A Radically Modern Approach to Introductory Physics — Volume 2

David J. Raymond
Physics Department
New Mexico Tech
Socorro, NM 87801

January 7, 2009

Contents

Preface to April 2006 Edition

This text has developed out of an alternate beginning physics course at New Mexico Tech designed for those students with a strong interest in physics. The course includes students intending to major in physics, but is not limited to them. The idea for a "radically modern" course arose out of frustration with the standard two-semester treatment. It is basically impossible to incorporate a significant amount of "modern physics" (meaning post-19th century!) in that format. Furthermore, the standard course would seem to be specifically designed to discourage any but the most intrepid students from continuing their studies in this area — students don't go into physics to learn about balls rolling down inclined planes — they are (rightly) interested in quarks and black holes and quantum computing, and at this stage they are largely unable to make the connection between such mundane topics and the exciting things that they have read about in popular books and magazines.

It would, of course, be easy to pander to students — teach them superficially about the things they find interesting, while skipping the "hard stuff". However, I am convinced that they would ultimately find such an approach as unsatisfying as would the educated physicist.

The idea for this course came from reading Louis de Broglie's Nobel Prize address.[1] De Broglie's work is a masterpiece based on the principles of optics and special relativity, which qualitatively foresees the path taken by Schrödinger and others in the development of quantum mechanics. It thus dawned on me that perhaps optics and waves together with relativity could form a better foundation for all of physics than does classical mechanics.

Whether this is so or not is still a matter of debate, but it is indisputable that such a path is much more fascinating to most college freshmen interested in pursing studies in physics — especially those who have been through the

[1] Reprinted in: Boorse, H. A., and L. Motz, 1966: *The world of the atom.* Basic Books, New York, 1873 pp.

usual high school treatment of classical mechanics. I am also convinced that the development of physics in these terms, though not historical, is at least as rigorous and coherent as the classical approach.

The course is tightly structured, and it contains little or nothing that can be omitted. However, it is designed to fit into the usual one year slot typically allocated to introductory physics. In broad outline form, the structure is as follows:

- Optics and waves occur first on the menu. The idea of group velocity is central to the entire course, and is introduced in the first chapter. This is a difficult topic, but repeated reviews through the year cause it to eventually sink in. Interference and diffraction are done in a reasonably conventional manner. Geometrical optics is introduced, not only for its practical importance, but also because classical mechanics is later introduced as the geometrical optics limit of quantum mechanics.

- Relativity is treated totally in terms of space-time diagrams — the Lorentz transformations seem to me to be quite confusing to students at this level ("Does gamma go upstairs or downstairs?"), and all desired results can be obtained by using the "space-time Pythagorean theorem" instead, with much better effect.

- Relativity plus waves leads to a dispersion relation for free matter waves. Optics in a region of variable refractive index provides a powerful analogy for the quantum mechanics of a particle subject to potential energy. The group velocity of waves is equated to the particle velocity, leading to the classical limit and Newton's equations. The basic topics of classical mechanics are then done in a more or less conventional, though abbreviated fashion.

- Gravity is treated conventionally, except that Gauss's law is introduced for the gravitational field. This is useful in and of itself, but also provides a preview of its deployment in electromagnetism. The repetition is useful pedagogically.

- Electromagnetism is treated in a highly unconventional way, though the endpoint is Maxwell's equations in their usual integral form. The seemingly simple question of how potential energy can be extended to the relativistic context gives rise to the idea of potential momentum.

The potential energy and potential momentum together form a four-vector which is closely related to the scalar and vector potential of electromagnetism. The Aharonov-Bohm effect is easily explained using the idea of potential momentum in one dimension, while extension to three dimensions results in an extension of Snell's law valid for matter waves, from which the Lorentz force law is derived.

- The generation of electromagnetic fields comes from Coulomb's law plus relativity, with the scalar and vector potential being used to produce a much more straightforward treatment than is possible with electric and magnetic fields. Electromagnetic radiation is a lot simpler in terms of the potential fields as well.

- Resistors, capacitors, and inductors are treated for their practical value, but also because their consideration leads to an understanding of energy in electromagnetic fields.

- At this point the book shifts to a more qualitative treatment of atoms, atomic nuclei, the standard model of elementary particles, and techniques for observing the very small. Ideas from optics, waves, and relativity reappear here.

- The final section of the course deals with heat and statistical mechanics. Only at this point do non-conservative forces appear in the context of classical mechanics. Counting as a way to compute the entropy is introduced, and is applied to the Einstein model of a collection of harmonic oscillators (conceptualized as a "brick"), and in a limited way to an ideal gas. The second law of thermodynamics follows. The book ends with a fairly conventional treatment of heat engines.

A few words about how I have taught the course at New Mexico Tech are in order. As with our standard course, each week contains three lecture hours and a two-hour recitation. The book contains little in the way of examples of the type normally provided by a conventional physics text, and the style of writing is quite terse. Furthermore, the problems are few in number and generally quite challenging — there aren't many "plug-in" problems. The recitation is the key to making the course accessible to the students. I generally have small groups of students working on assigned homework problems during recitation while I wander around giving hints. After all groups have

completed their work, a representative from each group explains their problem to the class. The students are then required to write up the problems on their own and hand them in at a later date. In addition, reading summaries are required, with questions about material in the text which gave difficulties. Many lectures are taken up answering these questions. Students tend to do the summaries, as their lowest test grade is dropped if they complete a reasonable fraction of them. The summaries and the associated questions have been quite helpful to me in indicating parts of the text which need clarification.

I freely acknowledge stealing ideas from Edwin Taylor, Archibald Wheeler, Thomas Moore, Robert Mills, Bruce Sherwood, and many other creative physicists, and I owe a great debt to them. My colleagues Alan Blyth and David Westpfahl were brave enough to teach this course at various stages of its development, and I welcome the feedback I have received from them. Finally, my humble thanks go out to the students who have enthusiastically (or on occasion unenthusiastically) responded to this course. It is much, much better as a result of their input.

There is still a fair bit to do in improving the text at this point, such as rewriting various sections and adding an index ... Input is welcome, errors will be corrected, and suggestions for changes will be considered to the extent that time and energy allow.

Finally, a word about the copyright, which is actually the GNU "copyleft". The intention is to make the text freely available for downloading, modification (while maintaining proper attribution), and printing in as many copies as is needed, for commercial or non-commercial use. I solicit comments, corrections, and additions, though I will be the ultimate judge as to whether to add them to my version of the text. You may of course do what you please to your version, provided you stay within the limitations of the copyright!

David J. Raymond
New Mexico Tech
Socorro, NM, USA
raymond@kestrel.nmt.edu

Chapter 13

Newton's Law of Gravitation

In this chapter we study the law which 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 plus 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.

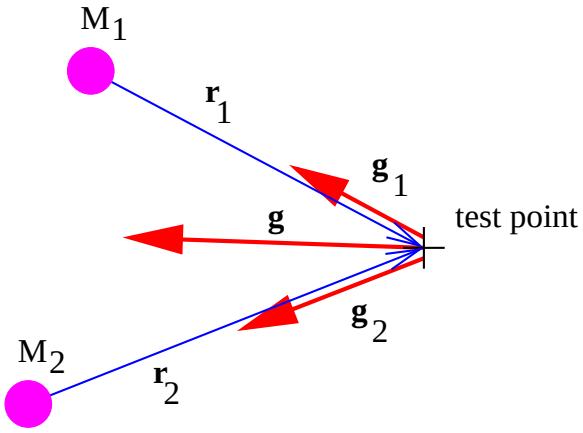

Figure 13.1: Sketch showing the addition of gravitational fields at a test point resulting from two masses.

13.2 Gravitational Field

The gravitational field at any point is 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 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 $\mathbf{r}$ 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.

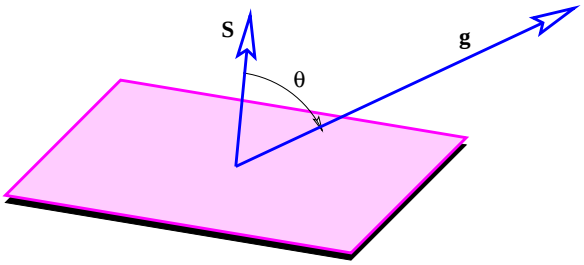

Figure 13.2: Definition sketch for the gravitational flux through the directed area **S**.

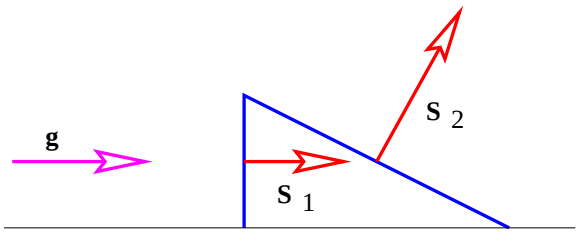

Figure 13.3: Two areas with the same projected area normal to **g**. Is the flux through area 2 greater than, less than, or equal to the flux through area 1? (The two areas are being viewed edge-on and are assumed to have some dimension d in the direction normal to the page.)

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. 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. It's 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 $\mathbf{g}$ 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 $\mathbf{S}$) 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, $\mathbf{S}_2 \cdot \mathbf{g}$.

The significance of the directedness of the area is now clear. If the vector $\mathbf{S}$ 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 $\mathbf{S}$ does — a positive flux is defined as going from the side opposite $\mathbf{S}$ to the side of $\mathbf{S}$.

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. It has magnitude $g = GM/R^2$, so if we

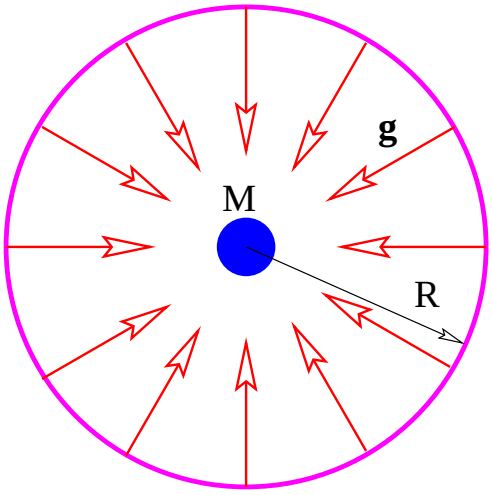

Figure 13.4: Calculation of the gravitational flux through the surface of a sphere with a mass at the center.

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. 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

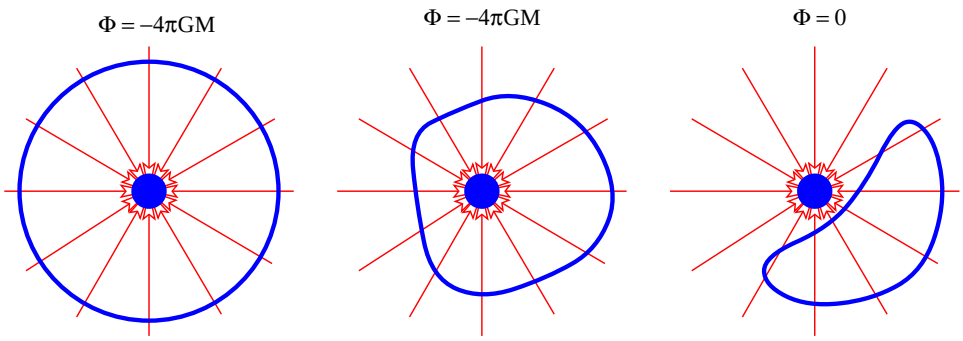

Figure 13.5: Three cases of a mass M and a closed surface. In the left and center examples the mass is inside the closed surface and the outward flux through the surface is $\Phi_g = -4\pi GM$. In the right example the mass is outside the surface and the outward flux through the surface is zero.

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

$$\Phi_g = -4\pi G \sum_{inside} M_i \quad \text{(Gauss's law)}. \tag{13.5}$$

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 $\Phi_g = -4\pi GM$ 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

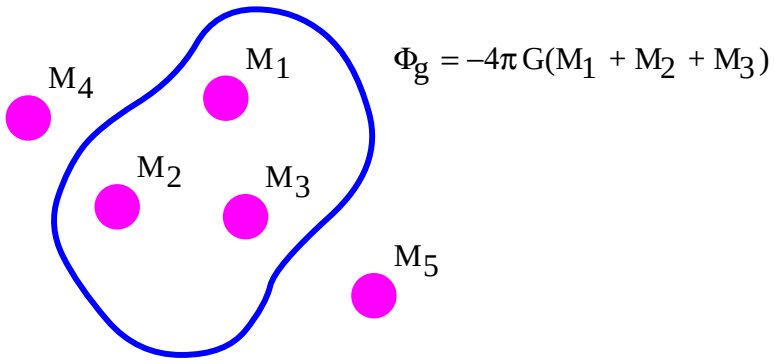

Figure 13.6: Gauss's law applied to more than one mass. The masses M_1, M_2, and M_3 contribute to the outward gravitational flux through the surface shown. The masses M_4 and M_5 don't contribute.

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 $\mathbf{g}$ resulting from a point mass M. Recall that $\mathbf{r}$ 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 to which we can transform which is everywhere inertial.

- 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 this geometry is a great circle. This is an example of a *geodesic curve*. 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 develop apparatus to detect gravitational waves produced by violent cosmic events such as the explosion of a supernova.

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,

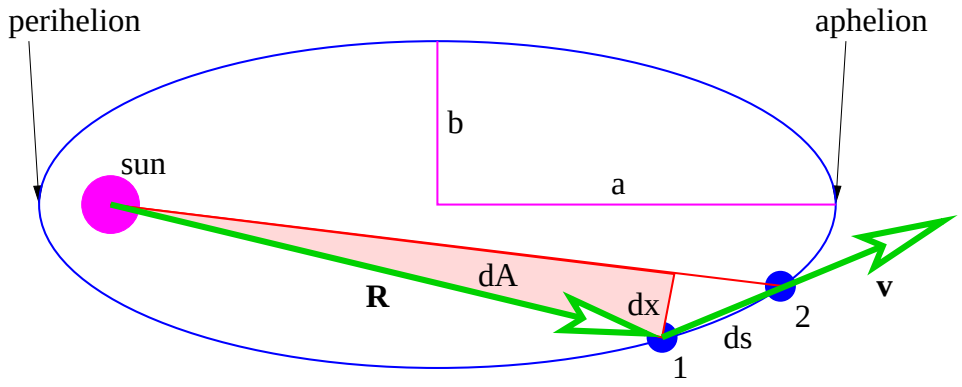

Figure 13.7: Illustration of elliptical orbit of a planet with the sun at the left focus. The *semi-major* and *semi-minor axes* are denoted by a and b. The shaded triangular area element is needed for the discussion of Kepler's second law. *Perihelion* and *aphelion* are respectively the points on the 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.

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. 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. Solution 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. 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

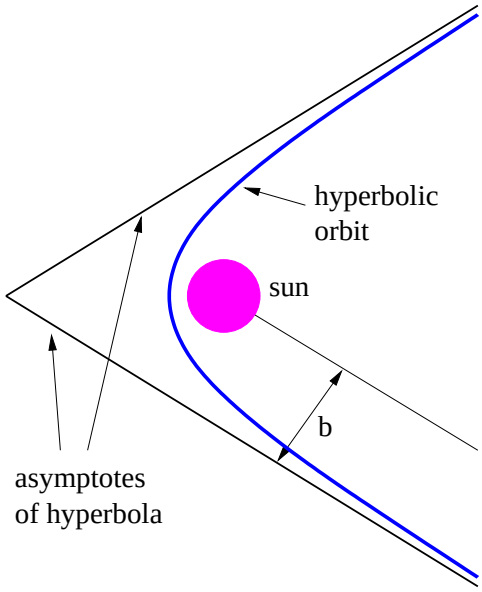

Figure 13.8: Example of a hyperbolic orbit of a positive energy object passing by the sun. The sun sits at the focus of the hyperbola. The quantity b is called the *impact parameter*.

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:

$$-E_b = \frac{L^2}{2mr^2} - \frac{GMm}{r} \quad \text{(aphelion and perihelion)}. \tag{13.17}$$

This can be rearranged into a quadratic equation

$$r^2 - \frac{GMm}{E_b} r + \frac{L^2}{2mE_b} = 0, \tag{13.18}$$

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 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.

13.8 Problems

1. Assume a mass M is located at $(-2\text{ m}, 0\text{ m})$ and a mass $2M$ is located at $(0\text{ m}, 3\text{ m})$. Find the (vector) gravitational field at the point $(1\text{ m}, 1\text{ m})$.

2. If two equal masses M are located at $\mathbf{x}_1 = (-3\text{ m}, -4\text{ m})$ and $\mathbf{x}_2 = (-3\text{ m}, +4\text{ m})$, determine where a third mass M must be placed to result in zero gravitational force at the origin.

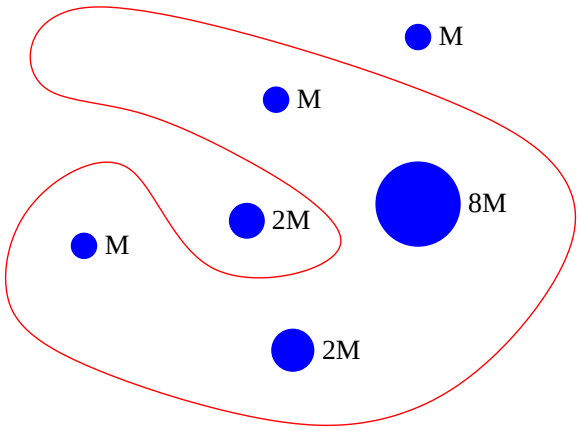

Figure 13.9: Various masses inside and outside a Gaussian surface.

3. Given the value of g at the Earth's surface, the radius of the Earth (look it up), and the universal gravitational constant G, determine the mass of the Earth.

4. Given the situation in figure 13.9:

   (a) What is the gravitational flux through the illustrated surface?

   (b) Explain why you cannot use this information to compute the gravitational field on the surface in this case.

5. Suppose mass is distributed uniformly with density μ kg m^{-1} in a thin line along the z axis. Try to figure out a way of using Gauss's law plus symmetry arguments to predict the gravitational field resulting from this mass distribution.

6. If the Earth is of uniform density, ρ, use Gauss's law to determine the gravitational field inside the Earth as a function of distance from the center.

7. Using the results of the above problem, determine the motion of an object moving through an evacuated hole drilled through the center of the Earth.

8. Two infinite thin sheets of mass, each with σ mass per unit area, are aligned perpendicular to each other. Determine the gravitational field from this combination. Hint: Compute $\mathbf{g}$ from each sheet separately and add vectorially.

9. Suppose that the universal law of gravitation says that the (attractive) gravitational force takes the form $F = -M_1 M_2 G r$, where r is the separation between the two masses M_1 and M_2 and G is a constant. Find the relationship between the orbital radius and the period for a circular orbit of a planet around the sun in this case.

10. An alien spaceship enters the solar system at distance D from the sun with speed v_0. (D may be considered to be very far from the sun.) It coasts through the solar system, approaching within a distance $d \ll D$ of the sun.

    (a) Find its speed at the point of closest approach.

    (b) Find the angular momentum of the spaceship with respect to the center of the sun.

    (c) What was the tangential component of the spaceship's velocity (i. e., the component normal to the radius vector) when it entered the solar system at distance D?

11. As a result of tidal torques, the spin angular momentum of the Earth is gradually being converted into orbital angular momentum of the moon, which causes the radius of its (circular) orbit to increase. Hint: Recall that for a solid sphere the moment of inertia is $I = 2mr^2/5$.

    (a) Obtain a relationship between the moon's orbital velocity and its distance from the Earth, assuming that the orbit is circular.

    (b) If the Earth's rotation rate is cut in half due to this effect, what will the new radius of the moon's orbit be?

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 electricity and magnetism 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 which 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

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 something (U) from the energy E without subtracting something from the momentum $\mathbf{\Pi}$ as well. However, $\underline{\Pi} = (\mathbf{\Pi}, E/c)$ 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 $\underline{Q}$ 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. In order to avoid confusion, we rename $\mathbf{\Pi}$ the *total momentum*.[1] 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 and 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*[2] and the *kinetic momentum* are related to the particle's velocity $\mathbf{v}$:

$$E - U = \frac{mc^2}{(1 - v^2/c^2)^{1/2}} \qquad \mathbf{p} = \mathbf{\Pi} - \mathbf{Q} = \frac{m\mathbf{v}}{(1 - v^2/c^2)^{1/2}}, \tag{14.10}$$

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 which depends on the existence of potential momentum. If the potential energy of a particle is zero and both the kinetic

[1] In advanced mechanics, $\mathbf{\Pi}$ is called the *canonical momentum*.

[2] In relativity, the quantity $E - U$ is actually equal to the kinetic plus the rest energy. This quantity ought to have a separate name but it does not.

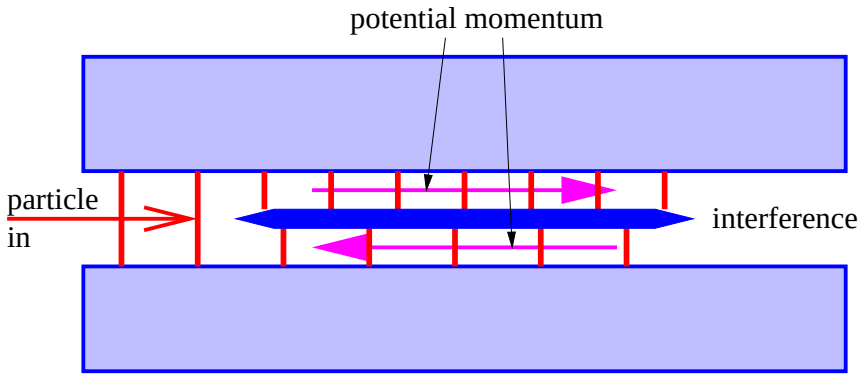

Figure 14.1: Setup for the Aharonov-Bohm effect. The particle moves through a channel which has a divided segment with non-zero potential momenta pointing in opposite directions in the two sub-channels. The vertical line segments show the wave fronts for the particle.

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 which splits into two sub-channels for an interval. 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

$$\mathbf{F} = -\left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z}\right) - \frac{\partial \mathbf{Q}}{\partial t} + \mathbf{u} \times \mathbf{P}, \tag{14.12}$$

where $\mathbf{u}$ 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 $\mathbf{u} \times \mathbf{P}$. 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.

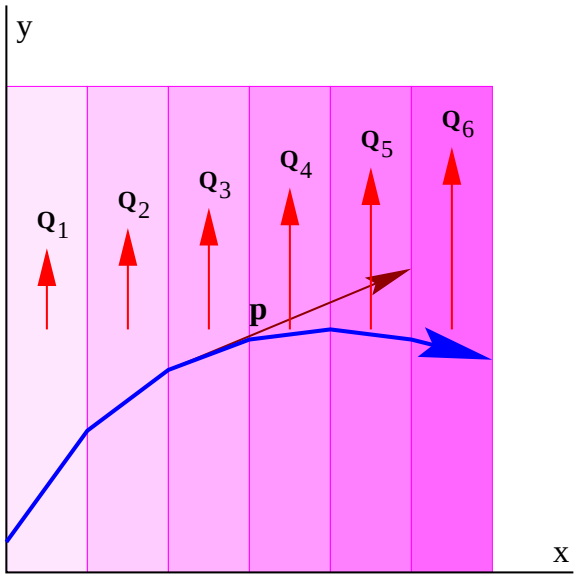

Figure 14.2: Trajectory of a wave packet through a region of variable potential momentum $\mathbf{Q}$. $\mathbf{Q}$ points in the y direction and increases in magnitude by steps with increasing x. The kinetic momentum $\mathbf{p}$ indicates the direction of motion of the associated particle at each point along the trajectory.

Furthermore, neither does the component of the wave vector parallel to the discontinuity. These two conditions together insure 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 $\mathbf{Q}$. 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:

$$F_y = \frac{dp_y}{dt} = \frac{dp_y}{dx}\frac{dx}{dt} = \frac{dp_y}{dx}u_x = -\frac{dQ_y}{dx}u_x. \tag{14.14}$$

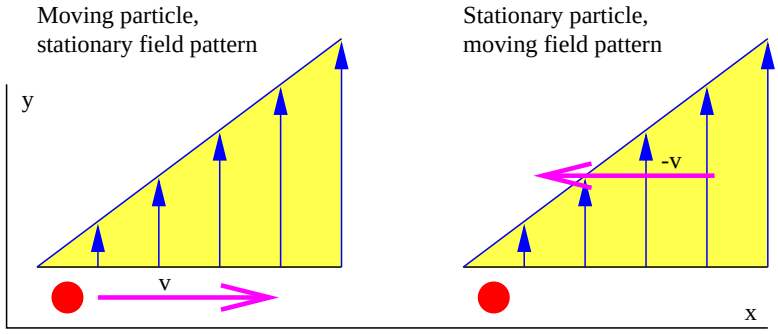

Figure 14.3: A moving particle and a stationary pattern of potential momentum **Q** must be equivalent to a stationary particle and a moving pattern of potential momentum according to the principle of relativity.

In the last step 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 the influence of the potential momentum:

$$F_x = \frac{dp_x}{dt} = \frac{dp_x}{dx}\frac{dx}{dt} = -\frac{p_y}{(p^2 - p_y^2)^{1/2}}\frac{dp_y}{dx}u_x = \frac{p_y}{p_x}\frac{dQ_y}{dx}u_x = \frac{dQ_y}{dx}u_y. \quad (14.15)$$

Aside from assuming that $p^2 = \textit{constant}$, 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. 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 $\mathbf{u}\times\mathbf{P} = 0$. 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 $\mathbf{u} \times \mathbf{P}$ in the original reference frame and the term $-\partial\mathbf{Q}/\partial t$ in the frame moving with the particle.

14.4 Lorentz Condition

It turns out that the four components of the potential four-momentum are not independent, but are subject to the condition

$$\frac{\partial Q_x}{\partial x} + \frac{\partial Q_y}{\partial y} + \frac{\partial Q_z}{\partial z} + \frac{1}{c^2}\frac{\partial U}{\partial t} = 0. \tag{14.16}$$

This is called the *Lorentz 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 subnuclear interactions are all of this type.

Gravity is the only fundamental force which 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. However, this theory has yet to be experimentally tested.

14.6 Conservation of Four-Momentum Again

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.[3] 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:

$$\underline{p} = \sum_i \underline{p}_i. \tag{14.18}$$

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. 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

[3] We use the symbol $\mathbf{p}$ for kinetic momentum here. However, in collisions we assume that the potential momentum and energy are only non-zero when the particles are very close together. Thus, when the particles are reasonably well separated, the distinction between kinetic and total momentum is unimportant.

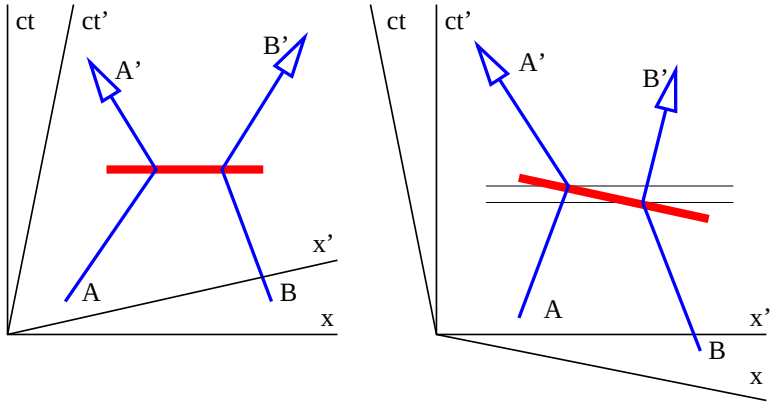

Figure 14.4: The trouble with action at a distance. View of the remote exchange of four-momentum from the point of view of two different coordinate systems. The fat line in both pictures is the line of simultaneity in the unprimed frame which is coincident with the exchange of four-momentum between the two particles.

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 action 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. 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 mediating particle exceeds the speed of light. This is reflected in the fact that different reference frames yield contradictory results as to whether the

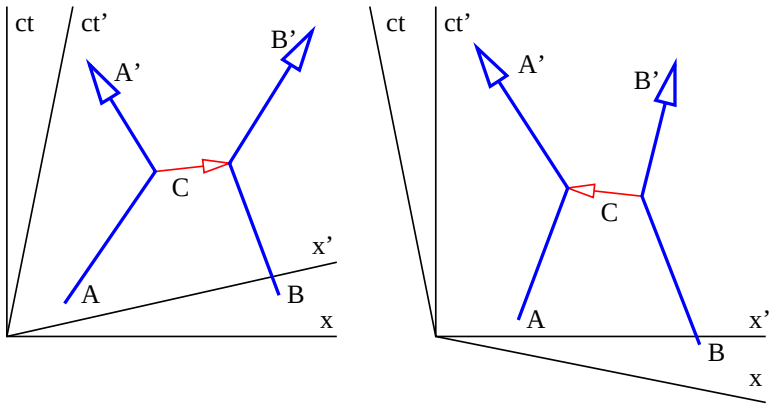

Figure 14.5: Mediation of action at a distance by a third particle. Notice that the world line of the mediating particle has a slope less than unity, which means that it is nominally moving faster than the speed of light.

mediating 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 $\underline{q}$ 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 particle C can move from its source point while keeping its erroneous mass hidden by the uncertainty principle:

$$\Delta I \approx \frac{\hbar}{(m_C^2c^2 + q^2)^{1/2}}. \tag{14.24}$$

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 mediating particle mass m_C.

Two distinct possibilities exist. If the mediating 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 mediating particle has mass, the range is simply $\Delta I \approx \hbar/m_Cc$ when $q \ll m_Cc$. The range is thus constant and inversely proportional to the mass of the mediating particle for low momentum transfers. For large momentum transfer, i. e., when $q \gg m_Cc$, the range decreases from this value with increasing momentum transfer, as in the case of a massless mediating particle.

14.8 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 mediating interactions. Thus a larger potential four-momentum at some point means a higher probability of finding the related virtual particles at that point.

14.9 Negative Energies and Antiparticles

Figure 14.5 illustrates another oddity in the role of mediating particles in collisions. In the unprimed frame, 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. Solution of equation $E^2 = p^2c^2 + m^2c^4$ for the energy has a sign ambiguity which 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 $\underline{p}$ is equivalent to the corresponding *antiparticle* with four-momentum $-\underline{p}$. 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.

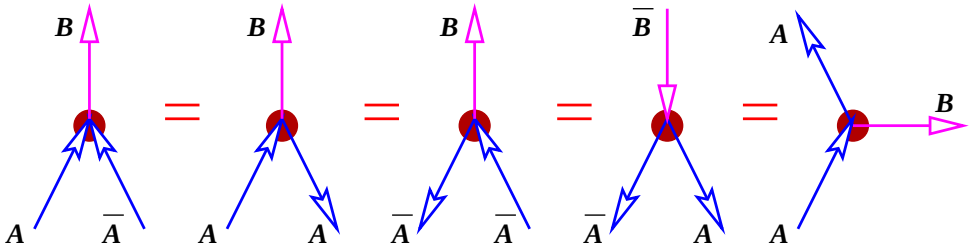

Figure 14.6: Equivalence of different processes according to Feynman's picture.

Suppose a particular kind of particle, call it an A particle, produces a B particle when it annihilates with its antiparticle $\overline{A}$. This is illustrated in the left panel in figure 14.6. 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 $\overline{A}$ backward in time by a B particle, the creation of an $A\overline{A}$ pair moving backward in time by a $\overline{B}$ 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 mediating particle going from left to right is a C particle, then the mediating particle going from right to left in the other picture is a $\overline{C}$ particle, or an anti$C$. It is immaterial whether the arrow representing either the C or the $\overline{C}$ 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.

14.10 Problems

1. An alternate way to modify the energy-momentum relation while maintaining relativistic invariance is with a "potential mass", $H(x)$:

$$E^2 = p^2c^2 + (m + H)^2c^4.$$

   If $|H| \ll m$ and $p^2 \ll m^2c^2$, show how this equation may be approximated as

$$E = something + p^2/(2m)$$

   and determine the form of "*something*" in terms of H. Is this theory distinguishable from the theory involving potential energy at nonrelativistic velocities?

2. For a given channel length L and particle speed in figure 14.1, determine the possible values of potential momentum $\pm Q$ in the two channels which result in destructive interference between the two parts of the particle wave.

3. Show that equations (14.14) and (14.15) are indeed recovered from equations (14.12) and (14.13) when $\mathbf{Q}$ points in the y direction and is a function only of x.

4. Show that the force $\mathbf{F} = \mathbf{u} \times \mathbf{P}$ is perpendicular to the velocity $\mathbf{u}$. Does this force do any work on the particle? Is this consistent with the fact that the force doesn't change the particle's kinetic energy?

5. Show that the potential momentum illustrated in figure 14.2 satisfies the Lorentz condition, assuming that $U = 0$. Would the Lorentz condition be satisfied in this case if $\mathbf{Q}$ depended only on x and pointed in the x direction?

6. A mass m moves at non-relativistic speed around a circular track of radius R as shown in figure 14.7. The mass is subject to a potential momentum vector of magnitude Q pointing counterclockwise around the track.

   (a) If the particle moves at speed v, does it have a longer wavelength when it is moving clockwise or counterclockwise? Explain.

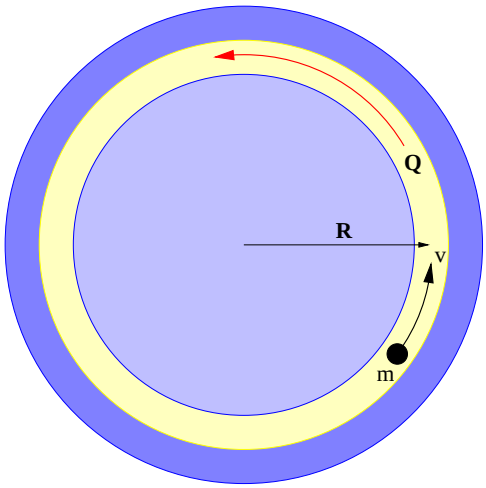

Figure 14.7: The particle is constrained to move along the illustrated track under the influence of a potential momentum **Q**.

(b) Quantization of angular momentum is obtained by assuming that an integer number of wavelengths n fits into the circumference of the track. For given $|n|$, determine the speed of the mass (i) if it is moving clockwise ($n < 0$), and (ii) if it is moving counterclockwise ($n > 0$).

(c) Determine the kinetic energy of the mass as a function of n.

7. Suppose momentum were conserved for action at a distance in a particular reference frame between particles 1 cm apart as in the left panel of figure 14.4 in the text. If you are moving at velocity 2×10^8 m s^{-1} relative to this reference frame, for how long a time interval is momentum apparently not conserved? Hint: The 1 cm interval is the *invariant* distance between the kinks in the world lines.

8. An electron moving to the right at speed v collides with a positron (an antielectron) moving to the left at the same speed as shown in figure 14.8. The two particles annihilate, forming a virtual photon, which then decays into a proton-antiproton pair. The mass of the electron is m and the mass of the proton is $M = 1830m$.

(a) What is the mass of the virtual photon? Hint: It is *not* $2m$. Why?

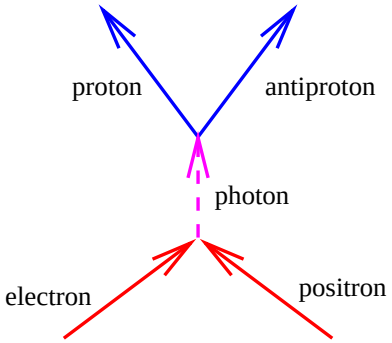

Figure 14.8: Electron-positron annihilation leading to proton-antiproton production.

(b) What is the maximum possible lifetime of the virtual photon by the uncertainty principle?

(c) What is the minimum v the electron and positron need to have to make this reaction energetically possible? Hint: How much energy must exist in the proton-antiproton pair?

9. A muon (mass m) interacts with a proton as shown in figure 14.9, so that the velocity of the muon before the interaction is v, while after the interaction it is $-v/2$. The interaction is mediated by a single virtual photon. Assume that $v \ll c$ for simplicity.

(a) What is the momentum of the photon?

(b) What is the energy of the photon?

10. A photon with energy E and momentum E/c collides with an electron with momentum $p = -E/c$ and mass m. The photon is absorbed, creating a virtual electron. Later the electron emits a photon with energy E and momentum $-E/c$. (This process is called Compton scattering and is illustrated in figure 14.10.)

(a) Compute the energy of the electron before it absorbs the photon.

(b) Compute the mass of the virtual electron, and hence the maximum proper time it can exist before emitting a photon.

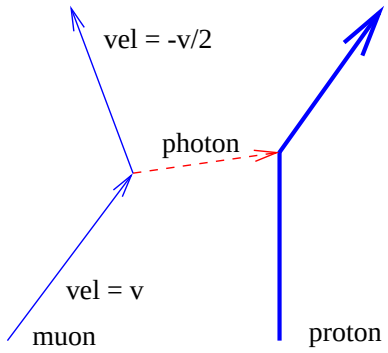

Figure 14.9: Collision of a muon with a proton, mediated by the exchange of a virtual photon.

(c) Compute the velocity of the electron before it absorbs the photon.

(d) Using the above result, compute the energies of the incoming and outgoing photons in a frame of reference in which the electron is initially at rest. Hint: Using $E_{photon} = \hbar\omega$ and the above velocity, use the Doppler shift formulas to get the photon frequencies, and hence energies in the new reference frame.

11. The dispersion relation for a negative energy relativistic particle is

$$\omega = -(k^2c^2 + \mu^2)^{1/2}.$$

Compute the group velocity of such a particle. Convert the result into an expression in terms of momentum rather than wavenumber. Compare this to the corresponding expression for a positive energy particle and relate it to Feynman's explanation of negative energy states.

12. The potential energy of a charged particle in a scalar electromagetic potential ϕ is the charge times the scalar potential. The total energy of such a particle at rest is therefore

$$E = \pm mc^2 + q\phi$$

where q is the charge on the particle and $\pm mc^2$ is the rest energy, the $\pm$ corresponding to positive and negative energy states. Assume that $|q\phi| \ll mc^2$.

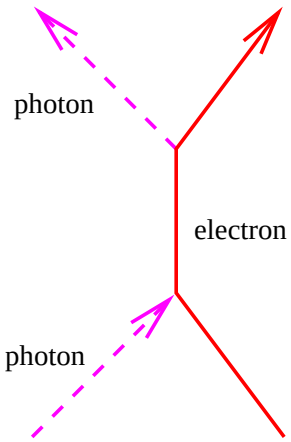

Figure 14.10: Compton scattering.

(a) Given that a particle with energy $E < 0$ is equivalent to the corresponding antiparticle with energy equal to $-E > 0$, what is the potential energy of the antiparticle?

(b) From this, what can you conclude about the charge on the antiparticle?

Hint: Recall that the total energy is always rest energy plus kinetic energy (zero in this case) *plus* potential energy.

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 which 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 ϕ/c such that $\underline{a} = (\mathbf{A}, \phi/c)$. 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 Lorentz condition written in terms of $\mathbf{A}$ 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

The forces caused by electric and magnetic fields are mostly what we can actually measure in electromagnetism. These vector quantities are related to the scalar and vector potentials as follows:

$$\mathbf{E} = -\left(\frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z}\right) - \frac{\partial \mathbf{A}}{\partial t} \quad \text{(electric field)} \tag{15.4}$$

$$\mathbf{B} \equiv \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) \quad \text{(magnetic field)}. \tag{15.5}$$

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 A Note on Units

We use the International System of units, often called SI units after the French translation of "International System". In SI units the meter (m), kilogram (kg), and second (s) are fundamental units of length, mass, and time. As previously noted, the unit of force is the Newton ($1\ \text{N} = 1\ \text{kg m s}^{-2}$) and the unit of energy is the Joule ($1\ \text{J} = 1\ \text{N m} = 1\ \text{kg m}^2\ \text{s}^{-2}$).

Electromagnetism introduces a new fundamental unit, the Coulomb (C), which is the unit for electric charge. The Lorentz force law tells us that the

electric field has the units N C^{-1}. The magnetic field has its own derived unit, i. e., one which can be expressed in terms of fundamental units, namely the Tesla (T): 1 T = 1 N s C^{-1} m^{-1}. The vector potential has units T m, while the scalar potential again has its own derived unit, the volt (V): 1 V = 1 N m C^{-1} = 1 J C^{-1}. The electric field can also be expressed in units of V m^{-1}. A commonly used unit for magnetic field is the Gauss (G). This non-SI unit is related to the Tesla as follows: 1 G = 10^{-4} T.

The charge on the electron is $-e = -1.60\times10^{-19}$ C. (The sign is arranged so that e is positive.) A commonly used non-SI unit for energy is the electron volt (eV). This is the energy gained by an electron passing through a potential difference of 1 V. Thus, 1 eV = 1.60×10^{-19} J. Commonly used multiples of the electron volt are 1 KeV = 10^3 eV, 1 MeV = 10^6 eV, 1 GeV = 10^9 eV, and 1 TeV = 10^{12} eV.

The electric current is the amount of charge passing some point per unit time. The unit of current is the Ampère (A): 1 A = 1 C s^{-1}.

15.4 Charged Particle Motion

We now explore some examples of the motion of charged particles under the influence of electric and magnetic fields.

15.4.1 Particle in Constant Electric Field

Suppose a particle with charge q is exposed to a constant electric field E_x 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.

15.4.2 Particle in Conservative Electric Field

If $\partial\mathbf{A}/\partial t = 0$, then the electric force on a charged particle is

$$\mathbf{F}_{electric} = -q\left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right), \qquad \frac{\partial\mathbf{A}}{\partial t} = 0. \tag{15.7}$$

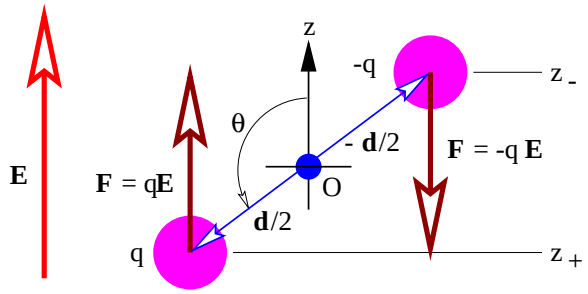

Figure 15.1: Definition sketch for an electric dipole. Two charges, q and $-q$ are connected by an uncharged bar of length d. The vectors $\mathbf{d}/2$ and $-\mathbf{d}/2$ give the positions of the two charges relative to the central point between them. The two forces are due to the electric field $\mathbf{E}$.

This force is conservative, with potential energy $U = q\phi$. 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.

15.4.3 Torque on an Electric Dipole

Let us now imagine a "dumbbell" consisting of positive and negative charges of equal magnitude q separated by a distance d, as shown in figure 15.1. If there is a uniform electric field $\mathbf{E}$, 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 $\mathbf{d}$ 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, but not either quantity individually. 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 $\mathbf{p}$ is parallel to the electric field $\mathbf{E}$. 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 electrostatic potential associated with the electric field is $\phi = -Ez$ 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 $U = q\phi = -qEz$. 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.

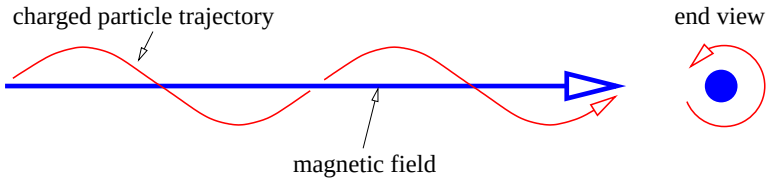

Figure 15.2: Spiraling motion of a charged particle in the direction of the magnetic field. This is composed of a circular motion about the field vector plus a translation along the field.

15.4.4 Particle in Constant Magnetic Field

The magnetic force on a particle with charge q moving with velocity $\mathbf{v}$ is $\mathbf{F}_{magnetic} = q\mathbf{v} \times \mathbf{B}$, 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 $F_{magnetic} = qvB\sin(\theta)$ 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 $F_{magnetic} = qvB$. 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: $v^2/R = a = F_{magnetic}/m = qvB/m$ 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

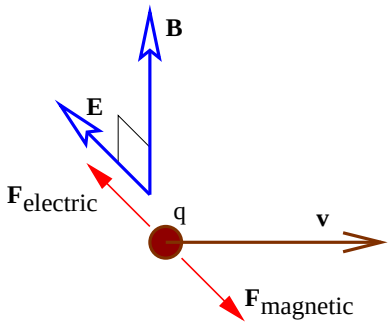

Figure 15.3: With crossed electric **E** and magnetic **B** fields (i. e., fields perpendicular to each other), a charged particle can move at a constant velocity **v** with magnitude equal to $v = |\mathbf{E}|/|\mathbf{B}|$ and direction perpendicular to both **E** and **B**. This is because the electric and magnetic forces, $\mathbf{F}_{electric} = q\mathbf{E}$ and $\mathbf{F}_{magnetic} = q\mathbf{v} \times \mathbf{B}$, balance each other in this case.

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.

15.4.5 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. 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 which 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

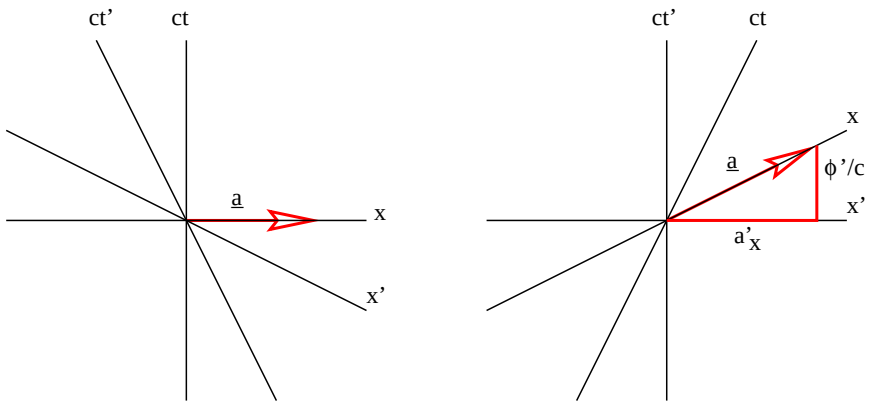

Figure 15.4: The four-potential $\underline{a}$ and its components in two different reference frames. In the unprimed frame the four-potential is purely space-like. The primed frame is moving in the x direction at speed U relative to the unprimed frame. The four-potential points along the x axis.

frame is not necessarily the same as the electric field perceived in another frame. Figure 15.4 shows why this is so. 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 $\underline{a}$ 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.

15.5 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 sur-

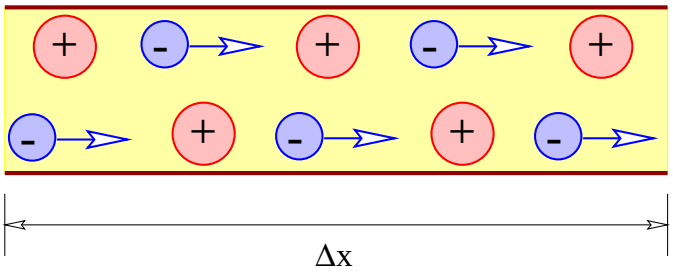

Figure 15.5: Fixed positive charge and negative charge moving to the right with speed v in blown up segment of wire.

rounded 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 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.

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

$$\mathbf{F} = iL\mathbf{n} \times \mathbf{B}. \tag{15.14}$$

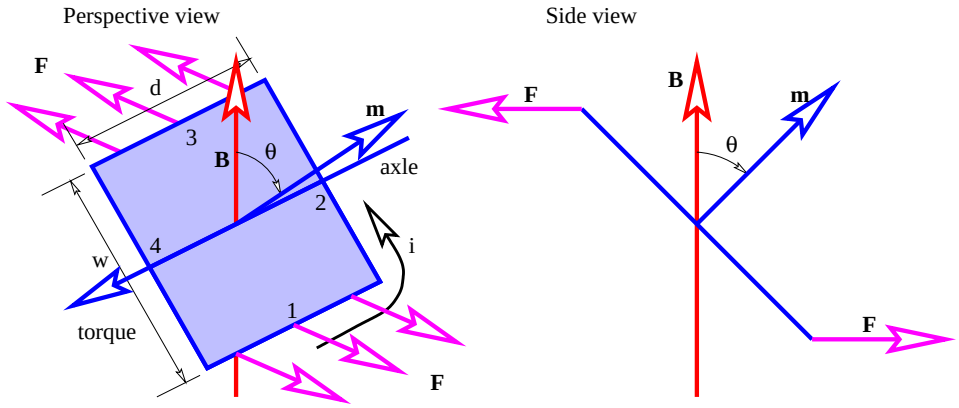

Figure 15.6: Perspective and side views of a rectangular loop of wire mounted on an axle in a magnetic field. Forces on the currents in loop segments 1 and 3 generate a torque about the axle.

15.6 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. 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 $\mathbf{m}$ 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 $\mathbf{m}$.

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.

15.7 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 $U = q\phi$, 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

$$\mathbf{E} = (-Cy, Cx, 0) \qquad \mathbf{B} = (0, 0, -2Ct). \tag{15.20}$$

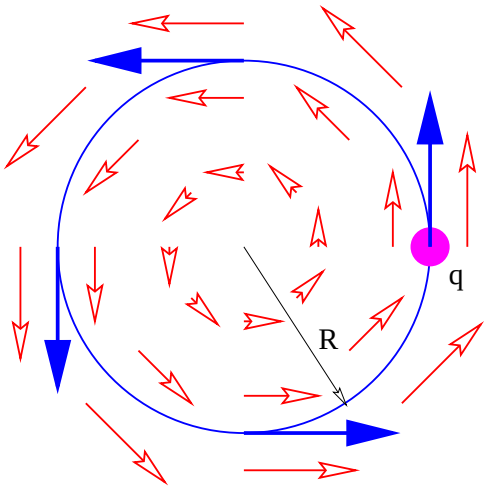

Figure 15.7: Illustration of the electric field pattern $\mathbf{E} = (-Cy, Cx, 0)$ (small arrows). A charged particle moving in a circle as shown continually gains energy.

The magnetic field points in the $-z$ direction and increases in magnitude with time. The electric field vectors are shown in figure 15.7. 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 $B_z = -2Ct$. Note that the time derivative of the magnetic field is just $\partial B_z/\partial t =$

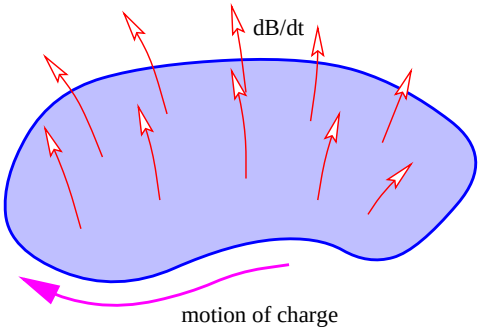

Figure 15.8: Sketch for Faraday's law. The arrows passing through the loop indicate the direction of the *time rate of change of the magnetic field*. The arrow going around the loop indicates the direction a positive charge would be pushed by the electric field.

$-2C$. 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 $\mathbf{B}$ perpendicular to the plane of the loop. The loop area times the normal component of $\mathbf{B}$ 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 which 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.

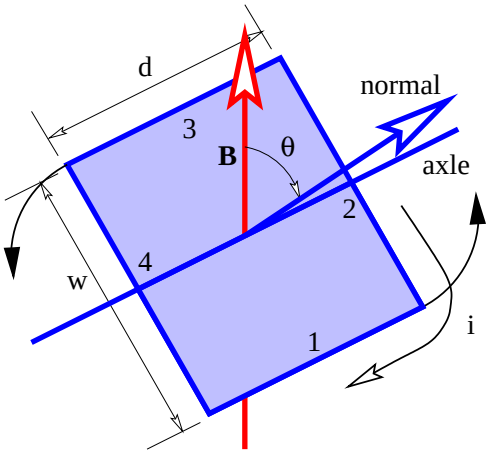

Figure 15.9: Rotating wire loop in a magnetic field. At the instant illustrated the magnetic flux is increasing with time, which means that an EMF tends to drive a current as illustrated.

The 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 which rotates at an angular rate Ω. 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.

15.8 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:

$$\mathbf{E} = -\left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right) - \frac{\partial\mathbf{A}}{\partial t}. \tag{15.24}$$

$\Delta\phi = \phi_2 - \phi_1$ 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.

Aside from the different sign conventions, there is one other fundamental difference between the two quantities: $\Delta\phi$ 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 $\phi_1 = \phi_2$. This condition doesn't necessarily apply to the EMF. ΔV often is non-zero for closed particle paths. The electric generator which 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 $\Delta V - \Delta\phi$. The $\Delta\phi$ term drops out if the path is closed.

15.9 Problems

1. Given a four-potential $\underline{a} = (Cyt, -Cxt, 0, 0)$ where C is a constant:

   (a) Determine whether this four-potential satisfies the Lorentz condition.

   (b) Compute the electric and magnetic fields from this four-potential.

2. Given $\underline{a}_1 = (C_1 zt, 0, 0, C_2 x)$, find the electric and magnetic field components. Compare with the fields you get from $\underline{a}_2 = (C_1 zt, C_3 y, -C_3 z, C_2 x)$. C_1, C_2, and C_3 are constants. Can one have more than one four-potential field giving rise to the same electric and magnetic fields?

3. Suppose that in the rest frame we have a four-potential of the form $\underline{a} = (0, 0, 0, Ky)$ where K is a constant.

   (a) Find the electric and magnetic fields in this frame.

   (b) Find the components of $\underline{a}$ in a reference frame moving in the $-x$ direction at speed U. Hint: Draw a spacetime diagram showing the $\underline{a}$ vector and resolve into components in the moving frame using the spacetime Pythagorean theorem.

   (c) Find the electric and magnetic fields in the moving frame.

4. Assume a four-potential of the form $\underline{a} = (\mathbf{A}, \phi/c)$, where $\mathbf{A} = (Ky, 0, 0)$ and $\phi = 0$ in the rest frame, K being a constant.

   (a) Compute the electric and magnetic fields in the rest frame.

   (b) Find the components of the four-potential in a reference frame moving in the $-x$ direction at speed U.

   (c) Compute the electric and magnetic fields in the moving frame using the above results.

5. Using the right-hand rule, show that the electric torque acting on an electric dipole tries to align the dipole so that it is in its state of lowest potential energy.

6. The net electric force on an electric dipole is zero in a *uniform* electric field. However, if the field varies with position, this is not necessarily true. Consider an electric field which has the form $\mathbf{E} = E_0(1 + \alpha z)\mathbf{k}$ along the z axis, where E_0 and α are positive constants.

   (a) An electric dipole consisting of charges $\pm q$ spaced by a distance d is centered at the origin. If the dipole is aligned with the electric field, determine the direction and magnitude of the net force on the dipole.

   (b) Determine the force on the dipole if it is *anti-aligned* with (i. e., pointing in the opposite direction from) the electric field.

7. Suppose that a charged particle is moving under the influence of electric and magnetic fields such that it periodically returns to some point P. If the four-potential is independent of time, will the kinetic energy of the particle be the same or different every time it returns to P? Explain.

8. Given constant electric and magnetic fields $\mathbf{E} = E\mathbf{j}$ and $\mathbf{B} = B\mathbf{k}$:

   (a) Find the velocity (magnitude and direction) of a charged particle for which the Lorentz force is zero.

   (b) Using this result, describe how you would build a setup to select out only those particles in a beam moving at a certain velocity.

9. Determine qualitatively how a charged particle moves in crossed electric and magnetic fields in the general case in which it is not moving at

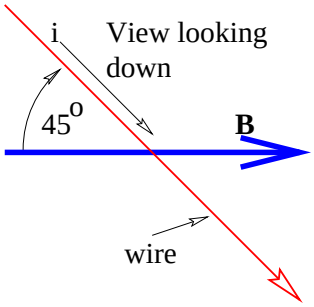

Figure 15.10: Horizontal wire with current i in a magnetic field.

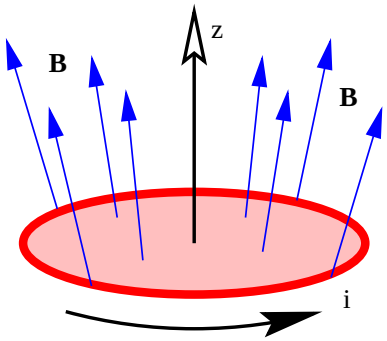

Figure 15.11: Magnetic dipole (current loop) in an inhomogeneous magnetic field.

constant velocity. For the sake of definiteness, assume that the magnetic field points in the $+z$ direction and the electric field in the $+x$ direction. Hint: Is there a reference frame in which the electric field vanishes? If there is, describe the motion in this reference frame and then determine how this motion looks in the original reference frame.

10. A horizontal wire of mass per unit length 0.1 kg m^{-1} passes through a horizontal magnetic field of strength $B = 0.1$ T with an orientation of 45° to the field as shown in figure 15.10. What current must the wire carry for the magnetic force on the wire to just balance gravity?

11. Figure 15.11 shows a current loop in a magnetic field. The magnetic field diverges with increasing z, so that its magnitude decreases with

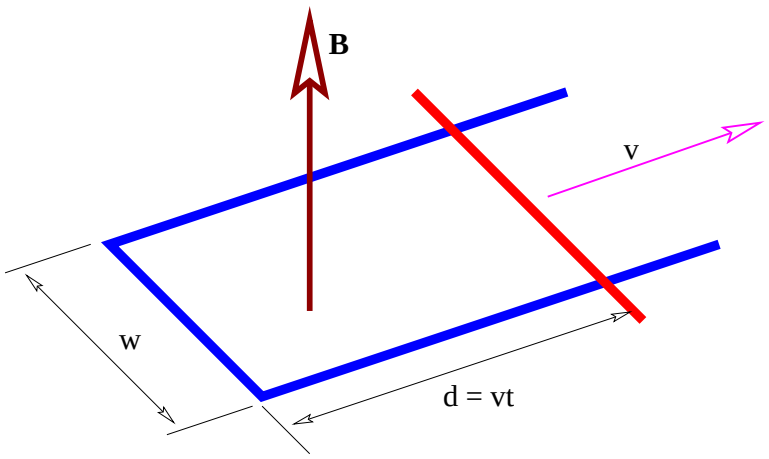

Figure 15.12: A moving crossbar on a U-shaped wire in a magnetic field.

height.

(a) Which way does the magnetic dipole vector due to the current loop point?

(b) Is this dipole oriented so as to have maximum or minimum potential energy, or is it somewhere in between?

(c) Is there a net force on the dipole? If so, what direction does it point? Hint: Determine the direction of the $\mathbf{v} \times \mathbf{B}$ force at each point on the current loop. What direction does the sum of all these forces point?

12. A charged particle moving in a circle in a magnetic field constitutes a circular current which forms a magnetic dipole.

(a) Determine whether the dipole moment produced by this current is aligned or anti-aligned with the initial magnetic field.

(b) Do charged particles moving in a non-uniform magnetic field as shown in figure 15.11 tend to accelerate toward regions of stronger or weaker field?

13. Why do electric motors have many turns of wire around the loop which cuts the magnetic field instead of just one? Hint: Magnetic fields

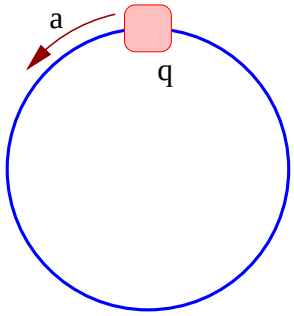

Figure 15.13: The charged bead continuously accelerates around the loop due to electromagnetic fields.

in normal motors are of order 0.1 T and currents are typically a few amps. Estimate the torque on a reasonably sized current loop for these conditions. Compare this to the torque you could expect to exert with your hand acting on a 1 m moment arm.

14. Imagine a stationary U-shaped conductor with a moving conducting bar in contact with the U as shown in figure 15.12. A uniform magnetic field exists normal to the plane of the U and has magnitude B. The bar is moving outward along the U at speed v as shown.

    (a) Using the fact that the charged particles in the moving bar are subject to a Lorentz force due to the motion of the bar through a magnetic field, compute the EMF around the closed loop consisting of the bar and the U. Hint: Recall that the EMF is the work done per unit charge on a charged particle moving around the loop.

    (b) Compute the EMF around the above loop using Faraday's law. Is the answer the same as obtained above?

15. A bead on a loop has a positive charge q and accelerates continuously around the loop in the counterclockwise direction, as shown in figure 15.13. Explain qualitatively what this information tells you about

    (a) the vector potential in the vicinity of the loop, and

    (b) the magnetic flux through the loop.

Chapter 16

Generation of Electromagnetic Fields

In this section 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 plus 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 times 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 obtained by multiplying the electric field produced by one charge 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 $\mathbf{S}$ is the directed area through which the flux passes. 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:

$$\Phi_E = q_{inside}/\epsilon_0 \quad \text{(Gauss's law for electricity)}, \tag{16.6}$$

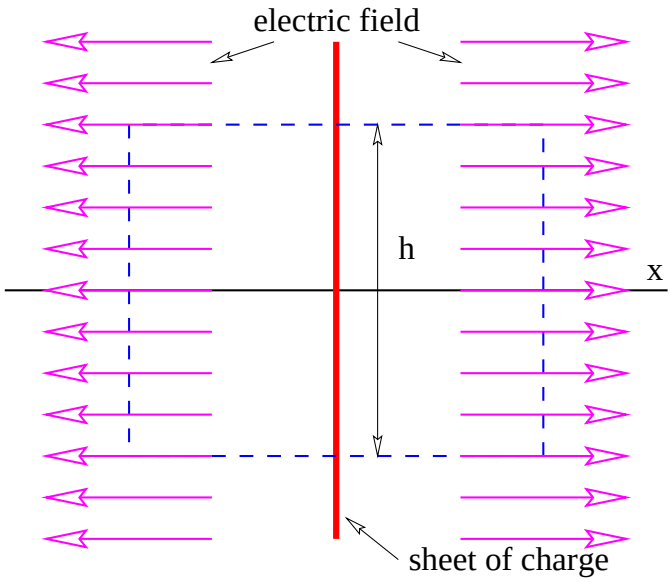

Figure 16.1: Definition sketch for use of Gauss's law to obtain the electric field due to an infinite sheet of surface charge. The dashed line shows the Gaussian box, which is of height h and depth d into the page.

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. 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.

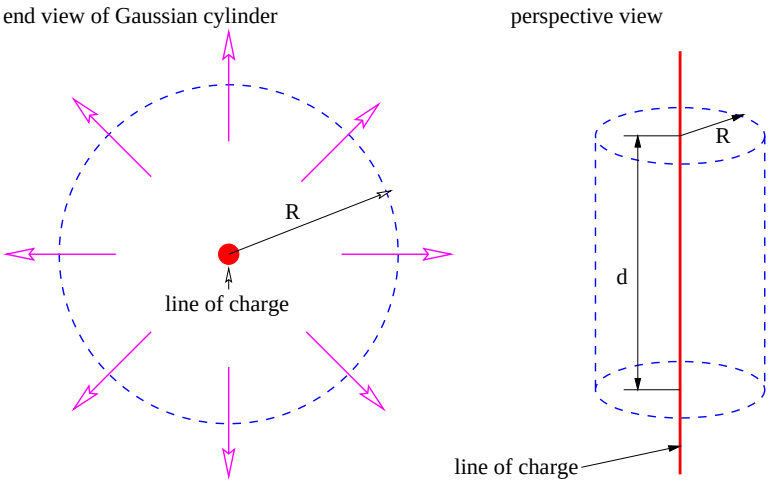

Figure 16.2: Definition sketch for use of Gauss's law to obtain the electric field due to an infinite line charge oriented normal to the page. The dashed line shows the Gaussian cylinder, which is of radius R and length d into the page. The outward-pointing arrows show the electric field.

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 $E = -\partial\phi/\partial x$, 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

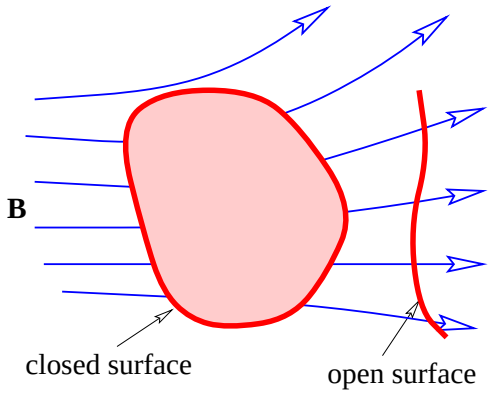

Figure 16.3: Illustration for Gauss's law for magnetism. The net flux out of the closed surface is zero, but the flux through the open surface is not.

with the line of charge is shown in figure 16.2. 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 $E = -\partial\phi/\partial r$, 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.

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 time-like 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 space-like 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. 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, $\phi'^2/c^2 = \phi^2/c^2 - A_x^2$, and relating

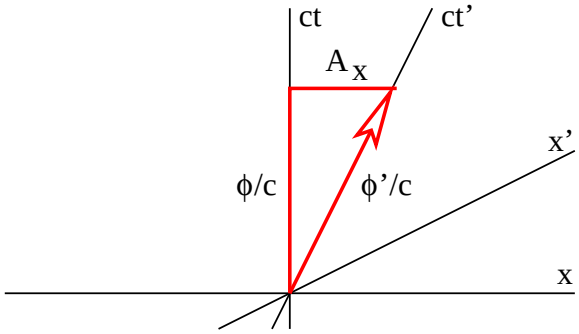

Figure 16.4: Finding the space and time components of the four-potential produced by a particle moving at the velocity of the primed reference frame. The ct' axis is the world line of the charged particle which generates the four-potential.

slope of the ct' axis to the components of the four-potential, $c/v = (\phi/c)/A_x$, 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

$$\phi' = -\frac{\lambda'}{2\pi\epsilon_0}\ln(r) \tag{16.15}$$

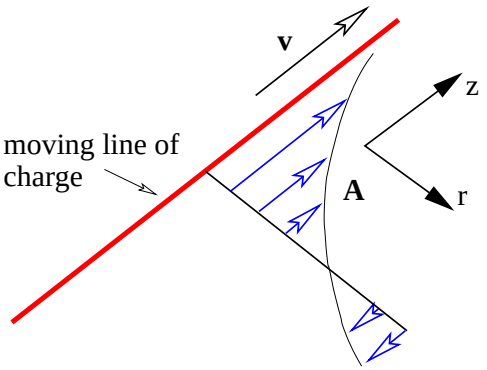

Figure 16.5: Vector potential from a moving line of charge. The distribution of vector potential around the line is cylindrically symmetric.

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.

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. 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 $\lambda = \gamma\lambda'$.

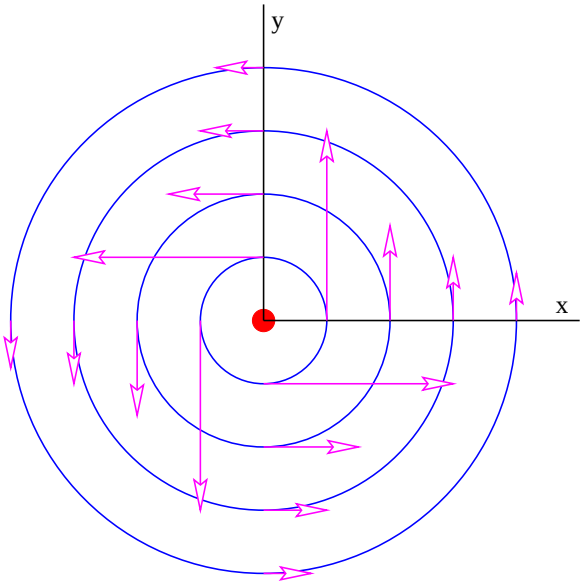

Figure 16.6: Magnetic field from a moving line of charge. The charge is moving along the z axis out of the page.

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 electromagetism 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

$$E'_z = \frac{\sigma'}{2\epsilon_0}\text{sgn}(z) \tag{16.21}$$

where we define

$$\text{sgn}(z) \equiv \begin{cases} -1 & z < 0 \\ 0 & z = 0 \\ 1 & z > 0 \end{cases} . \tag{16.22}$$

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).

The scalar potential in this frame is

$$\phi' = -\frac{\sigma'|z|}{2\epsilon_0}. \tag{16.23}$$

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

$$B_x = 0 \qquad B_y = \frac{dA_x}{dz} = -\frac{v\sigma}{2\epsilon_0 c^2}\text{sgn}(z) \qquad B_z = 0 \tag{16.25}$$

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.

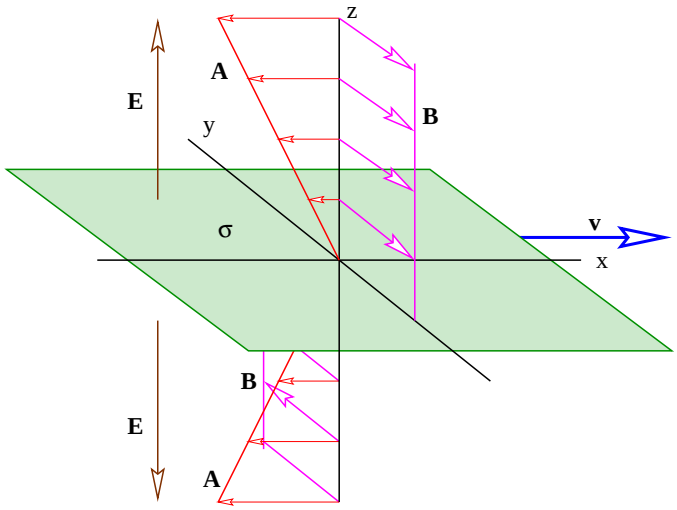

Figure 16.7: Vector potential **A**, electric field **E**, and magnetic field **B** from a moving sheet of charge. The charge is moving in the x direction.

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 which 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. 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 which can generate radiation occurs when two

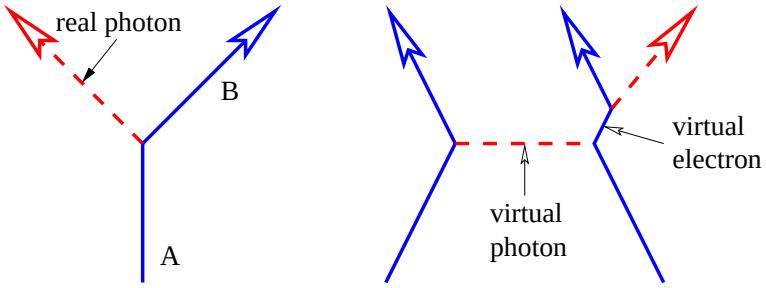

Figure 16.8: Feynman diagrams for two processes which potentially might produce real photons and hence electromagnetic radiation. The process in the left panel turns out to be impossible if the masses of particles A and B are the same, for reasons discussed in the text. The process in the right panel occurs commonly. Solid lines represent electrons while dashed lines represent photons. Particles are taken to be real unless otherwise labeled.

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 which 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 $\underline{a}_0 = (\mathbf{A}_0, \phi_0/c)$ 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 $\mathbf{A}$ and ϕ from equation (16.26) into the Lorentz condition, we find that

$$k_x A_x - \omega\phi/c^2 = 0 \quad \text{(Lorentz condition for plane wave).} \tag{16.27}$$

Thus, the Lorentz condition requires that the scalar potential ϕ be related to the x or *longitudinal component* of the vector potential, A_x, i. e., the

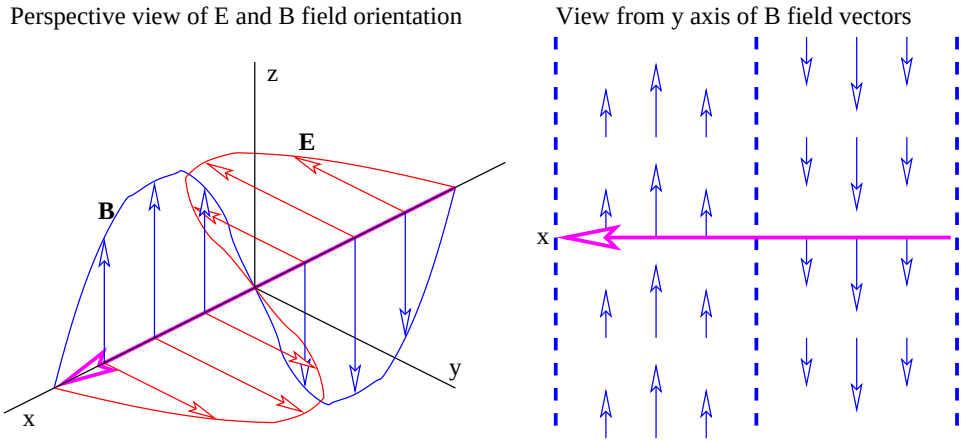

Figure 16.9: Electric and magnetic fields in a horizontally polarized plane wave, i. e., with $A_z = 0$, moving in the direction of the large arrow. The left panel shows how the electric and magnetic fields point, while the right panel shows the distribution of the magnetic field in space.

component pointing in the direction of wave propagation. The *transverse components*, A_y and A_z, are unconstrained by the Lorentz 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 $\mathbf{E}$ and $\mathbf{B}$ 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 which 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 $\mathbf{E}$ and $\mathbf{B}$ are perpendicular to each other. Furthermore, both $\mathbf{E}$ and $\mathbf{B}$ 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 $A_z = 0$. 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 Lorentz condition forces A_x to cancel with the term containing ϕ in the expression for E_x.

16.7 The Lorentz Condition

We are now in a position to see what the Lorentz condition means. For an isolated stationary charge, the scalar potential is given by equation (16.1) and the vector potential $\mathbf{A}$ is zero. The Lorentz 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 Lorentz 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 Lorentz condition is a consequence of 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 Lorentz condition is an equation which 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 Lorentz 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 Lorentz condition, as was done in equation (16.27) in the previous section. In this case the Lorentz condition reduces to $\underline{k} \cdot \underline{a} = 0$. Since the dot product of two four-vectors is a

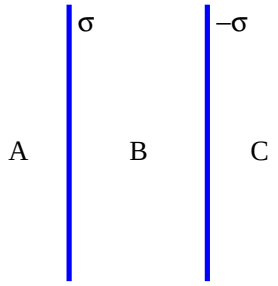

Figure 16.10: Two parallel sheets of charge, one with surface charge density σ, the other with $-\sigma$.

relativistic scalar, the Lorentz condition is equally valid in all frames. Since we believe that charge is indeed conserved in all circumstances, the Lorentz condition must always be satisfied.

16.8 Problems

1. Imagine that an electron actually consists of *two* point charges, each with charge $e/2$, separated by a distance D, where e is the charge on the electron. Compute D such that the potential energy of the two charges equals the rest energy of the electron. Look up the constants and compute a numerical value for D. Finally, compute the force between the two charges and compare to the gravitational force between two masses each equal to half the electron mass separated by this distance.

2. Verify that the equations for the scalar potentials associated with a sheet and a line of charge, (16.8) and (16.10), yield the corresponding electric fields.

3. Two sheets of charge, one with charge density σ, the other with $-\sigma$, are aligned as shown in figure 16.10. Compute the electric field in each of the regions A, B, and C.

4. Positive charge is distributed uniformly on the upper surface of an infinite conducting plate with charge per unit area σ as shown in figure 16.11. Use Gauss's law to compute the electric field above the plate. Hint: Is there any electric field inside the plate?

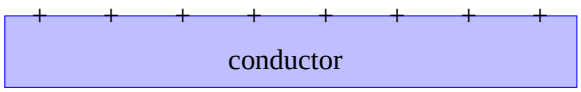

Figure 16.11: A charged metal plate.

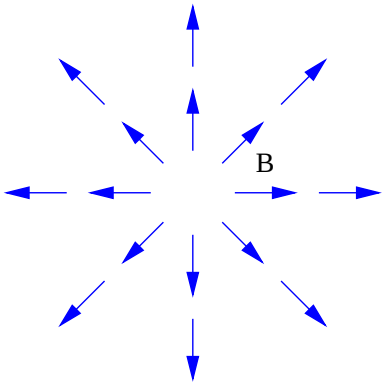

Figure 16.12: Hypothesized magnetic field. Does it satisfy Gauss's law for magnetism?

5. Suppose a student proposes that a magnetic field can take the form shown in figure 16.12. Is the proposed form of the magnetic field consistent with Gauss's law for magnetism? Explain.

6. The magnetic flux through the sides of the cone illustrated in figure 16.13 is zero. The magnetic field may be assumed to be approximately normal to the ends of the cone and the magnetic flux into the left end is Φ_B. The areas of the left and right ends of the cone are S_a and S_b.

   (a) What is the magnetic flux out of the right end of the cone?

   (b) What is the value of the magnetic field B on the left end of the cone?

   (c) What is the value of B on the right end?

7. In the lab frame a wire has negative charge with linear charge density $-\lambda$ moving at speed $-U$ corresponding to a current $i = \lambda U$ as shown in figure 16.14. Positive charge is stationary, and has charge density λ, so the net charge is zero.

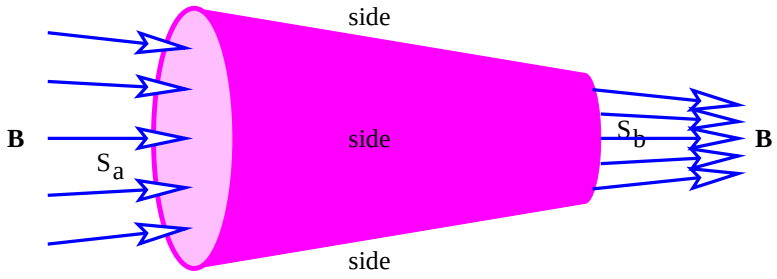

Figure 16.13: Converging magnetic field passing through a closed surface.

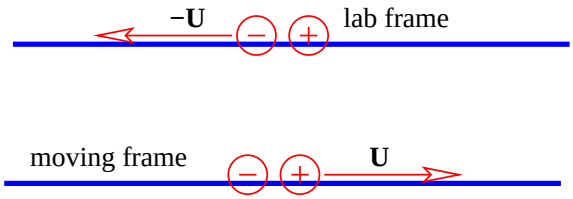

Figure 16.14: A horizontal wire with current i viewed in two different reference frames.

(a) What are the electric and magnetic fields produced by the charge in the wire in the stationary frame?

(b) In a reference frame moving at velocity $-U$ in the x direction, such that the negative charge is stationary, what is the apparent linear charge density of (1) the negative charge, and (2) the positive charge? Hint: The Lorentz contraction must be taken into account here.

(c) What is the electric field produced by the charge in the wire in the moving frame? Hint: Do the charge densities from the positive and negative charge cancel in this frame?

(d) What is the current in the wire in the moving frame, and hence, what is the magnetic field around the wire in this frame? Hint: Is the positive or negative charge causing the current in this frame?

(e) Explain why the net force on a separate charged particle some distance from the wire and stationary in the lab frame is zero in both reference frames.

8. The left panel of figure 16.8 shows a real charged particle A emitting a real photon, turning into a possibly different real particle B after the emission. If particle A and particle B have the same mass, show that this process is energetically impossible. Hint: Work in a reference frame in which particle A is stationary.

9. Given the four-potential for an electromagnetic plane wave, show why the longitudinal component of the magnetic field is zero.

10. Referring to figure 16.9, show that the vector $\mathbf{E} \times \mathbf{B}$ points in the direction of propagation of a plane electromagnetic wave.

11. Referring to figure 16.9, what direction and speed must a charged particle move in the presence of a free electromagnetic wave such that the net electromagnetic force on it is zero?

Chapter 17

Capacitors, Inductors, and Resistors

Various electronic devices are considered in this section. 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 treated. 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

In this subsection we first discuss a device which 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. This consists of two conducting plates of area S separated by distance d. 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

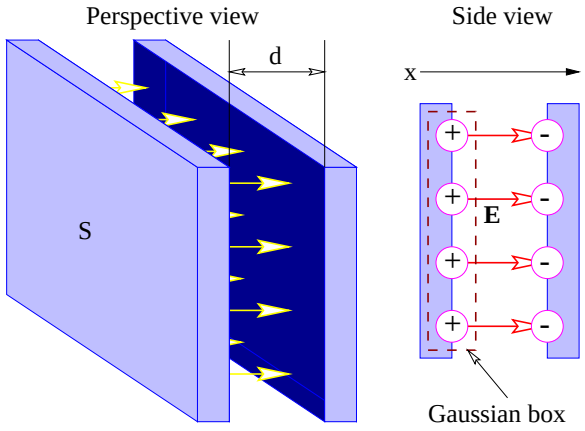

Figure 17.1: Two views of a parallel plate capacitor.

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 $-\sigma$. All charge is assumed to reside on the inside surfaces and thus contribute 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 $E_x = -\partial\phi/\partial x$ across the gap between the conducting plates, we find that the potential difference between the plates is $\Delta\phi = E_x d = qd/(\epsilon_0 S)$, since E_x is known to be constant in this case. This equation indicates that the potential difference $\Delta\phi$ 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

$$\Delta\phi = q/C \quad \text{or} \quad C = q/\Delta\phi. \tag{17.2}$$

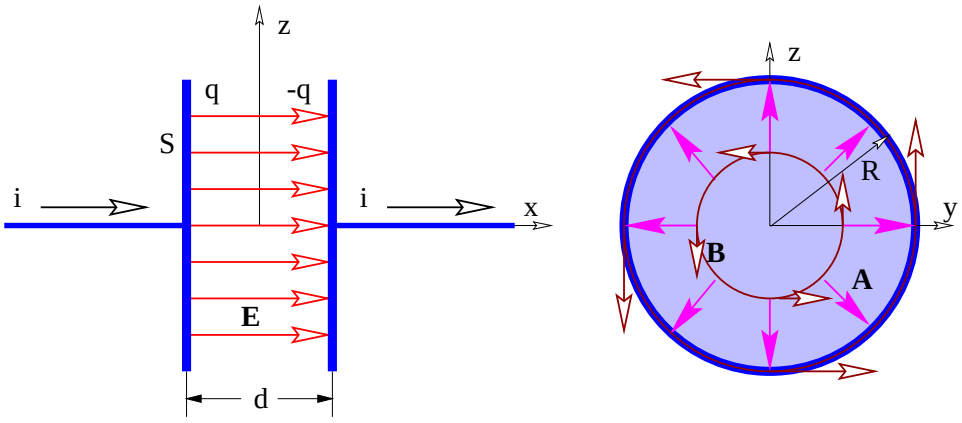

Figure 17.2: Parallel plate capacitor with circular plates in a circuit with current i flowing into the left plate and current i flowing out of the right plate. The magnetic field which occurs when the charge on the capacitor is increasing with time is shown at right as vectors tangent to circles. The 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.

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 which which 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 which 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. This is necessarily accompanied by an electric field which 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 which is a function of time: $\phi = -itx/(\epsilon_0 S)$. However, the Lorentz

condition

$$\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0 \tag{17.3}$$

demands that some component of the vector potential $\mathbf{A}$ be non-zero under these circumstances, since $\partial\phi/\partial t$ is non-zero.

How much can we infer about the vector potential from the geometry of the capacitor and equation (17.3)? Substituting $\phi = -itx/(\epsilon_0 S)$ 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 $\mathbf{A}$. 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 which is consistent with the shape of the capacitor and which satisfies the Lorentz 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 the previous section 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

$$\Gamma_E = -\frac{d\Phi_B}{dt} \quad \text{(Faraday's law)}. \tag{17.7}$$

[1]The terminology comes from fluid dynamics where the concept is used with the fluid velocity field. The idea of circulation is so useful in fluid dynamics that it seems worthwhile to generalize it to vector fields in other areas of physics.

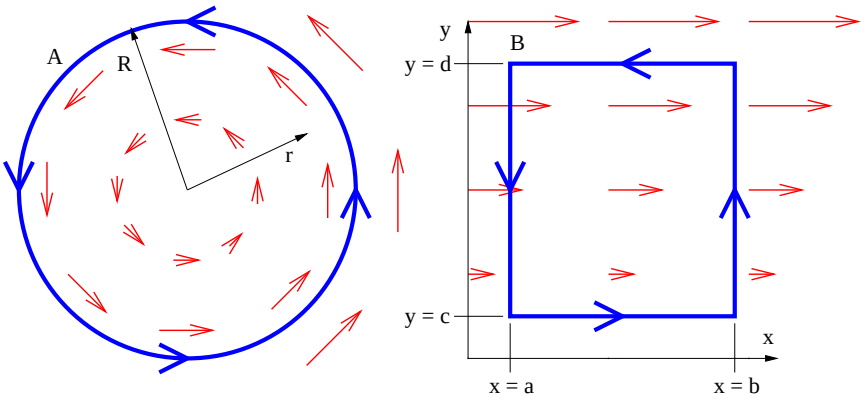

Figure 17.3: Two examples of circulation paths in a vector field.

In the simple case of a circular loop with the field directed along the loop, the circulation is just the magnitude of the field times the circumference of the loop, as illustrated in the left panel of figure 17.3. 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 $\mathbf{B}$ 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 $\Gamma_B = 2\pi rB = \mu_0 i$.

Notice that the same 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 $E = \sigma/\epsilon_0$ times 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:

$$\Gamma_B = \mu_0 \left(i + \epsilon_0 \frac{d\Phi_E}{dt} \right) \quad \text{(Ampère's law)}. \tag{17.8}$$

This states that the magnetic circulation around a loop equals the sum of two contributions, (1) μ_0 times the electric current through the loop and (2) $\mu_0\epsilon_0$ times 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 Lorentz 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 the time-changing electric flux!

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

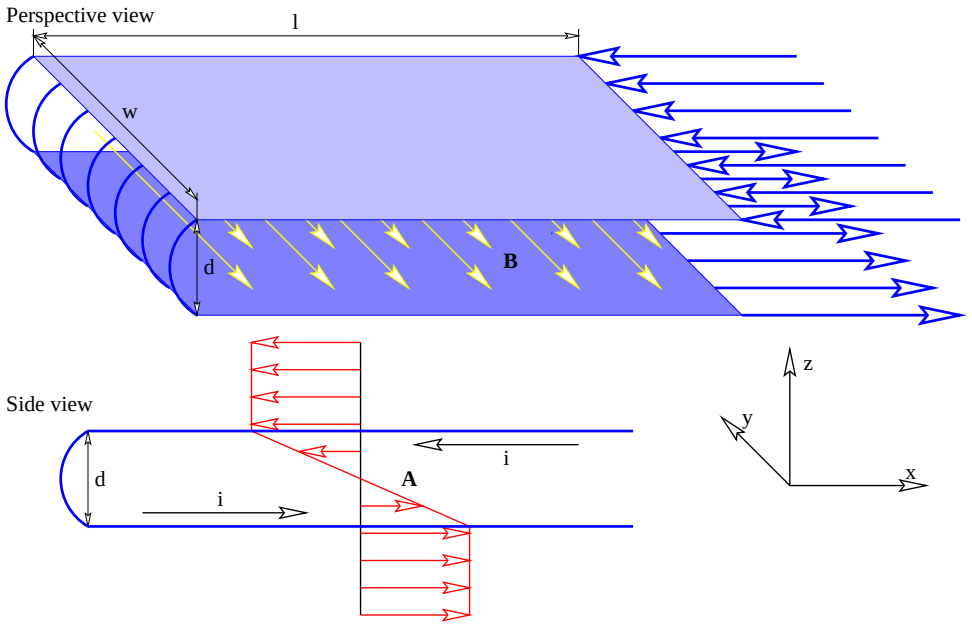

Figure 17.4: Magnetic field and vector potential for two parallel plates carrying equal currents in opposite directions.

of electromagnetism in terms of the four-potential, the dispersion relation for free electromagnetic waves, the Lorentz 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

Inductance 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. The vector potential between the plates is

$$\mathbf{A} = \left(-\frac{\mu_0 i z}{w}, 0, 0\right), \tag{17.9}$$

where w is the width of the plates, as illustrated in 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

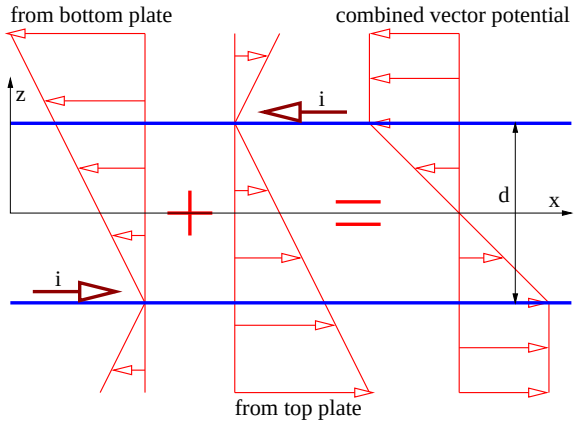

Figure 17.5: Illustration of the addition of the vector potentials from two current sheets with the left-moving current located above the x axis and the right-moving current below. The sum is obtained by the vector addition of the two components. Notice how the vector potential varies with z between the current sheets, but is constant outside of them.

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 A_x from each of the charge sheets add together to form a combined profile for the two sheets together.

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 device 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 $\Delta V = E_{x-upper}(-l) + E_{x-lower}(+l) = -(\mu_0 dl/w)(di/dt)$, 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. However, 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:

$$\Delta\phi = \frac{\mu_0 dl}{w}\frac{di}{dt} \quad \text{(parallel plate inductor).} \tag{17.13}$$

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

$$L = \frac{\mu_0 dl}{w} \quad \text{(parallel plate inductor).} \tag{17.15}$$

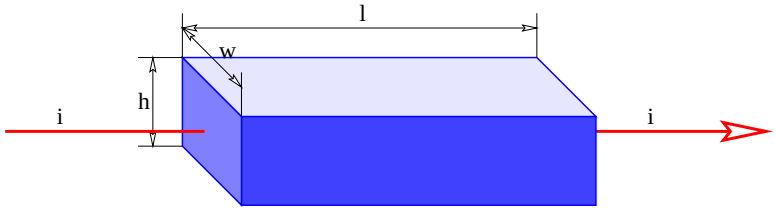

Figure 17.6: Rectangular resistor with a current i flowing through it.

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 $\Delta\phi$ 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 $\Delta\phi$ 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. 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 $q\Delta\phi$. This energy is converted to heat. The work done per unit time, which equals the power dissipated by a resistor is therefore

$$P = i\Delta\phi = i^2R = \Delta\phi^2/R. \tag{17.18}$$

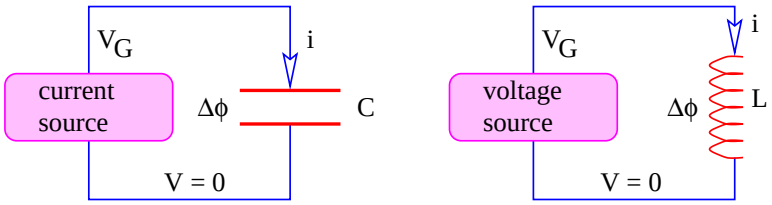

Figure 17.7: Capacitor (left) and inductor (right) being charged respectively by constant sources of current and voltage.

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. 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 $\Delta\phi$ 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 $dW = V_G dq = V_G i dt$. 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 $q = \epsilon_0 ES$ 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:

$$U_B = \frac{Li^2}{2} = \frac{\mu_0 dli^2}{2w} \quad \text{(parallel plate inductor)}. \tag{17.25}$$

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

$$U_B = \frac{dlwB^2}{2\mu_0}. \tag{17.26}$$

The volume between the inductor plates is just dlw, so again we can write an energy density, this time for the magnetic field:

$$u_B = \frac{U_B}{dlw} = \frac{B^2}{2\mu_0} \quad \text{(magnetic energy density)}. \tag{17.27}$$

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:

- The current entering one end of a wire connecting two devices is equal to the current leaving the other end. Thus, the current out of the generator in the left panel of figure 17.7 is assumed to equal the current into the capacitor.

- The net work done on a charged particle passing completely around a circuit loop is zero. Thus, 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.

These are called *Kirchhoff's laws*. They are used extensively in electronic circuit design.

It is important to realize that Kirchhoff's laws are only approximations which 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.

17.6 Problems

1. Compute the capacitance of an isolated conducting sphere of radius R. Hint: Consider the other electrode to be a spherical shell surrounding the conducting sphere at very large radius.

2. Given a parallel plate capacitor with plate area S, charge $\pm q$ on the plates, and the possibly variable plate separation x:

   (a) Is the force between the plates attractive or repulsive?

   (b) Compute the magnitude of the force of each plate on the other. Hint: You know both the electric field and the charge.

   (c) Make an alternate computation of the force as follows: Compute the energy U in the electric field between the plates. The force is $F = -dU/dx$.

   (d) You probably found that the above two calculations of the force didn't agree. Which is correct? Explain. Hint: In doing part (b), what part of the electric field acting on (say) the negative charge is due to itself, and what part is due to the positive charge? Only the latter part can exert a net force on the negative charge!

3. Compute the circulation of the vector field around the illustrated circle in the left panel of figure 17.3. Assume that the magnitude of the vector field equals Kr where K is a constant.

4. Compute the circulation of the vector field around the illustrated rectangle in the right panel of figure 17.3. Assume that the x component of the vector field equals Ky where K is a constant.

5. The solar wind consists of a plasma (a gas consisting of charged particles with equal amounts of positive and negative charge) streaming out from the sun. In certain sectors of the solar wind the magnetic field points away from the sun while in other sectors it points toward the sun. What is the magnitude and direction of the current flowing through the loop defined by the dashed rectangle which spans a sector boundary as shown in figure 17.8?

6. A superconducting parallel plate inductor with plate dimensions 0.1 m by 0.1 m and spacing 0.01 m is held together by connectors with max-

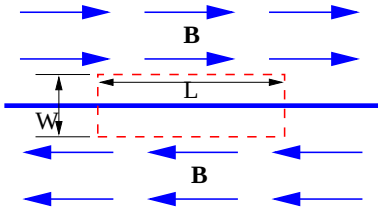

Figure 17.8: Magnetic field at a solar wind sector boundary.

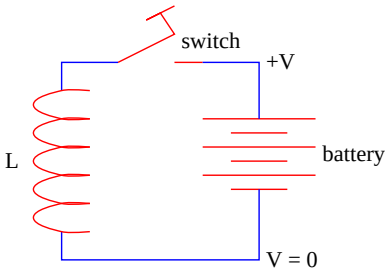

Figure 17.9: Battery in parallel with an inductor.

imum breaking strength 500 N and has the input and the output connected by a superconducting wire. A current i is circulating through the inductor.

(a) Is the force between the plates attractive or repulsive?

(b) What is the maximum magnetic field that the inductor can have between the plates without blowing apart? Hint: Find the energy in the magnetic field as a function of plate separation and compute the force between the plates as for the capacitor. The magnetic flux through the inductor remains constant as the plates move in this case, which means that the current can change.

(c) What is the current corresponding to the above maximum field?

7. Use Kirchhoff's laws to compute the net resistance of

(a) resistors in parallel, and

(b) resistors in series,

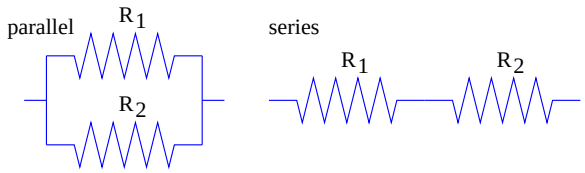

Figure 17.10: Resistors in parallel and in series.

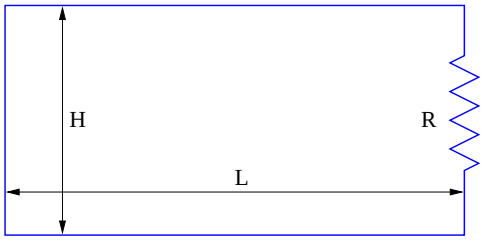

Figure 17.11: Circuit consisting of a shorted resistor.

as shown in figure 17.10. Hint: In the first case the voltage drop across the resistors is the same, in the second, the current through the resistors is the same. Recall that Ohm's law relates the current through a device to the voltage drop across it. (If you already know the answers, derive them, don't just write them down.)

8. Try to explain in physical terms why doubling the length of a resistor doubles its resistance, while doubling its cross-sectional area halves its resistance. Use this argument to justify equation (17.17).

9. Describe qualitatively what happens when

   (a) the switch is closed in the circuit in figure 17.9, and

   (b) when it is abruptly opened.

   The battery produces a voltage difference V, but also may be thought of as having a small internal resistance R.

10. Given the circuit shown in figure 17.11:

    (a) What do Kirchhoff's laws tell you about $\Delta\phi$ across the resistor?

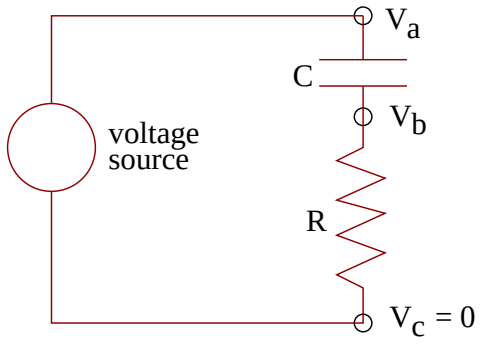

Figure 17.12: Simple RC circuit.

(b) Suppose a time-varying magnetic field $B = B_0 \sin(\omega t)$ is applied normal to the circuit loop, where B_0 is a constant. What is the (time-dependent) voltage drop $\Delta\phi$ across the resistor in this situation?

(c) Given the above $\Delta\phi$, what is the current through the resistor as a function of time?

You may ignore the effect of the current in creating an additional magnetic field.

11. In the circuit shown in figure 17.12, the voltage source is switched on at time $t = 0$, at which the voltage V_A goes from zero to some constant positive value. The capacitor initially has no charge.

(a) Just after the source is switched on what is the voltage V_B? Hint: Can the potential difference across the capacitor change instantaneously with the resistor in the circuit? Explain why or why not.

(b) After a very long time, what is the voltage V_B?

(c) Make a qualitative sketch of V_B as a function of time.

Chapter 18

Measuring the Very Small

To begin our study of matter we discuss experiments in the late 19th and early 20th centuries which led to proof of the existence of atoms and their constituents. We then introduce a fundamental idea about scattering of waves using the diffraction of light by small particles as a prototype. The famous Geiger-Marsden experiment which lead 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.[1]

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 which 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*.

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 dif-

[1]Many of the ideas in this chapter were taken from Aitchison, I. J. R., and A. J. G. Hey, 1989: Gauge theories in particle physics. IOP Publishing, 571 pp.

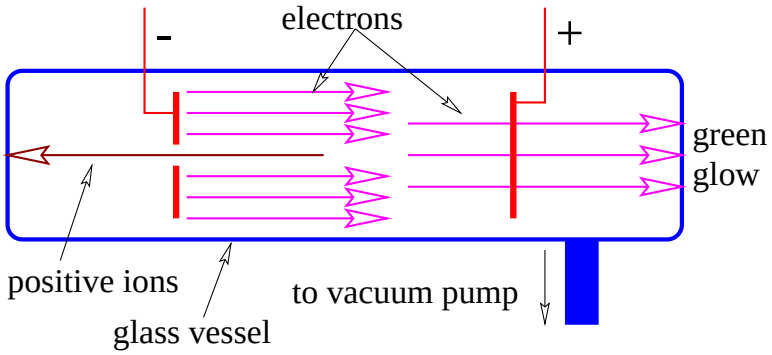

Figure 18.1: Crookes tube, the original particle accelerator. When potentials are applied to the plates as shown, electrons are emitted by the left electrode and accelerated to the right, some of which pass through holes in the right electrode. Positive ions, which are atoms missing one or more electrons, are created by collisions between electrons and residual gas atoms. These are accelerated to the left.

ference $\Delta\phi$ 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 which 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

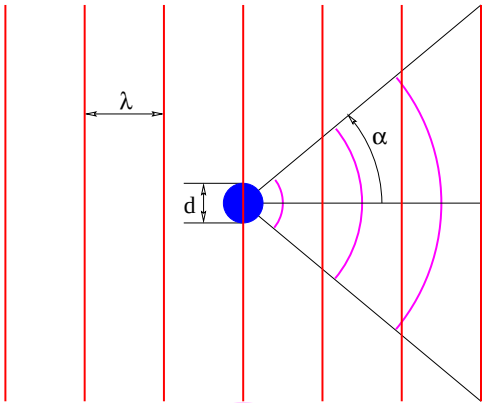

Figure 18.2: Scattering of an incident plane wave by a water droplet. The opening half-angle of the scattered wave is $\alpha \approx \lambda/(2d)$.

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

$$\alpha \approx \lambda/(2d). \tag{18.1}$$

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 section 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

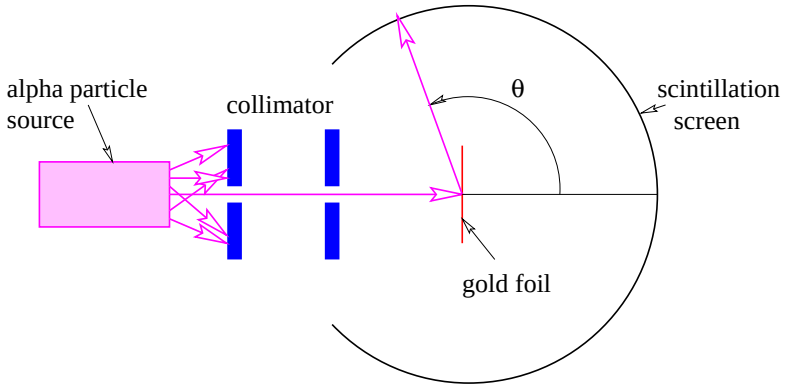

Figure 18.3: Schematic of Geiger-Marsden experiment. The radioactive source produces alpha particles which are collimated into a beam and directed at a gold foil. The alpha particles scatter off the foil and are detected by a flash of light when they hit the scintillation screen.

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. 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 $= 8\times10^{-13}$ 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.

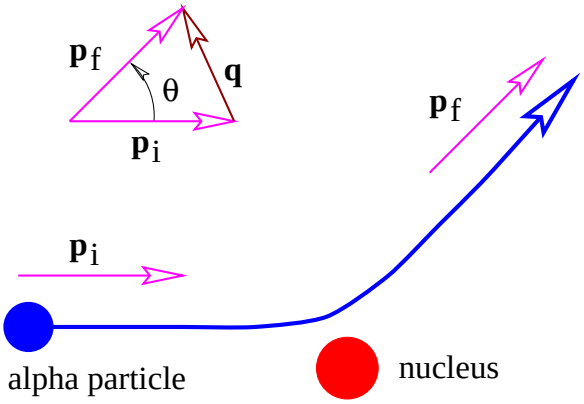

Figure 18.4: Illustration of alpha particle trajectory in Rutherford's model of the atom. The momentum transfer $\mathbf{q}$ from the nucleus to the alpha particle is equal to the change in the alpha particle's momentum.

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 $\approx 0.01°$.

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 $\mathbf{q}$ 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. The maximum momentum transfer for an incident alpha particle with momentum $\mathbf{p}$ is $2|\mathbf{p}|$, 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 Accelerator Experiments

18.4.1 Early Cosmic Ray Results

Earlier we indicated that particles interacted with each other via the exchange of a virtual intermediary particle which 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 which 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 plus an antineutrino, while the neutral pion decays into two gamma rays, or high energy photons. The antineutrino which 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 which 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 which 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 wave-

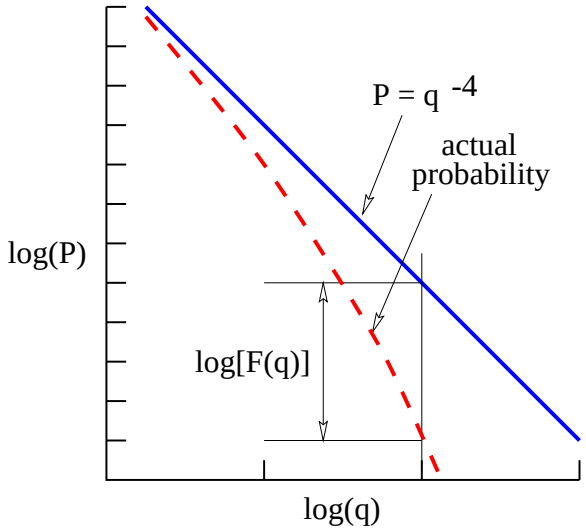

Figure 18.5: Schematic illustration of Robert Hofstadter's results for scattering of electrons off of atomic nuclei. The solid line shows the relative probability (in log-log coordinates) of elastic scattering as a function of the momentum transfer. The dashed curve illustrates the observed probability distribution. The difference between the curves is the logarithm of the form factor, $F(q)$.

lengths 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. 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 which 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 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

The 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 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. This leads to a subsequent set of reactions which produce the profusion of particles seen in the left panel of this figure. Feynman called the particles which make up the proton *partons*. However, we now know that they are actually *quarks*, spin 1/2 particles with fractional electronic charge which are thought to be the fundamental building blocks of matter, and *gluons*, the massless spin 1 intermediary particles which 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. 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.

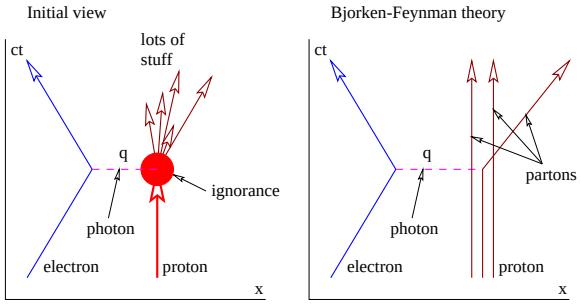

Figure 18.6: Deep inelastic scattering of a high energy electron by a proton occurs when the momentum transfer q is large and many particles are produced. According to the Bjorken-Feynman theory of this process, the proton consists of a number of partons flying in "loose formation". A sufficiently energetic photon, i. e., with large momentum transfer q, kicks out just one of these partons, leaving the others undisturbed.

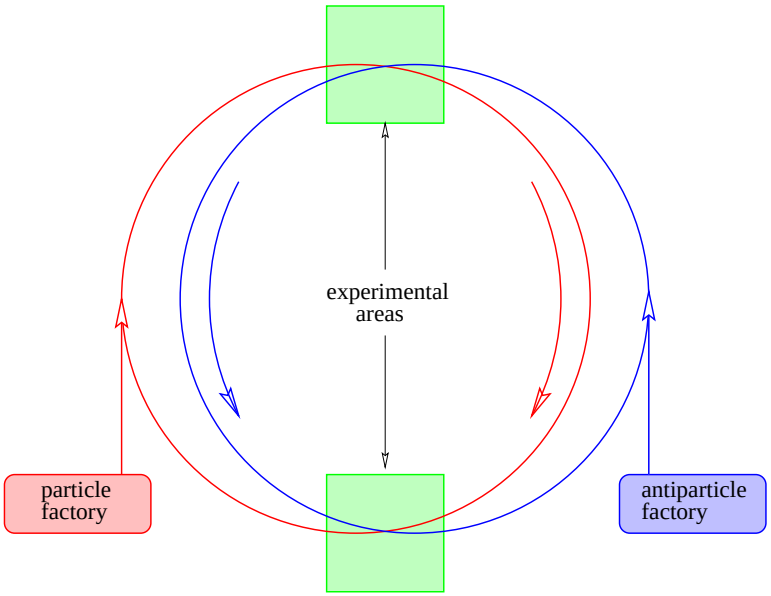

Figure 18.7: Schematic model of a particle-antiparticle collider. The particles and antiparticles are injected into the storage rings shown and are made to go in a circle by magnetic fields. The beams cross at two points and apparatus is set up around these points to observe the products of collisions.

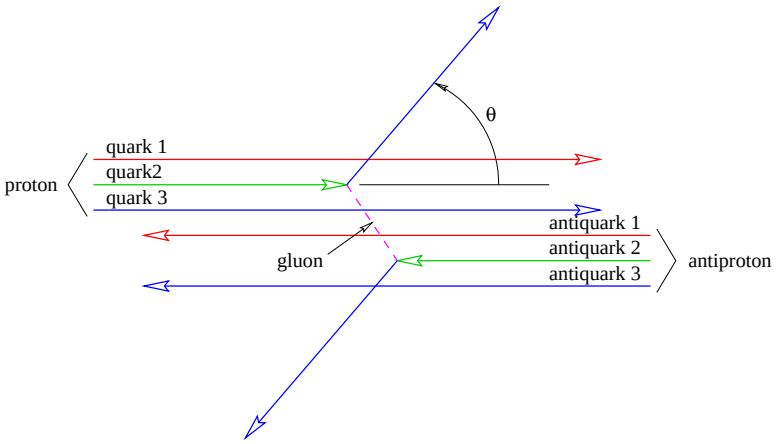

Figure 18.8: Illustration of what happens in a high energy collision between a proton and an antiproton according to the Bjorken-Feynman parton model.

An alternate type of collider has two storage rings which 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 which 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 col-

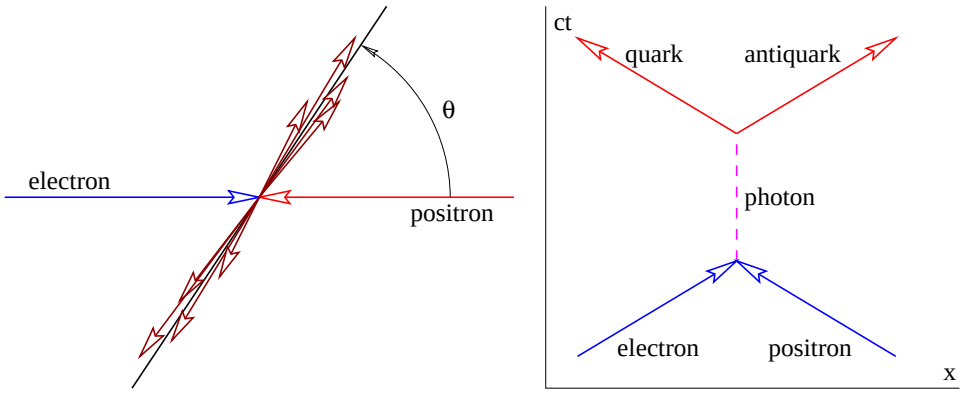

Figure 18.9: Two jet events resulting from the annihilation of high energy electrons and positrons. The virtual photon decays into a quark-antiquark pair which in turn generate the oppositely pointing jets of particles.

lision between highly energetic protons and antiprotons should operate as shown in figure 18.8. The actual collision is between individual partons. 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. 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 which 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.

18.6 Problems

1. If possible, observe the moon through a thin cloud layer and estimate the angular size of the disk of scattered light around the moon. From this, estimate the size of the particles doing the scattering.

2. Which particle can be used to investigate smaller scales, a proton or an electron, at

   (a) the same *velocity*, and

   (b) at the same *kinetic energy*? (Work non-relativistically in both cases.)

   (c) Now consider ultra-relativistic protons and electrons with the same *total energy*. Is there a significant difference between their ability to investigate very small scales?

3. Electron microscope:

   (a) What kinetic energy (in electron volts) must electrons in an electron microscope have to match the resolution of an optical microscope? (The resolutions match when the wavelengths of the electrons and the light are the same.)

   (b) If the electrons have kinetic energy 50 KeV, how much better resolution does the electron microscope have than the best optical microscope?

Hint: Use the non-relativistic kinetic energy and check whether this assumption is valid in retrospect.

4. Integrated circuits are made by a system in which the circuit pattern is engraved on a silicon wafer using a photochemical process working with an optical imaging device which projects the circuit image on the wafer.

   (a) Assuming visible light is used, estimate the size of the smallest feature which could be produced on the silicon by this system.

   (b) Do the same for 1 KeV X-rays.

   Hint: Recall that the smallest feature resolvable by a wave is of order the wavelength.

5. The rest energy of two colliding particles is just c^2 times the mass of the single particle created by the colliding particles sticking together.

   (a) Compute the rest energy (in GeV) of a particle resulting from a 100 GeV energy proton colliding with a stationary proton.

   (b) Compute the rest energy of the particle resulting from two 50 GeV protons colliding head-on.

   Hint: These calculations are relativistic, since the rest energy of the proton is about 0.9 GeV.

6. Relativistic charged particle in magnetic field: Assume that a relativistic particle of mass m and charge e is moving in a circle under the influence of the magnetic field $\mathbf{B} = (0, 0, -B)$. The position of the particle as a function of time is given by $\mathbf{x} = [R\cos(\omega t), R\sin(\omega t), 0]$.

   (a) Compute the (vector) velocity of the particle and show that its speed is $v = \omega R$.

   (b) Compute the (relativistic) momentum (again in vector form) of the particle using the above results.

   (c) Compute the magnetic force $\mathbf{F}$ on the particle.

   (d) Using the relativistic version of Newton's second law, $\mathbf{F} = d\mathbf{p}/dt$, determine how the rotational frequency ω depends on the speed of the particle, the magnetic field B, and the particle's charge and mass. Examine particularly the limits where $v \ll c$ and $v \approx c$.

(e) Eliminate ω between the above result and the speed formula to get an equation for the radius R of the circle. Show that this takes the particularly simple form $R = p/(eB)$ when written in terms of the magnitude of the momentum $p = mv\gamma$.

7. A 30 GeV electron is scattered by a virtual photon through an angle of 60° without changing its energy.

   (a) Compute its momentum vector before and after the scattering.

   (b) Compute the momentum transfer to the electron by the photon in the scattering event.

   (c) Compute the wavelength of the virtual photon.

   (d) What is the virtual photon's energy?

   (e) What is the virtual photon's mass?

8. Find α, β, and γ such that $\hbar^\alpha c^\beta G^\gamma$ has the units of length. (G is the universal gravitational constant.) Compute the numerical value of this length, which is called the *Planck length*. Compare this value to the resolution available today in the highest energy accelerators.

Chapter 19

Atoms

In this section 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 $P(x_1, x_2)$ 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 *anti-symmetric* 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:

$$\psi(x_1, x_2) = \psi_1(x_1)\psi_2(x_2) - \psi_1(x_2)\psi_2(x_1) \quad \text{(non-interacting fermions)} \tag{19.2}$$

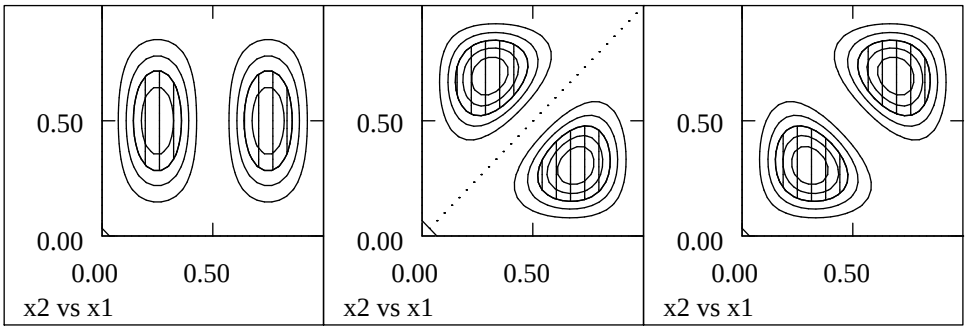

Figure 19.1: Joint probability distributions for two particles, one in the ground state and one in the first excited state of a one-dimensional box. Left panel: non-identical particles. Middle panel: identical fermions. Right panel: identical bosons. The curved lines are contours of constant probability. The vertical hatching shows where the probability is large.

for fermions and

$$\psi(x_1, x_2) = \psi_1(x_1)\psi_2(x_2) + \psi_1(x_2)\psi_2(x_1) \quad \text{(non-interacting bosons)} \quad (19.3)$$

for bosons.

Figure 19.1 shows the joint probability distribution for two particles in different energy states in an infinite square well: $P(x_1, x_2) = |\psi(x_1, x_2)|^2$. 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 con-

densation 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 times 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 $U = -2K$. 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* which 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 which 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 times the square of the fine structure constant.

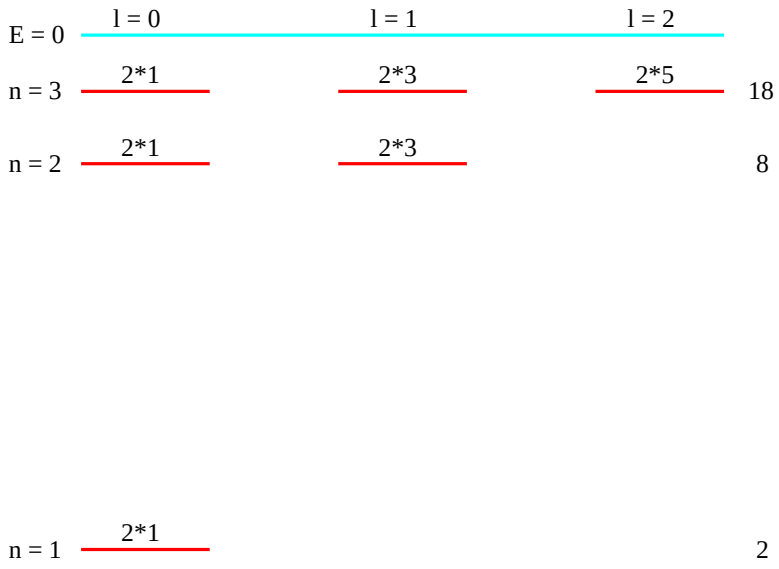

Figure 19.2: Energy levels of the hydrogen atom. Energy increases upward and angular momentum increases to the right. The numbers above each level indicate the spin orientation times the orbital orientation degeneracy for each level. The numbers at the right show the total degeneracy for each value of n. Only the first three values of n are shown.

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

$$E_n = -\frac{E_B}{n^2} = -\frac{\alpha^2 mc^2}{2n^2}, \quad n = 1, 2, 3, \ldots \quad \text{(hydrogen energy levels)}, \quad (19.11)$$

where n is called the *principal quantum number* of the hydrogen atom.

19.3 The Periodic Table of the Elements

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

$$N_n = 2\sum_{l=0}^{n-1}(2l+1) \tag{19.12}$$

states. In particular, for $n = 1, 2, 3, \ldots$, we have $N_n = 2, 8, 18, \ldots$. This is summarized in figure 19.2.

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 times 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 which 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 plus 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. 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

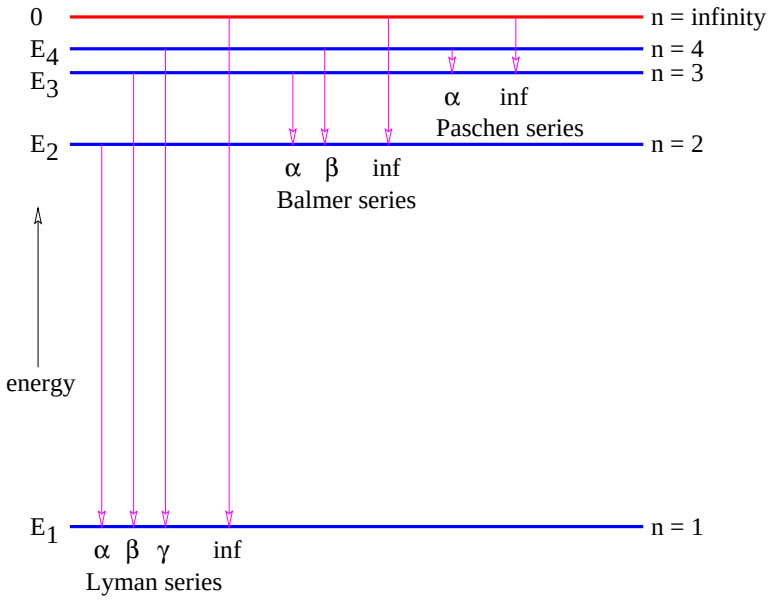

Figure 19.3: Spectral lines from transitions between electron energy states in hydrogen.

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 which 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.

19.5 Problems

1. The wave function for three non-identical particles in a box of unit length with one particle in the ground state, the second in the first excited state, and the third in the second excited state is

$$\psi(x_1, x_2, x_3) = \sin(\pi x_1)\sin(2\pi x_2)\sin(3\pi x_3).$$

   (a) From this write down the wave function for three identical bosons in the above mentioned states.

   (b) Do the same for three identical fermions.

   Hint: In each case there are six terms corresponding to the six permutations of x_1, x_2, and x_3. Exchanging any two particles leaves ψ unchanged for bosons but changes the sign for fermions.

2. Two identical particles with equal energies collide nearly head-on, so that they are both deflected through an angle θ, as shown in figure 19.4. A physicist calculates the amplitude ψ as a function of θ for this deflection to take place (using very advanced theory!), resulting in the solid curve shown in figure 19.4. However, measurements show that the actual amplitude as a function of θ (not probability!) is given by the dashed curve.

   (a) What did the physicist forget to take into account? Explain.

   (b) Are the particles fermions or bosons? Explain.

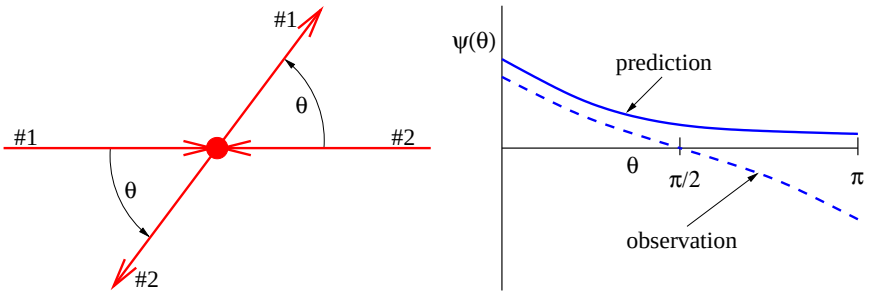

Figure 19.4: Incorrectly calculated and observed scattering amplitude for a collision between two identical particles.

Hint: If the outgoing particles (but not the incoming particles) are interchanged, how does the apparent deflection angle change?

3. Following the analysis made for the hydrogen atom, compute the "Bohr radius" and the ground state binding energy for an "atom" consisting of Z protons in the nucleus and one electron.

4. Upper and lower bounds on the binding energy of the last (outermost) electron in the sodium atom may be obtained by assuming (a) that the other electrons have no effect, or (b) that the other electrons neutralize all but one proton in the nucleus. Compute the binding energy of the last electron in sodium in these two limits. (The actual binding energy of the last electron in sodium is 5.139 eV.)

5. A uranium atom ($Z = 92$) has all its electrons stripped off except the first one.

   (a) What is the first electron's binding energy in electron volts?

   (b) What is the ground state radius of the electron orbit in this case?

6. The energy levels of a particle in a box are given by $E_n = E_0 n^2 = E_0, 4E_0, 9E_0, \ldots$ where E_0 is the ground state energy for the particle. Find the lowest possible total energy of a group of particles, expressed as a multiple of E_0, for the following particles in the box:

   (a) 5 identical spin 0 particles.

(b) 5 identical spin 1/2 particles.

(c) 5 identical spin 1 particles.

(d) 5 identical spin 3/2 particles.

7. A charged particle in a 1-D box has energy levels at $E_n = E_0 n^2 = E_0,\ 4E_0,\ 9E_0,\ 16E_0,\ 25E_0, \ldots$, where E_0 is the ground state energy of the particle. If the particle can absorb a photon with any of the energies $5E_0,\ 12E_0,\ 21E_0, \ldots$, what can you infer about the initial energy of the particle? Explain.

8. The X-rays in your dentist's office are produced when an energetic beam of free electrons knocks the most tightly bound electrons ($n = 1$) completely out of the target atoms. Electrons from the next level up ($n = 2$) then drop into the $n = 1$ level.

   (a) Estimate the energy in electron volts of the resulting photons for a copper target ($Z = 29$). Hint: For the inner electrons, you may ignore the effects of the other electrons to reasonable accuracy.

   (b) What minimum energy must the electron beam have in this case?

9. What is the shortest ultraviolet wavelength usable in astronomy? Hint: UV photons more energetic than the binding energy of the electron in hydrogen are strongly absorbed by this gas.

10. In the naive periodic table model, the first three closed shells occur for $Z = 2, 10, 28$. However, the first three noble gases have $Z = 2, 10, 18$. Explain why this is so.

Chapter 20

The Standard Model

In this section 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. However, with the standard model we have reached the frontier of what is known with any degree of certainty.

20.1 Quarks and Leptons

The standard model of quarks 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 which do not interact via the strong force. The electron, muon, and the electron and muon neutrinos are examples.
- *Hadrons* are particles which 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 which 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* which 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. 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 which 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

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.

leptons can only be created or destroyed in particle-antiparticle pairs.[1] 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 particles have an electric charge equal to $\pm e$ 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*.

[1] Actually, lepton conservation is even more restrictive, with conversion between electrons, muons, and tau particles being apparently forbidden. However, recent work shows that electron, muon, and tau neutrinos convert into each other on slow time scales. We also know from this work that neutrinos have small, but non-zero mass. The implications of these results are still being explored by the physics community.

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}
positive pion	+1	140	0	$u\overline{d}$	2.6×10^{-8}
negative pion	−1	140	0	$\overline{u}d$	2.6×10^{-8}
neutral pion	0	135	0	$u\overline{u}$ - $d\overline{d}$	8.7×10^{-17}
positive rho	+1	770	1	$u\overline{d}$	4×10^{-24}
positive kaon	+1	494	0	$u\overline{s}$	1.24×10^{-8}
neutral kaon	0	498	0	$d\overline{s}$	8.6×10^{-11}
J/psi	0	3097	1	$c\overline{c}$	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 times the half life) is in seconds. The composition is presented in terms of the flavors of quarks and antiquarks which make up the particle.

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. 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 $u\overline{d}$ system, while the positive pion is the ground state of

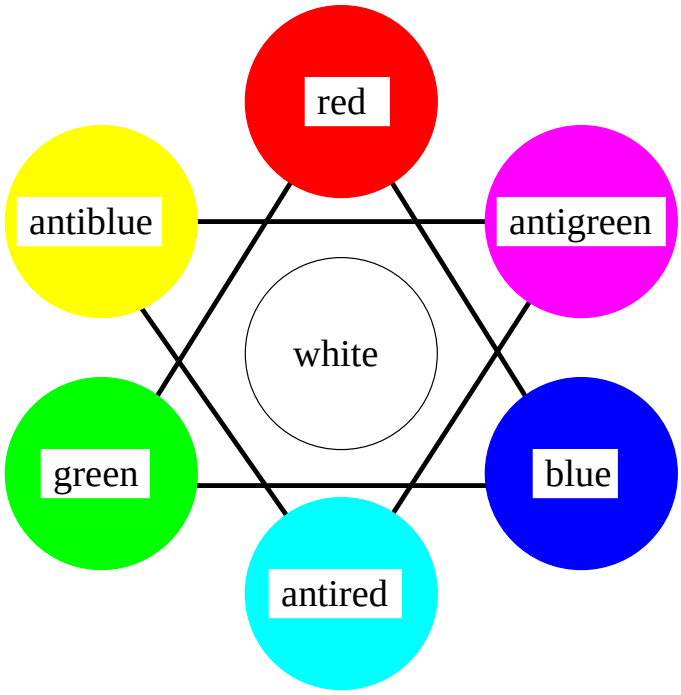

Figure 20.1: The color table of quantum chromodynamics. Quarks have colors red, green, and blue, while antiquarks have colors antired (cyan), antigreen (magenta), and antiblue (yellow). The combination of a quark and its corresponding antiquark is colorless or white, as is the combination of three quarks (or antiquarks) of three different colors.

this system.

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. Antiquarks 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 elec-

tromagnetic 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.

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 which 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

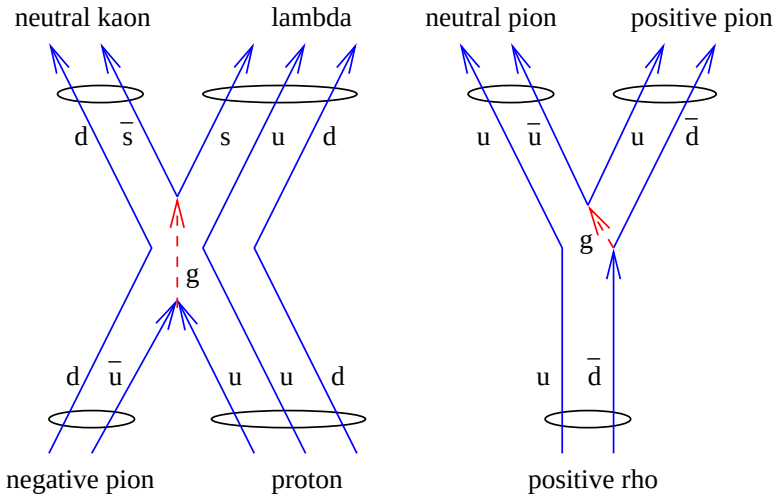

Figure 20.2: Some sample strong interactions illustrated in terms of gluon emission and absorption. The process on the left shows the reaction $\pi^- + p \rightarrow K^0 + \Lambda^0$, while the one on the right shows the decay $\rho^+ \rightarrow \pi^+ + \pi^0$. Quarks are labeled with solid lines while gluons are shown by dashed lines.

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. 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.

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).

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.

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 highly massive (several thousand GeV) spin zero boson called the Higgs particle. Due to its large mass, we have not yet determined whether the Higgs particle exists.

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

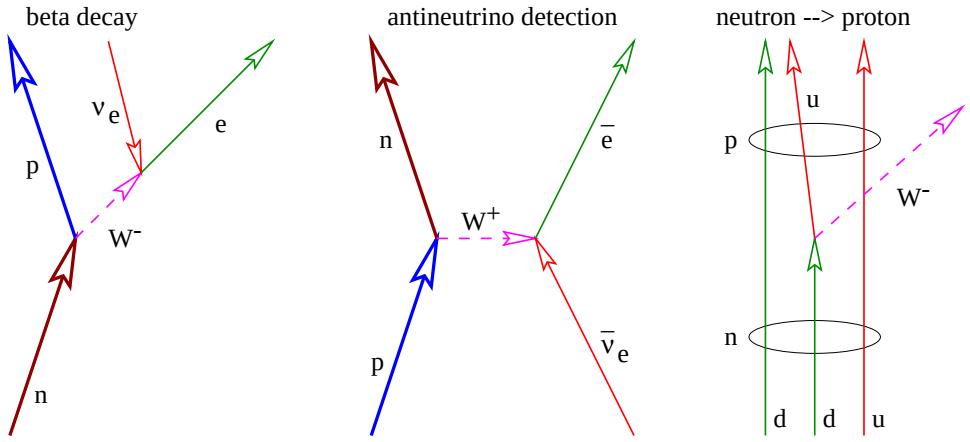

Figure 20.3: Illustration of two weak reactions. The left panel shows beta decay while the middle panel shows how electron antineutrinos can be detected by conversion to a positron. The right panel shows how W^- emission works according to the quark model, resulting in the conversion of a down quark to an up quark and the resulting transformation of a neutron into a proton.

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. 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.

Generation	Leptons	Quarks
1	electron electron neutrino	down up
2	muon mu neutrino	strange charm
3	tau tau neutrino	bottom top

Table 20.4: Generations of leptons and quarks. Members of each generation tend to fit together.

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?

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

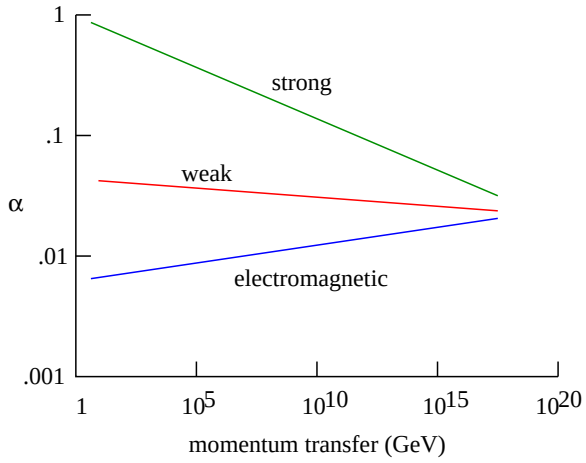

Figure 20.4: Speculated behavior of the dependence of the coupling constant α on momentum transfer for each of the forces. Extrapolation based on current measurements suggests that these constants come together to a common value at very high momentum transfer.

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 which 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.

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^{30} 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 which 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.

20.5 Problems

1. Verify that the quark model predicts the correct electric charge for the proton, the neutron, and all the pions. Also check to see if the spin angular momentum of each of these particles is consistent with its quark composition.

2. Draw a picture of how the negative pion decays into a muon and a mu antineutrino in terms of the quark model of the pion and our ideas about the weak interaction.

3. Draw a picture of how the muon decays into a mu neutrino, an electron, and an electron antineutrino in terms of our ideas about the weak interaction.

4. A mu antineutrino hits a proton, turning it into a neutron.

   (a) What other particle must be emitted from this reaction?

   (b) Could you use this result to distinguish between electron and mu antineutrinos?

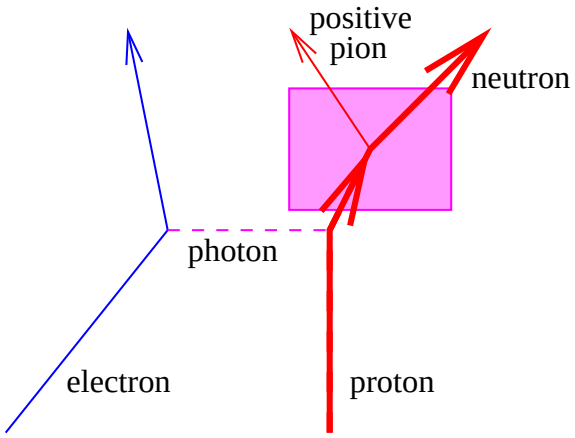

Figure 20.5: An example of inelastic electron-proton scattering.

(c) What minimum total energy in the center of momentum frame would you expect of the mu antineutrino for this reaction to be possible? Note that in this reference frame the kinetic energy of the initial proton will be nearly the same as that of the final neutron.

5. Suppose that the electron had a rest energy of $M = 500$ MeV rather than ≈ 0.5 MeV. Describe as best you can the many ways in which this would change the world and universe in which we live.

6. In the reaction shown in figure 20.5, specify what actually happens at the vertex in the shaded region in terms of the quark model of hadrons.

7. A solar neutrino detector in South Dakota consists of a huge tank of cleaning fluid, which has a large concentration of chlorine-37 ($Z = 17, A = 37$).

   (a) Will an electron neutrino more likely interact with a proton or a neutron in the chlorine-37 nucleus?

   (b) If this interaction occurs, what will the final products be?

   Note: $Z = 16$ is sulfur and $Z = 18$ is argon.

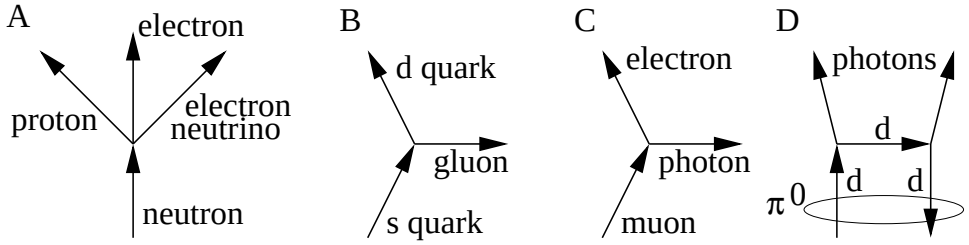

Figure 20.6: Reactions which may or may not be allowed.

8. An electron collides with an antimuon, resulting in the apparent disappearance of both particles. This seems to indicate that energy is not conserved.

   (a) What do you, the Sherlock Holmes of particle physics, suggest actually happened?

   (b) Is this likely to be a very common event? Why or why not?

9. A Λ particle consists of 3 quarks with flavors u, d, s. A possible decay mode is $\Lambda \to p + \pi^-$.

   (a) Is the Λ a fermion or a boson? Explain.

   (b) Draw a Feynman diagram showing how the above decay can happen at the quark level.

   (c) Is the above decay a strong or a weak process?

   Reminder: $p = u, u, d$; $\pi^- = \overline{u}, d$.

10. For each of the reactions shown in figure 20.6, determine whether it is allowed or not. If not, state what is wrong.

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 which are electrically neutral. The ways in which atoms form molecules are therefore complex and resistent 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 section, 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. 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, times the binding energy per nucleon pair, divided by 2. 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

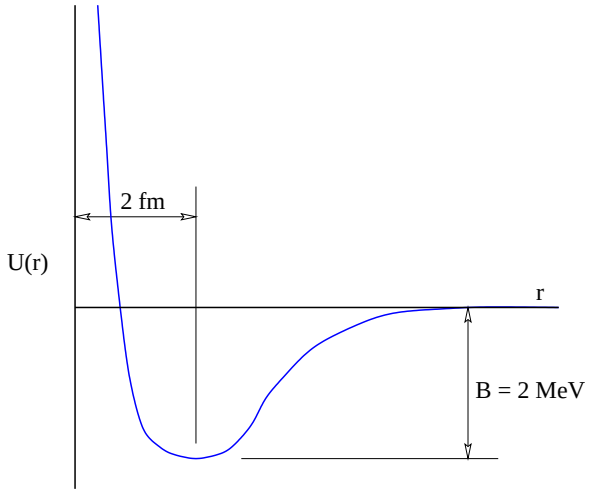

Figure 21.1: Approximate sketch of the strong force potential energy between two nucleons. 1 fm $= 10^{-15}$ m. The binding energy B is the energy required to separate the two nucleons. If the nucleons are bound together, the rest energy of the resulting combination, $M_{combo}c^2$ is less than the sum of the rest energies of the two nucleons, M_1c^2, M_2c^2, by the amount B: $M_{combo}c^2 = M_1c^2 + M_2c^2 - B$.

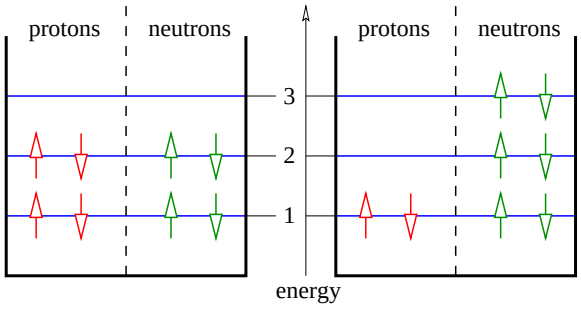

Figure 21.2: Effect of Pauli exclusion principle on two nuclei, each with 8 nucleons. The total energy of the nucleus on the left, which has an equal number of protons and neutrons is $2\times(1+2)+2\times(1+2)=12$. The nucleus on the right has total energy $2\times(1)+2\times(1+2+3)=14$.

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. 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 which 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:

$$B(Z,A)=a_vA-a_sA^{2/3}-a_cZ^2/A^{1/3}-a_a(2Z-A)^2/A \qquad (21.1)$$

where Z is the *atomic number* or the number of protons, N is the number of

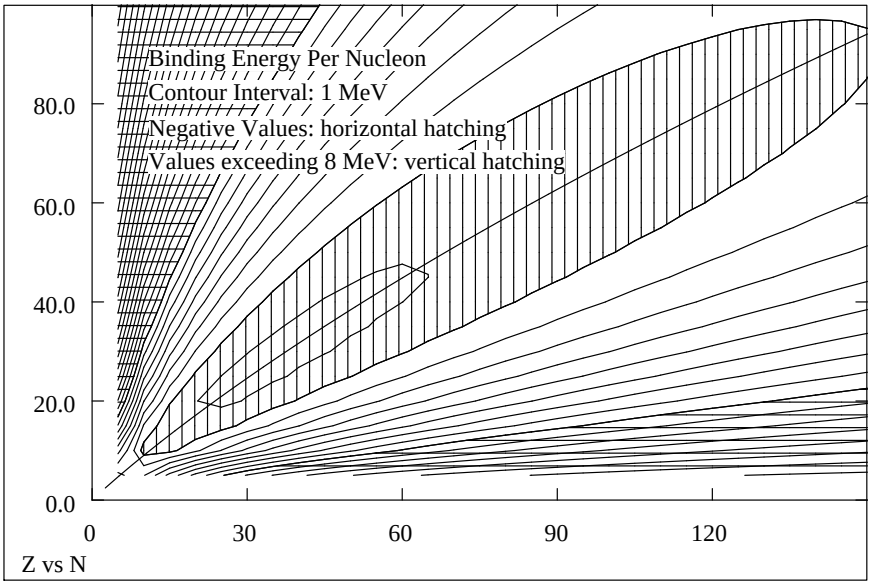

Figure 21.3: Nuclear binding energy per nucleon $B(Z,A)/A$, calculated from equation (21.1). The curved line starting near the origin gives the line of stability for atomic nuclei.

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. 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 spin up-spin down pairs of protons or neutrons 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

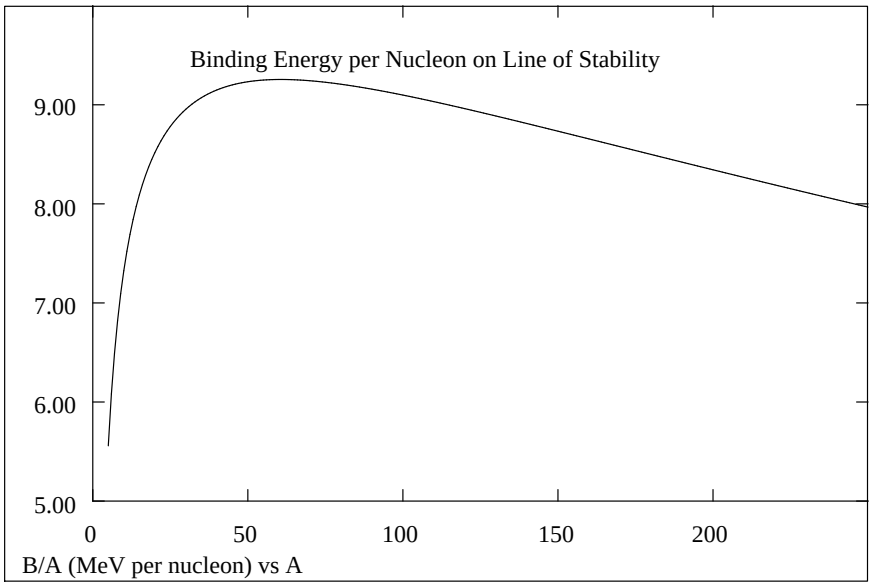

Figure 21.4: Binding energy per nucleon along line of stability according to equations (21.3) and (21.2).

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. 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 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 the last chapter. 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.

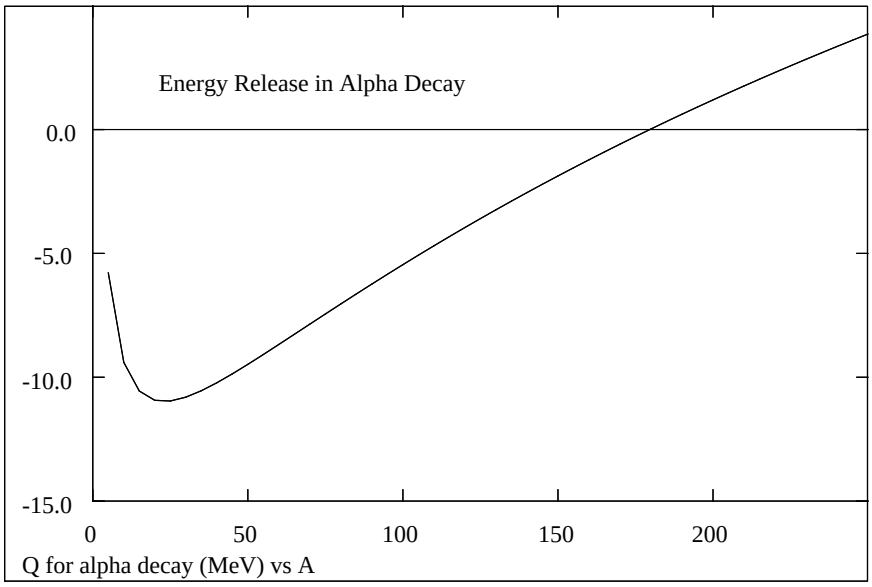

Figure 21.5: Approximate curve for the energy released in alpha decay of a nucleus on the line of stability. Decay is only possible if $Q > 0$.

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 21.6 shows schematically how alpha and beta decay transform atomic nuclei in the N-Z plane. As previously indicated, alpha decay decreases both Z and N by two. Ordinary beta decay (i. e., $n \rightarrow p^+ + e^- + \overline{\nu}_e$)

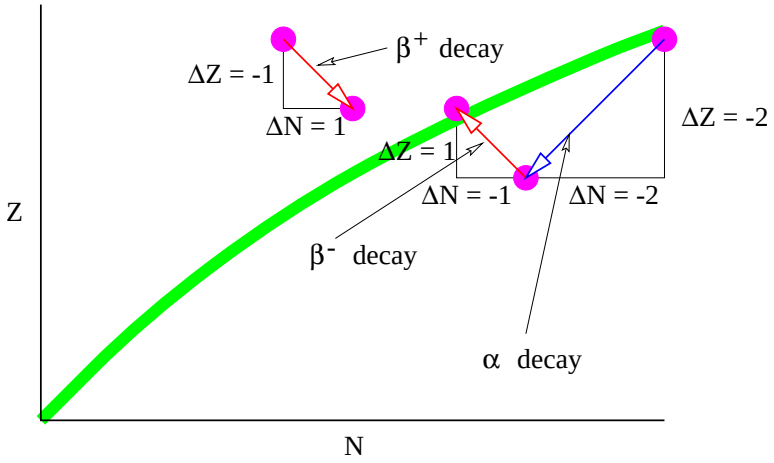

Figure 21.6: Schematic illustration of the paths of nuclear transformations in the N-Z plane due to alpha and beta decay. The thick line represents the line of stability. β^+ is the decay of a proton into a neutron, positron, and electron neutrino, while β^- is the decay of a neutron into a proton, electron, and electron antineutrino.

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 $p^+ \rightarrow n + e^+ + \nu_e$. 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 which 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, namely 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 which 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 which must be surmounted in order to release energy from fusion. 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

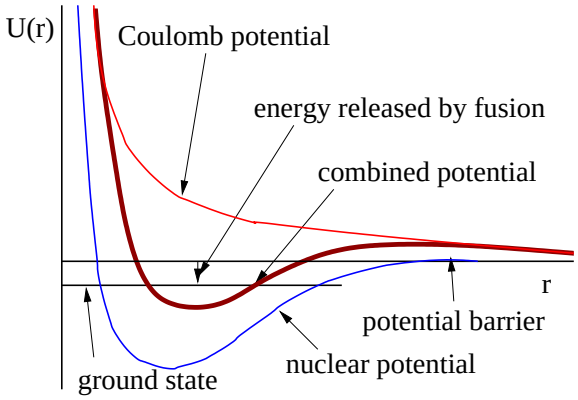

Figure 21.7: Combined nuclear and Coulomb potentials between two light nuclei. The resulting potential barrier repels the two nuclei unless their kinetic energy is very large. However, if the nuclei are able to overcome this barrier, substantial energy is released.

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

Table 21.1: Binding energies of light nuclei.

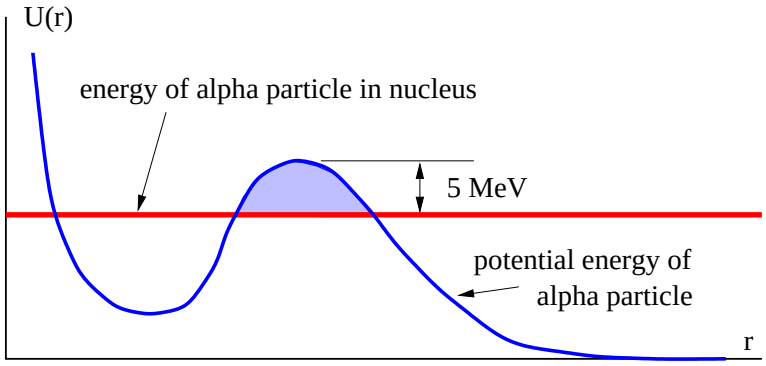

Figure 21.8: Spontaneous fission of a heavy nucleus into a slightly lighter nucleus and an alpha particle occurs when the alpha particle penetrates the potential barrier illustrated by the shading and leaves the nucleus. Other types of spontaneous fission occur in a similar manner. Compared to the case of two light nuclei in figure 21.7, the Coulomb potential is more important here, which makes the resultant force more repulsive.

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.

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:

$$Q \equiv B(Z-Z', A-A') + B(Z', A') - B(Z, A) > 0 \quad \text{(fission possible)}. \quad (21.9)$$

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. 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.

21.5 Problems

1. How would nuclear physics be different if the weak interactions didn't exist?

2. Suppose one started with 10^{20} radioactive atoms with a half life of 2 hr. How many half lives would one have to wait to be reasonably sure that none of the atoms were left?

3. One possible laboratory fusion reaction is $d + d \rightarrow \alpha + Q$ where d represents a deuteron $(Z = 1, A = 2)$, α an alpha particle, and Q the released

energy. Given the binding energies for the deuteron (2.22 MeV) and for the alpha particle (28.30 MeV), find the energy released by this reaction. For the purposes of this problem you may ignore the rest energy of the electrons and their binding energy.

4. Fusion in the sun is a complicated process, but the net effect is the conversion of four protons into an alpha particle, or a helium-4 nucleus. This is what powers the sun.

   (a) How much energy is released for every helium-4 nucleus created?

   (b) How many and what kind of neutrinos or antineutrinos are released for every helium-4 nucleus created?

   (c) At the earth's orbit we get about 1400 J m^{-2} s^{-1} from the sun. How many neutrinos or antineutrinos do we expect to get from the sun per square meter per second from solar fusion?

5. A neutrino has to pass within a distance $D \approx \hbar/(Mc)$ of a quark to have a chance of α_w^2 interact with it, where M is the mass of a W particle and α_w is the weak "fine structure constant".

   (a) What is the area of the circular "target" centered on the quark through which the neutrino has to pass in order to interact with the quark?

   (b) If the quarks are located in the nuclei of water molecules, how many quarks are there per molecule with which the neutrino can interact? Hint: The neutrino can only interact with d quarks in neutrons. Why?

   (c) Imagine a cylindrical water tank of end cross-sectional area A and length L, with neutrinos passing through the tank in a direction parallel to the axis of the cylinder. How many quarks of the right kind are needed in the tank to give a neutrino passing through the tank a 50% probability of interacting with a quark?

   (d) How big must L be in this case? Water has a density of about 1000 kg m^{-3}.

6. Suppose that fission of a uranium-235 nucleus induced by absorbing a slow neutron ultimately results in two equal nuclei plus two neutrons.

(a) How much energy is released for each fissioned uranium-235 nucleus? Hint: The fission products must beta decay until they reach the line of stability on the N-Z plot. Thus, the final state consists of the two free neutrons, two nuclei with the same value of A as the fission products, but with some of the neutrons converted to protons, and the resulting electrons and neutrinos.

(b) How many neutrinos or antineutrinos are released per second by a 100 MW nuclear power plant?

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*,

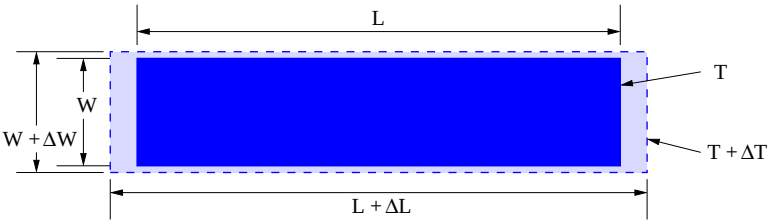

Figure 22.1: Most solid bodies expand by the same fractional amount in all directions when their temperature increases, so that $\Delta L/L = \Delta W/W$. Thus, the ratio $\alpha = \Delta L/(LdT)$ is the same for all objects constructed of the same material, generally over a considerable range of temperature.

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*.

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:

$$\frac{\Delta V}{V} = \beta \Delta T, \tag{22.2}$$

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° C and it boils (at sea level) at 100° C. More precise definition of the Celsius scale depends

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}

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 which is specificially formulated to have a low coefficient of thermal expansion.

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.

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.

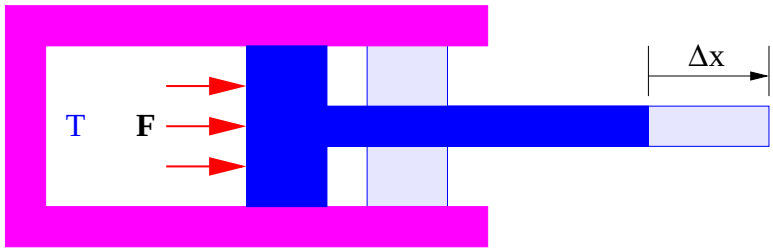

Figure 22.2: Conversion of internal energy of gas in the cylinder to macroscopic energy. The work done by the force of the gas on the piston as it moves outward results in a decrease in temperature of the gas.

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 which are apparently nonconservative. This is in marked contrast to the microscopic world, where forces are either conservative (gravity, electrostatics), or 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 which 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 which 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

Material	C (J kg^{-1} K^{-1})
brass	385
glass	669
ice	2092
steel	448
methyl alcohol	2510
glycerine	2427
water	4184

Table 22.2: Specific heats of common materials.

figure 22.2. As the piston moves out of the cylinder under the force exerted on it by the gas, work is done which 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.

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

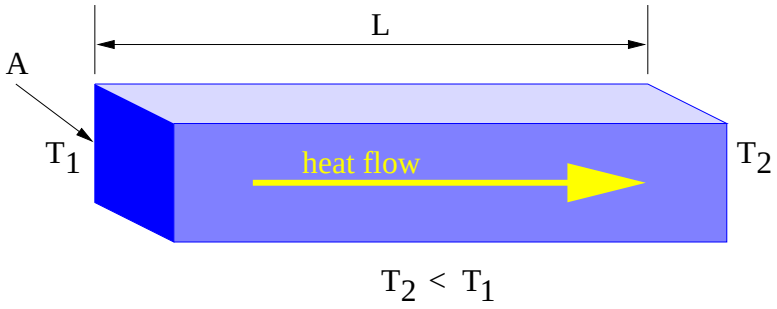

Figure 22.3: Geometry of heat flow problem. Heat flows from higher to lower temperature.

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 which 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

Material	κ (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.

empirically to be proportional to the temperature difference across the body. For a rectangular body it 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:

$$\frac{dQ}{dt} = \frac{\kappa A \Delta T}{L} \tag{22.6}$$

where 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.

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 $\omega_{thermal}$ which is a function only of the temperature T of the body which 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*

$$J_E = \varepsilon \sigma T^4 \quad \text{(Stefan's law)} \tag{22.8}$$

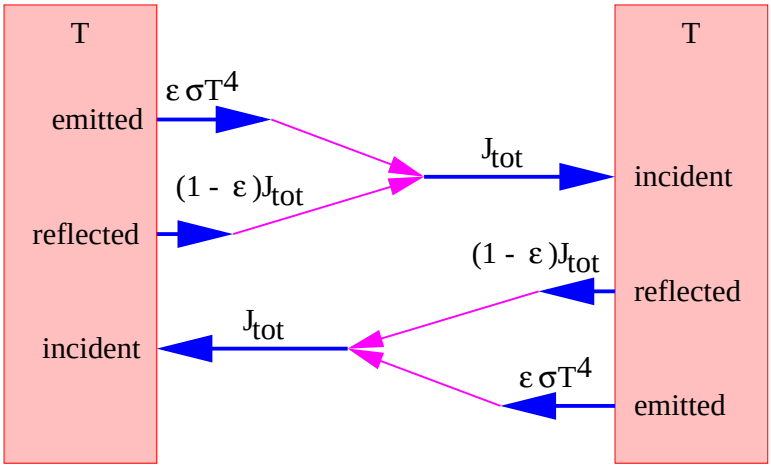

Figure 22.4: Two surfaces facing each other, each with emissivity ε and temperature T.

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 which 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. 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 which 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.[1] 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.) The frictional force is proportional to the normal force N pressing the objects together:

$$F_k = \mu_k N \quad \text{(kinetic friction)}. \tag{22.13}$$

[1] An empirical law is one which we cannot justify in terms of the fundamental principles of physics, but which is observed to be true in a wide variety of situations.

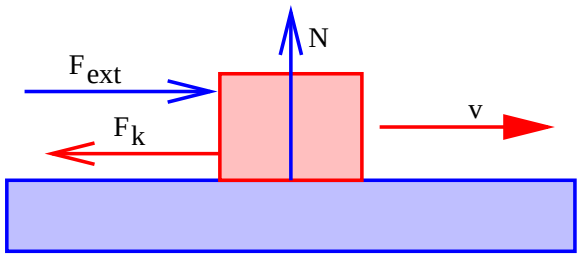

Figure 22.5: The kinetic frictional force F_k is exerted on the upper body by the stationary lower body. The upper body is moving with velocity v and is pressed together with the lower body by a normal force N. It may also be acted upon by an additional non-normal external force F_{ext}.

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

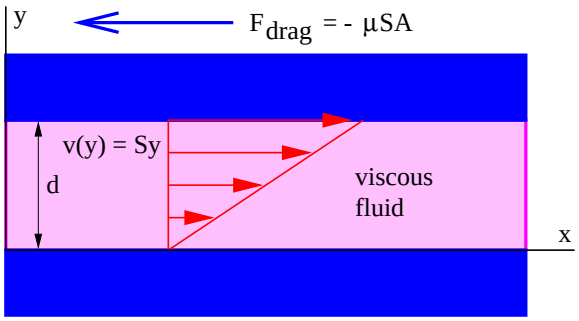

Figure 22.6: Two solid plates separated by a distance d, the gap being filled by a viscous fluid. The lower plate is stationary and the upper plane is moving to the right at speed $v_p = v(d) = Sd$. The fluid is sheared, with the fluid moving according to $v(y) = Sy$. The fluid velocity matches that 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.

story: 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° C to 0.0100 Poise at 20° C to 0.0028 Poise at 100° 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° 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 which 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.

22.4 Problems

1. The George Washington bridge, which spans the Hudson River between New York and New Jersey, is 4760 feet long and is made out of steel. How much does it expand in length between winter and summer? (Pick reasonable winter and summer temperatures.)

2. A volume coefficient of expansion β can be defined for solids as well as liquids. Show that $\beta = 3\alpha$ in this case, where α is the linear coefficient of expansion. Hint: Imagine a cube which increases the length of a side by a fractional amount $\alpha\Delta T \ll 1$ when the temperature increases by ΔT. Compute the fractional change in the volume of the cube.

3. Equal masses of brass and glass are put in the same insulating container, the brass initially at 300 K, the glass at 350 K. Assuming that the interior of the container has negligible heat capacity, what temperature does the material in the container reach after coming to equilibrium?

4. The gravitational potential energy of water going over Niagara Falls (60 m high) is converted to kinetic energy in the fall and then dissipated at the bottom. How much warmer does the water get as a result?

5. A normal-sized house has concrete walls and roof 0.1 m thick. About how much does it cost per month to heat the house electrically if electricity costs \$0.10 per kilowatt-hour? Estimate the wall and roof areas of a typical house and typical inside-outside temperature differences in winter.

6. Compute the thermal frequency $\omega_{thermal}$ and the power per unit area emitted by a surface with emissivity $\epsilon = 1$ for

   (a) $T = 3$ K (cosmic background temperature),

   (b) $T = 300$ K (earth's temperature),

   (c) $T = 6000$ K (sun's surface),

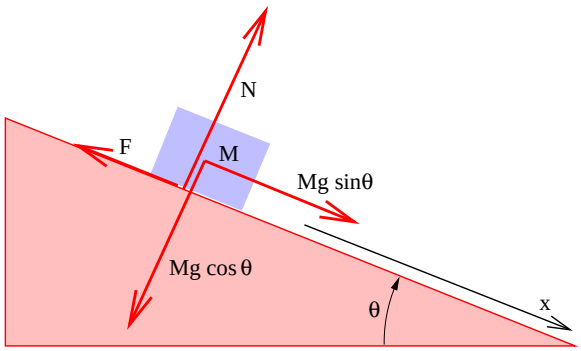

Figure 22.7: Mass M subject to gravity, friction (F), and a normal force (N) on a ramp tilted at an angle θ with respect to the horizontal.

(d) and $T = 2 \times 10^7$ K (sun's interior).

7. Derive an equation for the light pressure (force per unit area) acting on the walls of a box whose interior is at temperature T. Assume for simplicity that all photons being emitted and absorbed by a wall move in a direction normal to the wall. Compute this pressure for the interior of the sun. Hint: Recall that a photon with energy E has momentum E/c, and that both emitted and absorbed photons transfer momentum to the wall.

8. Imagine two plates each at temperature T as in figure 22.4, except that the left plate has emissivity ϵ_L and the right plate has emissivity ϵ_R. Show that the radiative energy flux incident on each plate from the other is still σT^4.

9. Two parallel plates facing each other, one at temperature T_1, the other at temperature T_2, each have emissivity $\epsilon = 1$. Assuming that $T_1 = 200$ K and $T_2 = 300$ K, compute the net radiative transfer of energy per unit area per unit time from plate 2 to plate 1.

10. Imagine a mass sliding down a ramp subject to frictional and normal forces as shown in figure 22.7. If the coefficient of kinetic friction is μ_k, determine the acceleration down the ramp.

11. Suppose the mass in the previous problem has been given a push so

that it is sliding *up* the ramp. Determine its acceleration *down* the ramp.

12. If the coefficient of static friction is μ_s, compute the maximum angle for which the mass in figure 22.7 will remain stationary.

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 based on classical mechanics which Gibbs developed. 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, brick A with internal energy between E and $E + \Delta E$, and brick B with 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 which 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, 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 which 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

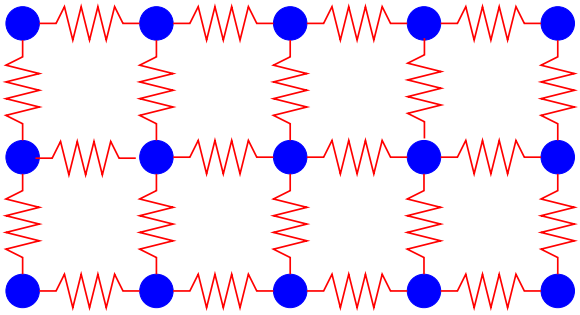

Figure 23.1: "Inner-spring mattress" model of the atoms in a solid body. Interatomic forces act like miniature springs connecting the atoms. As a result the whole system oscillates like a bunch of harmonic oscillators.

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 which 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). 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

$$E_n = (n + 1/2)\hbar\omega_0 \equiv (n + 1/2)E_0, \quad n = 0, 1, 2, \ldots. \tag{23.1}$$

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

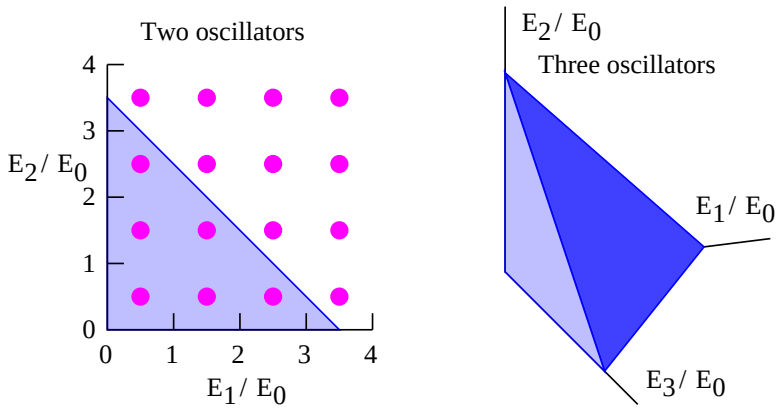

Figure 23.2: Diagrams for counting states of systems of two (left panel) and three (right panel) harmonic oscillators with the same classical oscillation frequency.

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, $\mathcal{N}$, 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. 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 $1\cdot 2\cdot 3\ldots N$ 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:

$$\Delta\mathcal{N} = \mathcal{N}(E+\Delta E) - \mathcal{N}(E) = \frac{\mathcal{N}(E+\Delta E)-\mathcal{N}(E)}{\Delta E}\Delta E \approx \frac{\partial \mathcal{N}}{\partial E}\Delta E. \tag{23.6}$$

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

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.

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 $\Delta E = E_0$. 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

$$\begin{aligned} \ln(\Delta\mathcal{N}) &= \ln\left(\frac{(E/E_0)^{N-1}}{(N-1)!}\frac{\Delta E}{E_0}\right) \\ &\approx \ln\left(\frac{(E/E_0)^{N}}{N!}\frac{\Delta E}{E_0}\right) \\ &= N\ln(E/E_0) - \ln(N!) + \ln(\Delta E/E_0). \end{aligned} \tag{23.8}$$

A useful mathematical result for large N is the *Stirling approximation*[1]:

$$\ln(N!) \approx N\ln(N) - N \quad \text{(Stirling approximation)}. \tag{23.9}$$

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

$$\ln(\Delta\mathcal{N}) = N\left[\ln\left(\frac{E}{NE_0}\right) + 1\right] + \ln\left(\frac{\Delta E}{E_0}\right) \quad (N \text{ oscillators}). \tag{23.10}$$

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

[1] To derive the Stirling approximation note that $\ln(N!) = \ln(1)+\ln(2)+\ldots+\ln(N)$. This sum can be approximated by the integral $\int_1^N \ln(x)dx = N\ln(N) - N + 1 \approx N\ln(N) - N$.

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 which 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. It's 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 which 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.

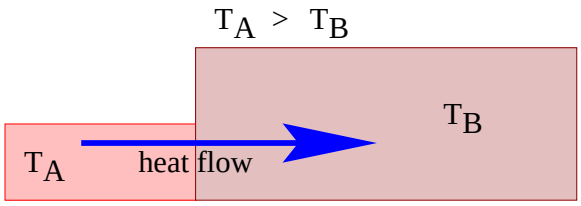

Figure 23.3: Two bricks in thermal contact, one at temperature T_A, the other at temperature T_B. If $T_A > T_B$, internal energy flows from brick A to brick B.

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:

$$S = S_A + S_B. \tag{23.17}$$

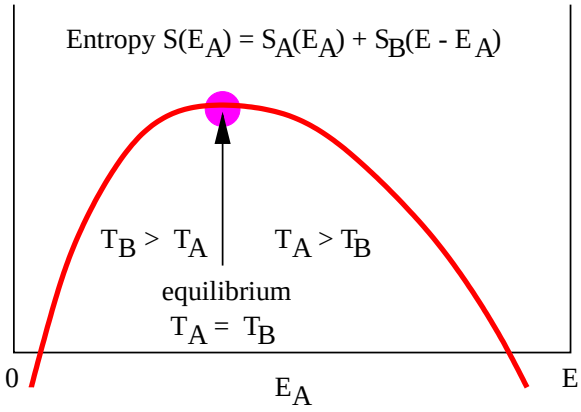

Figure 23.4: Total entropy of two systems for fixed total energy $E = E_A + E_B$ as a function of E_A, the energy of system A.

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 which typically looks something like figure 23.4. 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

$$\frac{\partial S}{\partial E_A} = \frac{\partial S_A}{\partial E_A} + \frac{\partial S_B}{\partial E_A} = \frac{\partial S_A}{\partial E_A} + \frac{\partial S_B}{\partial E_B}\frac{dE_B}{dE_A} = \frac{\partial S_A}{\partial E_A} - \frac{\partial S_B}{\partial E_B} = 0. \qquad (23.18)$$

(The partial derivatives indicate that parameters besides the energy are held constant while taking the derivative of entropy.) Thus,

$$\frac{\partial S_A}{\partial E_A} = \frac{\partial S_B}{\partial E_B} \quad \text{(equilibrium condition)} \qquad (23.19)$$

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 which 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 which 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. (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 which 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. 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

$$\Delta S = \Delta S_1 + \Delta S_2 = \Delta Q\left(\frac{1}{T_2} - \frac{1}{T_1}\right). \tag{23.26}$$

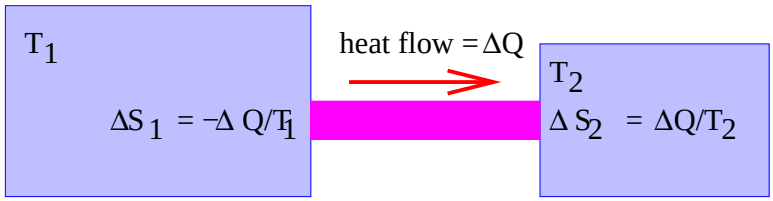

Figure 23.5: The two regions at temperatures T_1 and $T_2 < T_1$ are connected by a thin bar which conducts heat slowly from the first to the second region. For heat ΔQ transferred, the entropy of region 1 decreases according to $\Delta S_1 = -\Delta Q/T_1$, while the entropy of region 2 increases by $\Delta S_2 = \Delta Q/T_2$.

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.

23.7 Problems

1. Compute an approximate value for $N^N/N!$ using the Stirling approximation. (This gives the essence of $\Delta\mathcal{N}$ for N harmonic oscillators.) From this show that $\ln(\Delta\mathcal{N}) \propto N$.

2. States of a pair of distinguishable dice (i. e., one is red, the other is green):

(a) List all of the possible states of a pair of dice, i. e., all the possible combinations of face-up numbers.

(b) Given that each of the dice has six faces, does the total number of states equal that given by equation (23.16)?

3. There are $N!/[M!(N-M)!]$ ways of arranging N pennies with M heads up. Verify this for 2, 3, and 4 pennies. (Note that by definition $0! = 1$.)

4. Suppose we have N pennies on a shaking table which bounces the pennies around, flipping them over at random. The pennies are weighted so that the gravitational potential energy of a penny is zero with tails up and U with heads up.

(a) If M heads are up, what is the total energy E?

(b) How many "states", $\Delta\mathcal{N}$, are there with M heads up? Hint: Compute this directly from the statement of the previous problem, not by computing $d\mathcal{N}/dE$ as we did for N harmonic oscillators. What does the energy interval ΔE correspond to in this case?

(c) Compute the entropy of the system as a function of E and N. Hint: You will need to use the Stirling approximation to do this part.

(d) Compute the temperature as a function of E and N.

(e) Invert the temperature equation derived in the previous step to obtain E as a function of T and N. To understand this result, approximate it in the low and high limits, i. e., $k_BT/U \ll 1$ and $k_BT/U \gg 1$. Try to think of an explanation of the behavior of the pennies in these limits which would make sense to (say) an 8th grade student. In particular, how is the intensity of the shaking of the table related to the "temperature"? Hint: In the low temperature limit note that $\exp(U/k_BT) \gg 1$, while in the high temperature limit $\exp(U/k_BT) \approx 1$.

5. Suppose that two systems, A and B, have available states $\Delta\mathcal{N}_A = E_A^X$ and $\Delta\mathcal{N}_B = E_B^Y$, where $E = E_A + E_B = 2$. Compute and plot $\Delta\mathcal{N} = \Delta\mathcal{N}_A\Delta\mathcal{N}_B$ as a function of E_A over the range $0 < E_A < 2$ for:

(a) $X = Y = 1$;

(b) $X = Y = 5$;

(c) $X = Y = 25$;

(d) $X = 2$; $Y = 8$ — explain the position of the peak in terms of the values of X and Y.

How does the width of the peak change as X and Y get larger? Explain the consequences of this result for the reliability of the second law of thermodynamics as a function of the number of particles in each system.

6. Suppose we have a system of mass M in which $k_B T = AE^{1/2}$, where T is the temperature, E is the internal energy, k_B is Boltzmann's constant, and A is a constant.

   (a) Derive a formula for the entropy of the system as a function of internal energy. Hint: Remember the thermodynamic definition of temperature.

   (b) Compute the specific heat of this system.

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 which 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 section. 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 which 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

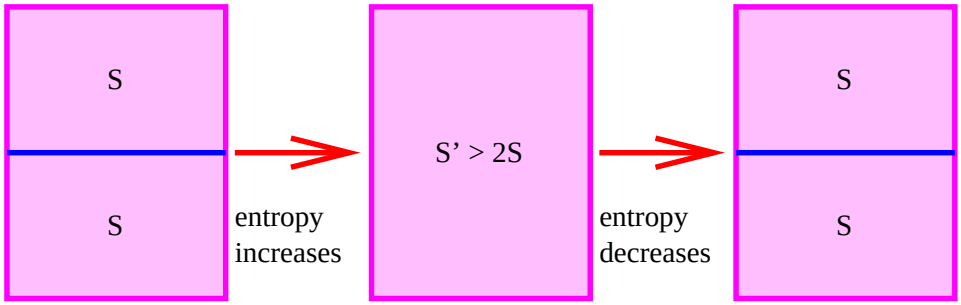

Figure 24.1: Consequence of the incorrect classical calculation of entropy of an ideal gas by Gibbs. Two parts of a container separated by a divider each contain the same type of gas at the same temperature and pressure. The total entropy is $2S$ where S is the classically calculated entropy of each 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 $2S$.

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. Imagine a container of gas of a certain type, temperature, and pressure which 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 which 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 which 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

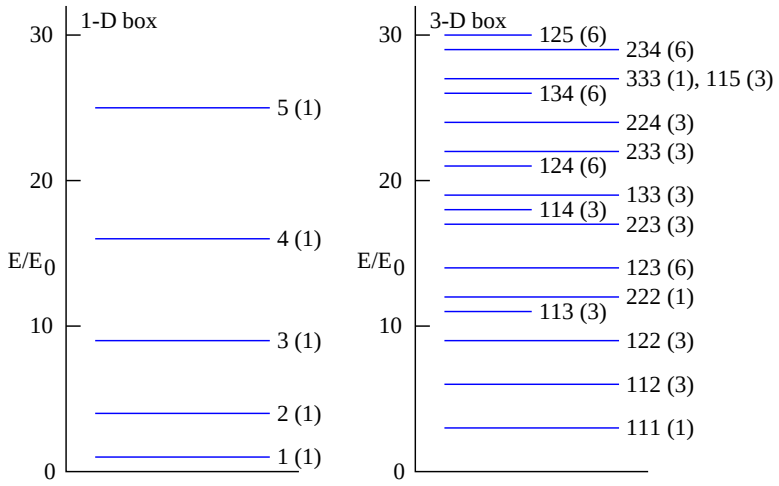

Figure 24.2: Energy levels for a non-relativistic particle in a one-dimensional and a three-dimensional box, each of side length a. The value E_0 is the ground state energy of the one-dimensional particle in a box of length a. The numbers to the right of the levels respectively give the values of l for 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).

components of the kinetic momentum are therefore $p_x = \hbar k_x$, 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. 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

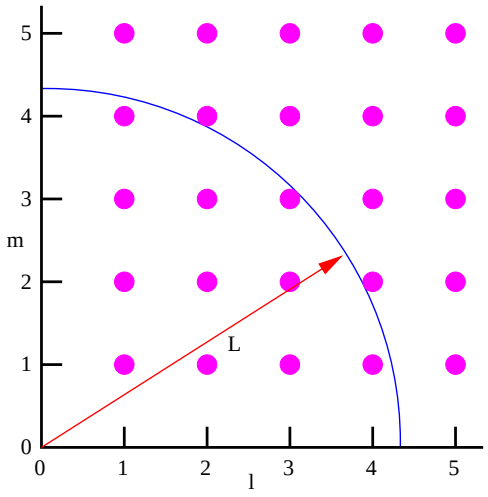

Figure 24.3: The states of a particle in a two-dimensional box. The dots indicate particular states associated with allowed values of the x and y direction quantum numbers, l and m. The pie-shaped segment bounded by the arc of radius L and the l and m axes encompasses all of the states with $l^2 + m^2 \leq L^2$.

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 which have that 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. 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 times some as yet unknown function of N, $F(N)$:

$$\mathcal{N} = F(N)\left(\frac{E}{E_0}\right)^{3N/2} \quad (N \text{ particles in 3-D box}). \tag{24.6}$$

(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 times volume to the two-thirds power, so we write this combination of constants in terms of constant reference values of E and V:

$$\pi^2\hbar^2/(2M) = E_{ref}V_{ref}^{2/3}. \tag{24.8}$$

Given 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.

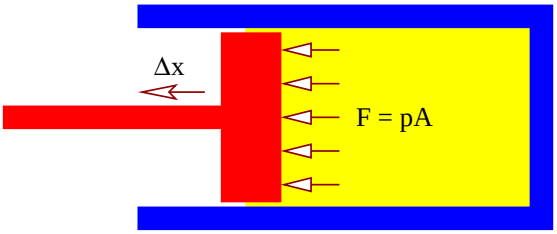

Figure 24.4: Gas in a cylinder with a movable piston. The force F exerted by the gas on the piston is the area A of the face of the piston times the pressure p.

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 which 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 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 24.4. 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 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

$$p \equiv -\frac{\partial E}{\partial V}. \tag{24.15}$$

The assumption that $\Delta Q = 0$ normally implies that the entropy S does not change as well. Thus, in the evaluation of $\partial E/\partial V$, the entropy is held constant. This turns out to be a non-trivial assumption and the conditions under which it is true are discussed in the next section. For now we shall assume that this assumption is valid.

We can determine the pressure for an ideal gas by solving equation (24.12) for E and taking the V derivative. This equation may be written in compact form as

$$E = \frac{N^{5/3}E_{ref}V_{ref}^{2/3}\exp[2S/(3Nk_B)]}{V^{2/3}} = \frac{B}{V^{2/3}} \quad \text{(ideal gas)} \tag{24.16}$$

where B contains all references to entropy, number of particles, etc. For isentropic (i. e., constant entropy) processes, we therefore have

$$EV^{2/3} = const \quad \text{(constant entropy processes)}. \tag{24.17}$$

The pressure can be rewritten as

$$p = -\frac{\partial E}{\partial V} = \frac{2}{3}\frac{B}{V^{5/3}} = \frac{2E}{3V} \quad \text{(ideal gas)} \tag{24.18}$$

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

$$pV = Nk_BT \quad \text{(ideal gas law)}, \tag{24.19}$$

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 formula for the entropy of an ideal gas (24.12), its temperature (24.13), and the ideal gas law (24.19) 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 vibrational and rotational degrees of freedom which 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) and (24.13) results in equations which apply to diatomic gases.

24.1.6 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 $\Delta E = (3/2)Nk_B\Delta T$. 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

$$C_V = \frac{1}{NM}\frac{3Nk_B}{2} = \frac{3k_B}{2M} \quad \text{(specific heat at const vol)} \tag{24.20}$$

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 (24.19) predicts that $p\Delta V =$

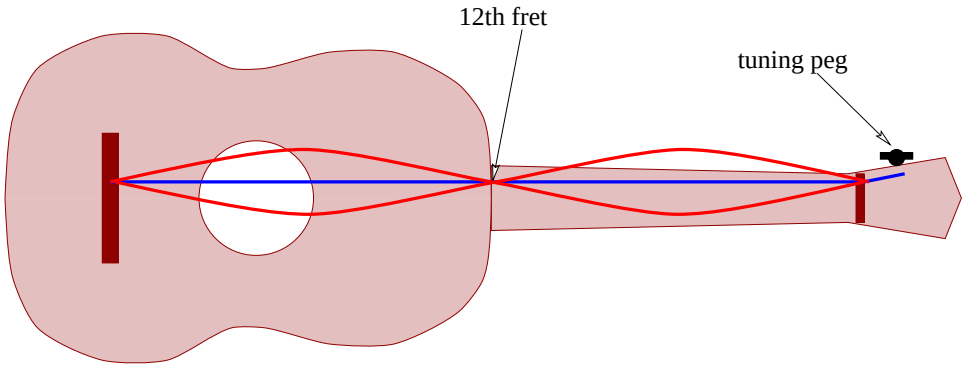

Figure 24.5: Harmonics on a guitar. Plucking a string while a finger rests lightly on the string at the 12th fret results in excitation of the first harmonic on the guitar string. Only one string is shown for clarity.

$Nk_B\Delta T$. Using this and the previous equation for ΔE results in the specific heat of an ideal gas at constant pressure:

$$C_P = \frac{1}{NM}\left(\frac{3Nk_B}{2} + Nk_B\right) = \frac{5k_B}{2M} \quad \text{(specific heat at const pres).} \quad (24.21)$$

24.2 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, 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 24.5, 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.

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

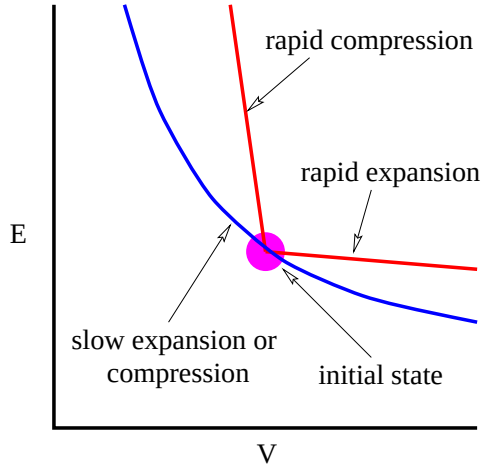

Figure 24.6: The curved line indicates the reversible adiabatic curve $E \propto V^{-2/3}$ for an ideal gas in a box. The two straight line segments indicate what happens in a rapid expansion or compression.

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 which changes the macroscopic condition of a system but which 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 $EV^{2/3} = const.$ Thus, the energy of the gas increases when the volume is decreased and vice versa. This behavior is illustrated in figure 24.6. 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 contraction, 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 24.6.

24.3 Heat Engines

Heat engines typically operate by transferring internal energy (i. e., heat) to and from 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:

$$\Delta S = \frac{\partial S}{\partial E}\Delta E + \frac{\partial S}{\partial V}\Delta V. \tag{24.22}$$

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

$$p\Delta V = \Delta W = T\Delta S - \Delta E \quad \text{(work by ideal gas)} \tag{24.23}$$

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.

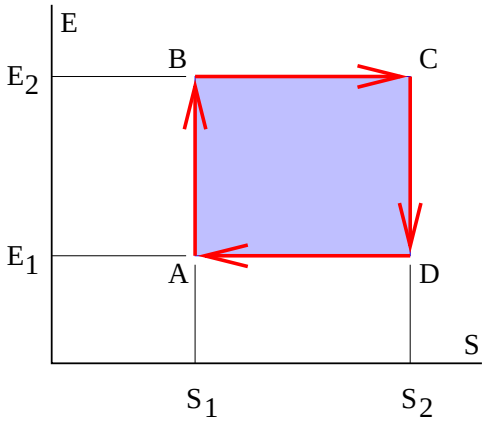

Figure 24.7: Plot of Carnot cycle for an ideal gas in a cylinder. Entropy-energy coordinates are used.

Unfortunately, a piston in a cylinder which can only extract energy during single expansion wouldn't be very useful — it would be like an automobile engine which 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 which results in a net conversion of internal energy to useful work. 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 re-

versible), moving from point B to point C in figure 24.7 in such a way that its internal energy doesn't change. From equation (24.23) 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:

$$\begin{aligned} W &= W_{AB} + W_{BC} + W_{CD} + W_{DA} \\ &= (E_1 - E_2) + T_2(S_2 - S_1) + (E_2 - E_1) + T_1(S_1 - S_2) \\ &= (T_2 - T_1)(S_2 - S_1) \quad \text{(Carnot cycle).} \end{aligned} \tag{24.24}$$

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 which is converted to useful work in this cycle is

$$\epsilon = \frac{W}{Q_2} = \frac{(T_2 - T_1)(S_2 - S_1)}{T_2(S_2 - S_1)} = 1 - \frac{T_1}{T_2} \quad \text{(thermodynamic efficiency).} \tag{24.25}$$

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 ΔQ_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 (24.25) 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 (24.24) and (24.25). 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.

24.4 Perpetual Motion Machines

Perpetual motion machines are devices which 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

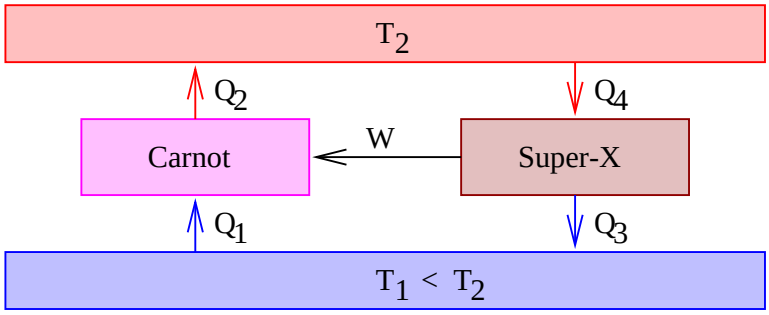

Figure 24.8: Perpetual motion machine of the second kind. The Super-X machine is advertised as having a thermodynamic efficiency greater than a Carnot engine. The output of the Super-X machine runs the Carnot engine backwards as a refrigerator, resulting in net transfer of heat from the lower temperature to the higher temperature reservoir.

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 which is more efficient than the Carnot cycle. In other words, does there exist a heat engine operating between temperatures T_2 and T_1 which 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. 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,

$$W < \epsilon Q_2 \quad \text{(real heat engines)}, \tag{24.26}$$

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

$$Q_2 < W/\epsilon \quad \text{(real refrigerators)}, \tag{24.27}$$

where W is the work done on the refrigerator.

24.5 Problems

1. Following the procedure for a three-dimensional gas, do the following for a two-dimensional gas in a box of area $A = a^2$, where a is the side length of the box.

   (a) Find $\mathcal{N}$ for N particles. Eliminate a in favor of the box area A.

   (b) Compute the entropy for this gas.

   (c) Compute the temperature T, as a function of N and the internal energy E. Invert to obtain the internal energy as a function of N and T.

   (d) Solve the entropy equation for energy and compute the "two-dimensional pressure", $\tau = -\partial E/\partial A$. What units does τ have?

   (e) Find the two-dimensional analog to the ideal gas equation.

   These calculations are relevant to atoms which can move freely around on a surface, but cannot escape it for energetic reasons.

2. Suppose your house has interior volume V. There are a few small air leaks, so that the inside air pressure p always equals the outside air pressure, which is assumed not to change.

   (a) Compute the internal energy of the air in your house.

   (b) Your roommate, trying to impress you with his knowledge of physics, says that he is going to turn up the thermostat to increase the internal energy of the air in the house. Will this work? Explain.

3. It has been proposed to extract useful work from the ocean by exploiting the temperature difference between deep ocean water at $\approx 0°$ C and tropical surface water at $\approx 30°$ C to run a heat engine. What thermodynamic efficiency would this process have?

4. Suppose your house is heated by a Carnot engine working as a refrigerator between an outdoor temperature of 273 K and an indoor temperature of 303 K. (This means you are cooling the outdoors to heat the indoors! Such devices are called *heat pumps*.) If your house loses heat at a rate of 5 kW, how much electrical energy must be used

to power the (perfectly efficient) electrical motor running the Carnot engine? Compare the monthly cost of running this Carnot engine to the cost of direct electric heating, i. e., via a big resistor.

5. Suppose an airplane engine is a heat engine which works between temperatures T_{air} and $T_{air} + \Delta T$, where T_{air} is the air temperature, and where ΔT is fixed. Other things being equal, is this engine more thermodynamically efficient in the summer or winter? Explain.

6. Suppose the spring constant k of a spring varies with temperature, so that $k = CT$, where C is a constant and T is the (Kelvin) temperature. Describe how this spring could be used to construct a heat engine.

7. Suppose a monatomic ideal gas at initial temperature T is allowed to expand very rapidly so that it's new volume is twice its original volume. It is then compressed isentropically (i. e., at constant entropy, which means it is done slowly) back to the original volume. What is its new temperature?

8. An inventor claims to have a refrigerator that extracts 100 W of heat from its interior which is kept at 150 K, rejecting the heat at room temperature (300 K). He claims that the refrigerator only consumes 10 W of externally supplied power. If this device works, does it violate the second law of thermodynamics?

Appendix A

Constants

This appendix lists various useful constants.

A.1 Constants of Nature

Symbol	Value	Meaning
h	6.63×10^{-34} J s	Planck's constant
$\hbar$	1.06×10^{-34} J s	$h/(2\pi)$
c	3×10^{8} m s^{-1}	speed of light
G	6.67×10^{-11} m^3 s^{-2} kg^{-1}	universal gravitational constant
k_B	1.38×10^{-23} J K^{-1}	Boltzmann's constant
σ	5.67×10^{-8} W m^{-2} K^{-4}	Stefan-Boltzmann constant
K	3.67×10^{11} s^{-1} K^{-1}	thermal frequency constant
ϵ_0	8.85×10^{-12} C^2 N^{-1} m^{-2}	permittivity of free space
μ_0	$4\pi \times 10^{-7}$ N s^2 C^{-2}	permeability of free space ($= 1/(\epsilon_0 c^2)$).

A.2 Properties of Stable Particles

Symbol	Value	Meaning
e	1.60×10^{-19} C	fundamental unit of charge
m_e	9.11×10^{-31} kg $= 0.511$ MeV	mass of electron
m_p	1.672648×10^{-27} kg $= 938.280$ MeV	mass of proton
m_n	1.674954×10^{-27} kg $= 939.573$ MeV	mass of neutron

A.3 Properties of Solar System Objects

Symbol	Value	Meaning
M_e	5.98×10^{24} kg	mass of earth
M_m	7.36×10^{22} kg	mass of moon
M_s	1.99×10^{30} kg	mass of sun
R_e	6.37×10^{6} m	radius of earth
R_m	1.74×10^{6} m	radius of moon
R_s	6.96×10^{8} m	radius of sun
D_m	3.82×10^{8} m	earth-moon distance
D_s	1.50×10^{11} m	earth-sun distance
g	9.81 m s^{-2}	earth's surface gravity

A.4 Miscellaneous Conversions

1 lb = 4.448 N
1 ft = 0.3048 m
1 mph = 0.4470 m s^{-1}
1 eV = 1.60×10^{-19} J
1 mol = 6.022×10^{23} molecules
(One mole of carbon-12 atoms has a mass of 12 g.)

Appendix B

GNU Free Documentation License

Version 1.1, March 2000

Preamble

The purpose of this License is to make a manual, textbook, or other written document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software

does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

B.1 Applicability and Definitions

This License applies to any manual or other work that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you".

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (For example, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, whose contents can be viewed and edited directly and straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup has been designed to thwart or

discourage subsequent modification by readers is not Transparent. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML designed for human modification. Opaque formats include PostScript, PDF, proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

B.2 Verbatim Copying

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

B.3 Copying in Quantity

If you publish printed copies of the Document numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts:

Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a publicly-accessible computer-network location containing a complete Transparent copy of the Document, free of added material, which the general network-using public has access to download anonymously at no charge using public-standard network protocols. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

B.4 Modifications

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if

there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

- List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has less than five).
- State on the Title page the name of the publisher of the Modified Version, as the publisher.
- Preserve all the copyright notices of the Document.
- Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- Include an unaltered copy of this License.
- Preserve the section entitled "History", and its title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

- In any section entitled "Acknowledgements" or "Dedications", preserve the section's title, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

- Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

- Delete any section entitled "Endorsements". Such a section may not be included in the Modified Version.

- Do not retitle any existing section as "Endorsements" or to conflict in title with any Invariant Section.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties – for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

B.5 Combining Documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections entitled "History" in the various original documents, forming one section entitled "History"; likewise combine any sections entitled "Acknowledgements", and any sections entitled "Dedications". You must delete all sections entitled "Endorsements."

B.6 Collections of Documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

B.7 Aggregation With Independent Works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution

medium, does not as a whole count as a Modified Version of the Document, provided no compilation copyright is claimed for the compilation. Such a compilation is called an "aggregate", and this License does not apply to the other self-contained works thus compiled with the Document, on account of their being thus compiled, if they are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one quarter of the entire aggregate, the Document's Cover Texts may be placed on covers that surround only the Document within the aggregate. Otherwise they must appear on covers around the whole aggregate.

B.8 Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License provided that you also include the original English version of this License. In case of a disagreement between the translation and the original English version of this License, the original English version will prevail.

B.9 Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

B.10 Future Revisions of This License

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.1 or any later version published by the Free Software Foundation; with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have no Invariant Sections, write "with no Invariant Sections" instead of saying which ones are invariant. If you have no Front-Cover Texts, write "no Front-Cover Texts" instead of "Front-Cover Texts being LIST"; likewise for Back-Cover Texts.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Appendix C

History

- Prehistory: The text was developed to this stage over a period of about 5 years as course notes to Physics 131/132 at New Mexico Tech by David J. Raymond, with input from Alan M. Blyth. The course was taught by Raymond and Blyth and by David J. Westpfahl at New Mexico Tech.
- 14 May 2001: First "copylefted" version of this text.
- 14 May 2003: Numerous small changes and corrections were made.
- 25 June 2004: More small corrections to volume one.
- 14 December 2004: More small corrections to volume one.
- 9 May 2005: More small corrections to volume two.
- 7 April 2006: More small corrections to both volumes.
- 7 July 2006: Finish volume 2 corrections for the coming year.
- 2 July 2008: Finish volume 1 corrections for fall 2008.
- 7 January 2009: Finish volume 2 corrections for spring 2009.