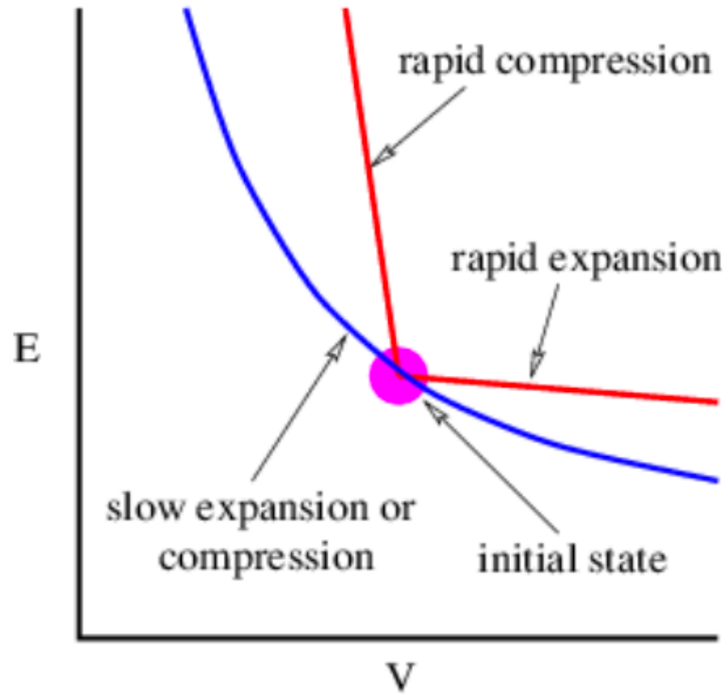
The word "isentropic" means at constant entropy, while "adiabatic" means that no heat flows in or out of the system.

If the entropy doesn't change as a result of a change in volume, then $E^{3/2}V = const$ according to equation (24.12).

Thus, the energy of the gas increases when the volume is decreased and vice versa.

This behavior is illustrated in figure 24.5.

Figure 24.5: The curved line indicates the reversible adiabatic curve $E \propto V^{-2/3}$ for an ideal gas in a box. The two straight line segments indicate what happens in a rapid expansion or compression.



The change in energy in both cases is a consequence of work done by the gas on the walls of the container as it changes volume — positive in expansion, meaning that the gas loses energy; and negative in compression, meaning that the gas gains energy.

This type of energy transfer is the means by which internal energy is converted to useful work.

A rapid expansion of the box has a completely different effect.

If the expansion is so rapid that the quantum mechanical waves trapped in the box undergo negligible evolution during the expansion, then the *internal energy* of the particles in the box does not change.

As a consequence, the particle quantum numbers *must* change to compensate for the change in volume.

Equation (24.12) tells us that if the volume increases and the internal energy doesn't change, the entropy must increase.

A rapid compression has the opposite effect — it does extra work on the material in the box, thus adding internal energy to the gas at a rate in excess of the reversible adiabatic rate.

The entropy increases in this type of process as well.

Both of these effects are illustrated in figure 24.5.

24.1.6 Work, Pressure, and Gas Law

The *pressure p* of a gas is the normal force per unit area exerted by the gas on the walls of the chamber containing the gas.

If a chamber wall is movable, the pressure force can do positive or negative work on the wall if it moves out or in.

This is the mechanism by which internal energy is converted to useful work.

We can determine the pressure of a gas from the entropy formula.

Consider the behavior of a gas contained in a cylinder with a movable piston as shown in figure 24.6.

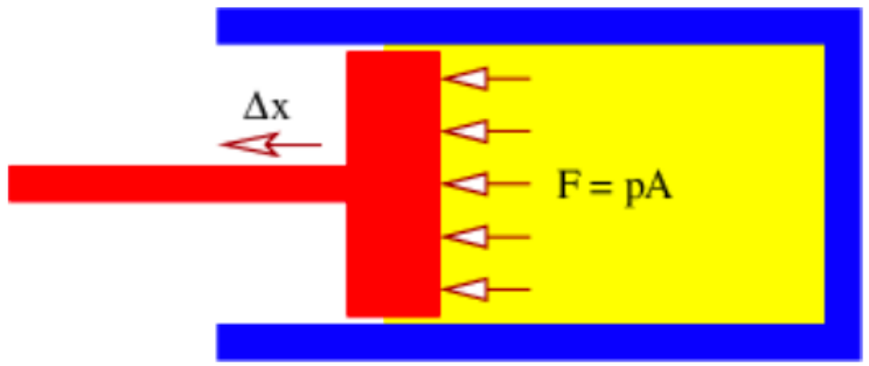


Figure 24.6: Gas in a cylinder with a movable piston. The force F exerted by the gas on the piston is the area A of the face of the piston multiplied by the pressure p.

The net force F exerted by gas molecules bouncing off of the piston results in work $\Delta W = F\Delta x$ being done by the gas if the piston moves (slowly) out a distance Δx.

The pressure is related to F and the area A of the piston by $p = F/A$.

Furthermore, the change in volume of the cylinder is $\Delta V = A\Delta x$.

If the gas does work ΔW on the piston, its internal energy changes by

$$\Delta E = -\Delta W = -F\Delta x = -\frac{F}{A}A\Delta x = -p\Delta V, \tag{24.14}$$

assuming that $\Delta Q = 0$, i. e., no heat is added or removed during the change in volume.

Solving this for p results in

$$p \equiv -\frac{\partial E}{\partial V}. \tag{24.15}$$

In the previous section we showed that as long as the change in volume is slow and $\Delta Q = 0$, the entropy does not change.

Thus, in the evaluation of $\partial E/\partial V$, the entropy is held constant.

We can determine the pressure for an ideal gas by solving equation (24.12) for E and taking the V derivative.

As we showed in the previous section, $E^{3/2}V$ is constant for constant entropy processes, which means that

$$E = BV^{-2/3} \quad \text{(constant entropy expansion)}, \tag{24.16}$$

where B is constant.

The pressure is then computed to be

$$p = -\frac{\partial E}{\partial V} = \frac{2}{3}\frac{B}{V^{5/3}} = \frac{2E}{3V} \quad \text{(ideal gas)}, \tag{24.17}$$

where B is eliminated in the last step using equation (24.16).

Employing equation (24.13) to eliminate the energy in favor of the temperature, this can be written

$$pV = Nk_BT \quad \text{(ideal gas law)}, \tag{24.18}$$

which relates the pressure, volume, temperature, and particle number of an ideal gas.

This equation is called the *ideal gas law* and jointly represents the observed relationships between pressure and volume at constant temperature (Boyle's law) and pressure and temperature at constant volume (law of Charles and Gay-Lussac).

The fact that we can derive it from statistical mechanics is evidence in favor of our quantum mechanical model of a gas.

The formulas for the entropy of an ideal gas (24.12), its temperature (24.13), and the ideal gas law (24.18) summarize our knowledge about ideal gases.

Actually, the entropy and temperature laws only apply to a particular type of ideal gas in which the molecules consist of single atoms.

This is called a *monatomic* ideal gas, examples of which are helium, argon, and neon.

The molecules of most gases consist of two or more atoms.

These molecules have rotational degrees of freedom that can store energy.

The calculation of the entropy of such gases needs to take these factors into account.

The most common case is one in which the molecules are *diatomic*, i. e., they consist of two atoms each.

In this case simply replacing factors of $3/2$ by $5/2$ in equations (24.12), (24.13), (24.16), and (24.17) results in equations that apply to diatomic gases at ordinary temperatures.

24.1.7 Specific Heat of an Ideal Gas

As previously noted, the specific heat of any substance is the amount of heating required per unit mass to raise the temperature of the substance by one degree.

For a gas one must clarify whether the volume or the pressure is held constant as the temperature increases — the specific heat differs between these two cases because in the latter situation the added energy from the heating is split between the production of internal energy and the production of work as the gas expands.

At constant volume all heating goes into increasing the internal energy, so $\Delta Q = \Delta E$ from the first law of thermodynamics.

From equation (24.13) we find that $\Delta E = (3/2)NkB\Delta T$. If the molecules making up the gas have mass M, then the mass of the gas is NM.

Thus, the specific heat at constant volume of an ideal gas is

$$C_V = \frac{1}{NM}\frac{3Nk_B}{2} = \frac{3k_B}{2M} \quad \text{(specific heat at const vol)}. \tag{24.19}$$

As noted above, when heat is added to a gas in such a way that the pressure is kept constant as a result of allowing the gas to expand, the added heat ΔQ is split between the increase in internal energy ΔE and the work done by the gas in the expansion $\Delta W = p\Delta V$ such that $\Delta Q = \Delta E + p\Delta V$.

In a constant pressure process the ideal gas law (24.18) predicts that $p\Delta V = Nk_B\Delta T$.

Using this and the previous equation for ΔE results in the specific heat of an ideal gas at constant pressure:

$$C_P = \frac{1}{NM}\left(\frac{3Nk_B}{2} + Nk_B\right) = \frac{5k_B}{2M} \quad \text{(specific heat at const pres)}. \tag{24.20}$$

24.2 Heat Engines

Heat engines typically operate by heating and cooling a volume of gas and by compressing or expanding the gas.

If these operations are done in a particular order, internal energy can be converted to useful work.

We therefore seek to understand how an ideal gas reacts to the addition and subtraction of internal energy and to the change in the volume of the gas.

The equation for the entropy of an ideal gas and the ideal gas law contain the information we need.

The entropy of an ideal gas is a function of its internal energy E and its volume V.

We assume that the number of molecules in the gas remains fixed.

Thus, a small change ΔS in the entropy can be related to small changes in the energy and volume as follows:

$$\Delta S = \frac{\partial S}{\partial E}\Delta E + \frac{\partial S}{\partial V}\Delta V. \tag{24.21}$$

We know that $\partial S/\partial E = 1/T$.

Using equation (24.12) we can similarly calculate $\partial S/\partial V = Nk_B/V = p/T$, where the ideal gas law is used in the last step to eliminate Nk_B in favor of p.

Substituting these into the above equation, multiplying by T, and solving for $p\Delta V$ results in

$$p\Delta V = \Delta W = T\Delta S - \Delta E \quad \text{(work by ideal gas)} \tag{24.22}$$

where we have recognized $p\Delta V = \Delta W$ to be the work done by the gas on the piston.

We are now in a position to investigate the conversion of internal energy to useful work.

If the gas is allowed to push the piston out in a reversible adiabatic manner, then $\Delta S = 0$ and energy is converted with 100% efficiency from internal form to work.

This work could in principle be used to run an electric generator, stretch springs, power an automobile, etc.

Unfortunately, a piston in a cylinder that can only extract energy during single expansion wouldn't be very useful — it would be like an automobile engine that only worked for half a turn of the crankshaft and then had to be replaced!

If the piston is simply pushed back into the cylinder, then the macroscopic energy gained from the initial expansion would be converted back into internal energy of the gas, resulting in zero net creation of useful work.

The trick to obtaining non-zero useful work from the expansion and contraction of a gas is to add heat to the gas before the expansion and extract heat from it before the recompression.

This makes the gas cooler in the compression than in the expansion.

The pressure is therefore less in the compression and the work needed to compress the gas is less than that produced in the expansion.

Figure 24.7 shows a particular way of executing a complete cycle of expansion and compression of the gas that results in a net conversion of internal energy to useful work.

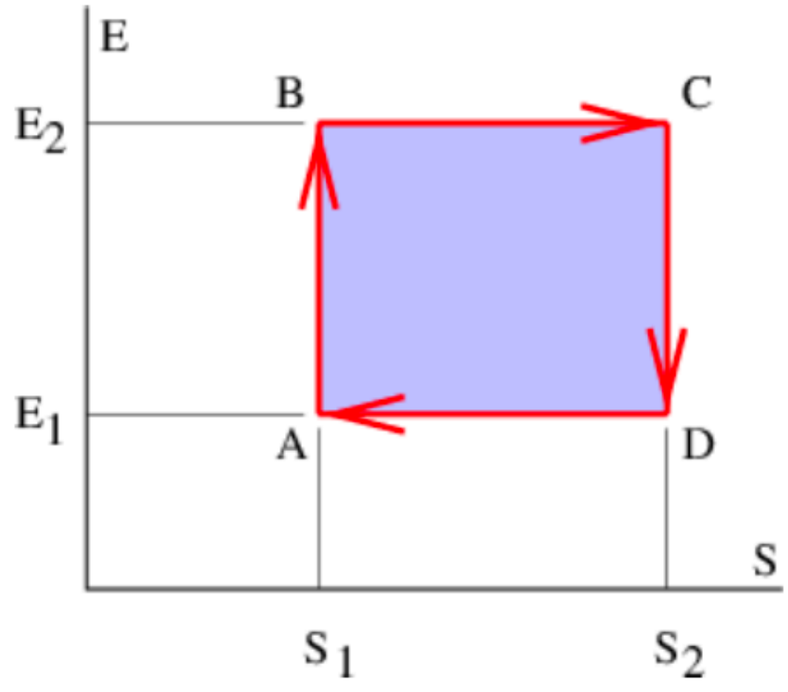


Figure 24.7: Plot of Carnot cycle for an ideal gas in a cylinder. Entropy-energy coordinates are used

Assuming that the gas has initial entropy and internal energy S_1 and E_1 at point A in figure 24.7, the gas is first compressed in reversible adiabatic fashion to point B.

The entropy doesn't change in this compression but the internal energy increases from E_1 to E_2.

The work done by the gas is negative and equals $W_{AB} = E_1 - E_2$.

The gas then is allowed to slowly expand (so that the expansion is reversible), moving from point B to point C in figure 24.7 in such a way that its internal energy doesn't change.

From equation (24.22) we see that $W_{BC} = T_2(S_2 - S_1)$ for this segment of the expansion.

However, heat must be added to the gas equal in amount to the work done in order to keep the internal energy fixed: $Q_2 = T_2(S_2 - S_1)$.
From point C to point D the gas expands further but in this segment the expansion is reversible adiabatic so that the entropy change is again zero.

Thus, $W_{CD} = E_2 - E_1$.

Finally, the gas is slowly compressed from point D to point A in a constant internal energy process.

Keeping the internal energy fixed means that the (negative) work done by the gas in this segment is $W_{DA} = T_1(S_1 - S_2)$.

Furthermore, heat equal to the work done *on* the gas *by* the piston must be removed from the gas in order to keep the internal energy constant: $Q_1 = -W_{DA} = T_1(S_2 - S_1)$.

The net work done by the gas over the full cycle is obtained by adding up the contributions from each segment:

$$\begin{aligned} W &= W_{AB} + W_{BC} + W_{CD} + W_{DA} \\ &= (E_1 - E_2) + T_2(S_2 - S_1) + (E_2 - E_1) + T_1(S_1 - S_2) \\ &= (T_2 - T_1)(S_2 - S_1) \quad \text{(Carnot cycle)}. \end{aligned} \tag{24.23}$$

The energy source for this work is internal energy at temperature T_2.

As demanded by energy conservation, $W = Q_2 - Q_1$.

The fraction of the internal energy Q_2 that is converted to useful work in this cycle is

$$\epsilon = \frac{W}{Q_2} = \frac{(T_2 - T_1)(S_2 - S_1)}{T_2(S_2 - S_1)} = 1 - \frac{T_1}{T_2} \quad \text{(thermodynamic efficiency)}. \tag{24.24}$$

This quantity ϵ is called the *thermodynamic efficiency* of the heat engine.

Notice that it depends only on the ratio of the lower and upper temperatures, expressed in absolute or Kelvin form.

The smaller this ratio, the larger the thermodynamic efficiency.

Heat engines normally work via repeated cycling around some loop such as described above.

The particular cycle we have discussed is called the *Carnot cycle* after the 19th century French scientist Sadi Carnot.

Heat is accepted from a high temperature heat source, created, for example, by burning coal in a power plant.

Excess heat is disposed of in the atmosphere or in some source of running water such as a river.

Notice that the ability to get rid of excess heat at low temperature is as important to a heat engine as the supply of heat at a high temperature.

Many cycles for converting heat to work are possible — these are represented by different closed trajectories in the S-E plane.

However, the Carnot cycle is special for two reasons: First, all heat absorbed by the system is absorbed at a single temperature T_2, and all heat rejected from the system is rejected at a single temperature T_1.

This allows the expression of the efficiency simply in terms of the two temperatures.

Second, the Carnot cycle is *reversible*, which means that no net entropy is generated.

A Carnot engine running backwards acts as a refrigerator.

Heat Q_1 is extracted at temperature T_1 from the box being cooled with the aid of externally supplied work W.

An amount of heat $Q_2 = W + Q_1$ is then transferred to the environment at temperature $T_2 > T_1$.

Equation (24.24) gives the ratio of W to Q_2 in this case, as well as when the heat engine is run in the forward direction.

This may be verified by tracing the cycle in figure 24.7 in reverse.

In analyzing heat engines and refrigerators it is generally easier to go back to basic principles than it is to use equations such as (24.23) and (24.24).

In particular, for a Carnot engine in which heat Q_2 is being extracted from the high temperature reservoir (T_2) and heat Q_1 is being added to the low temperature reservoir (T_1), conservation of energy says that the useful work extracted is $W = Q_2 - Q_1$, and that the total combined entropy change in the warm and cold reservoirs is $\Delta S = -Q_2/T_2 + Q_1/T_1 = 0$.

Note that the reservoir providing energy has a minus sign, while the reservoir accepting energy has a plus sign.

Given these two relationships, any two of Q_1, Q_2, and W can be determined if the third is known.

For a refrigerator, the higher temperature reservoir accepts energy while the lower temperature reservoir (generally the interior of the refrigerator) and the work term provide energy.

This changes the signs of all three energy flows.

If the machine is not a perfectly efficient Carnot engine, then $\Delta S > 0$ whether the machine is a heat engine or a refrigerator, and one deals with inequalities rather than equalities.

24.3 Perpetual Motion Machines

Perpetual motion machines are devices that are purported to create useful work for "nothing" by violating some physical principle.

Generally they are divided into two types, perpetual motion machines of the first and second kinds.

Perpetual motion machines of the first kind violate the conservation of energy, while perpetual motion machines of the second kind violate the second law of thermodynamics.

It is the latter type that we address here.

We commonly hear talk of an "energy crisis".

However, it is clear to all physicists that such a crisis, if it exists, is actually an "entropy crisis".

Energy beyond the most extravagant projected needs of mankind exists in the form of internal energy of the earth.

Furthermore, one cannot possibly "waste" energy, because energy can neither be created nor destroyed.

The real problem is that internal energy can only be tapped if two reservoirs of internal energy exist, one at high temperature and one at low temperature.

Heat engines depend on this temperature difference to operate and if all internal energy exists at the same temperature, no conversion of internal energy to useful work is possible, at least using the Carnot cycle.

One naturally inquires as to whether some cycle exists that is more efficient than the Carnot cycle.

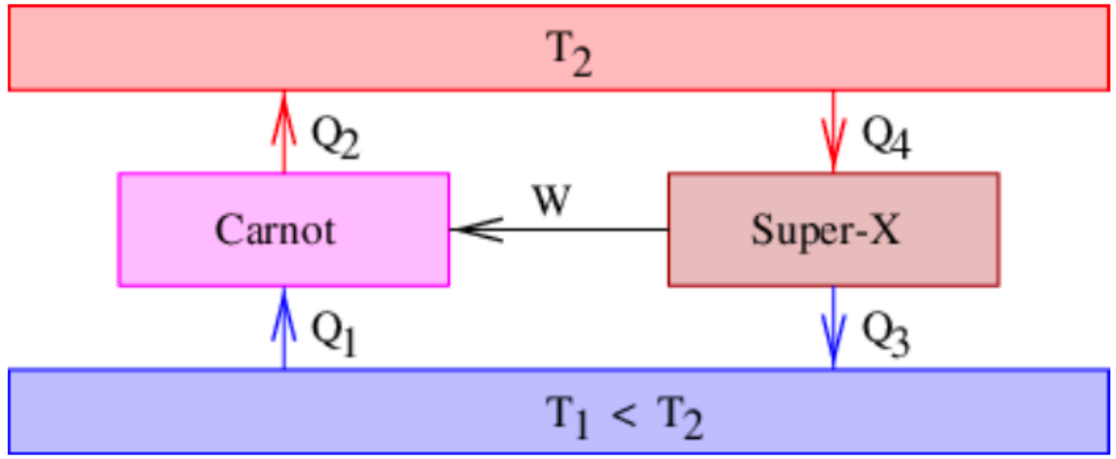
In other words, does there exist a heat engine operating between temperatures T_2 and T_1 that extracts more work from the high temperature heat input Q_2 than ϵQ_2?

Recall that $\epsilon = 1 - T_1/T_2$ is the thermodynamic efficiency of the Carnot engine.

Let's suppose that an inventor has presented us with the "Super-X machine", which is purported to have a thermodynamic efficiency greater than a Carnot engine.

Figure 24.8 shows how we could set up an experiment in our laboratory to test the inventor's claim.

Figure 24.8: Perpetual motion machine of the second kind. The Super-X machine is advertised as having a thermodynamic efficiency greater than a Carnot engine. The output of the Super-X machine runs the Carnot engine backwards as a refrigerator, resulting in net transfer of heat from the lower temperature to the higher temperature reservoir.



The Carnot engine runs in reverse as a refrigerator, emitting heat Q_2 to the upper reservoir, absorbing Q_1 from the lower reservoir, and using the work $W = Q_2 - Q_1 = \epsilon\, Q_2$ from the Super-X machine.

The Super-X machine is operated in heat engine mode, emitting Q_3 to the lower reservoir and absorbing $Q_4 = Q_3 + W < W/\epsilon$ from the upper reservoir.

The inequality indicates that the ratio of work produced and heat extracted from the upper reservoir, W/Q_4, is greater for the Super-X machine than for an equivalent Carnot engine.

Let us examine the net heat flow out of the upper reservoir, $Q_{upper} = Q_4 - Q_2$.

Since $Q_4 < W/\epsilon = Q_2$, we find that $Q_{upper} < 0$.

In other words, the Super-X machine is extracting less heat from the upper reservoir than the Carnot engine is returning to this reservoir using the work produced by the Super-X machine.

The source of this energy is the lower reservoir, from which an equivalent amount of heat is being extracted.

The net effect of these two machines working together is a *spontaneous transfer of heat from a lower to a higher temperature*, since no outside energy source or entropy sink is needed to make it operate.

This is a violation of the second law of thermodynamics.

Therefore, the Super-X machine, if it truly works, is a perpetual motion machine of the second kind.

Though there have been many claims, no perpetual motion machine has been convincingly demonstrated.

Thus, heat engines are apparently incapable of converting all of the internal energy supplied to them to useful work, as this would require either an infinite input temperature or a zero output temperature.

As we have demonstrated, this source of inefficiency is intrinsic to all heat engines and is in addition to the usual sources of inefficiency such as friction and heat loss from imperfect insulation.

No heat engine, no matter how perfectly designed, can overcome this intrinsic inefficiency.

As a result of the second law of thermodynamics, we see that real heat engines, which are always less efficient than Carnot engines, produce useful work, W,

$$W < \epsilon Q_2 \quad (\text{real heat engines}), \tag{24.25}$$

where Q_2 is the amount of heat energy extracted from the upper reservoir.

On the other hand, refrigerators transfer heat Q_2 to the upper reservoir in the amount

$$Q_2 < W/\epsilon \quad \text{(real refrigerators)}, \tag{24.26}$$

where W is the work done on the refrigerator.