Class: B. Tech (Unit I)

I have taken all course materials for Unit I from Book Introduction to Electrodynamics by David J. Griffith (Prentice- Hall of India Private limited).

Students can download this book form given web address;

Web Address : **https://b-ok.cc/book/5103011/55c730**

All topics of unit I (vector calculus & Electrodynamics) have been taken from **Chapter 1, Chapter 7 & Chapter 8** from said book (https://b-ok.cc/book/5103011/55c730). I am sending pdf file of Chapter 1 Chapter 7 & chapter 8.

Unit-I: Vector Calculus & Electrodynamics: (8 Hours)

Gradient, Divergence, curl and their physical significance. Laplacian in rectangular, cylindrical and spherical coordinates, vector integration, line, surface and volume integrals of vector fields, Gauss-divergence theorem, Stoke's theorem and Green Theorem of vectors. Maxwell equations, electromagnetic wave in free space and its solution in one dimension, energy and momentum of electromagnetic wave, Poynting vector, Problems.

INTRODUCTION TO

ELECTRODYNAMICS

Fourth Edition

$$\nabla^2 V = -\frac{1}{\epsilon_0}\rho$$

$$\nabla^2 \vec{A} = -\mu_0 \vec{J}$$

DAVID J. GRIFFITHS

CHAPTER

1 Vector Analysis

1.1 ■ VECTOR ALGEBRA

1.1.1 ■ Vector Operations

If you walk 4 miles due north and then 3 miles due east (Fig. 1.1), you will have gone a total of 7 miles, but you're *not* 7 miles from where you set out—you're only 5. We need an arithmetic to describe quantities like this, which evidently do not add in the ordinary way. The reason they don't, of course, is that **displacements** (straight line segments going from one point to another) have *direction* as well as *magnitude* (length), and it is essential to take both into account when you combine them. Such objects are called **vectors**: velocity, acceleration, force and momentum are other examples. By contrast, quantities that have magnitude but no direction are called **scalars**: examples include mass, charge, density, and temperature.

I shall use **boldface** (**A**, **B**, and so on) for vectors and ordinary type for scalars. The magnitude of a vector **A** is written $|\mathbf{A}|$ or, more simply, A. In diagrams, vectors are denoted by arrows: the length of the arrow is proportional to the magnitude of the vector, and the arrowhead indicates its direction. *Minus* **A** ($-\mathbf{A}$) is a vector with the same magnitude as **A** but of opposite direction (Fig. 1.2). Note that vectors have magnitude and direction but *not location:* a displacement of 4 miles due north from Washington is represented by the same vector as a displacement 4 miles north from Baltimore (neglecting, of course, the curvature of the earth). On a diagram, therefore, you can slide the arrow around at will, as long as you don't change its length or direction.

We define four vector operations: addition and three kinds of multiplication.

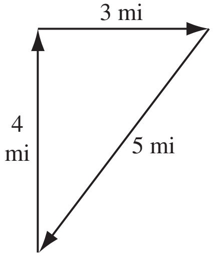


FIGURE 1.1

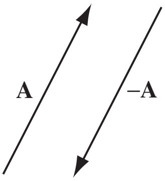


FIGURE 1.2

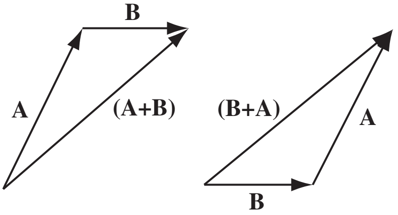


FIGURE 1.3

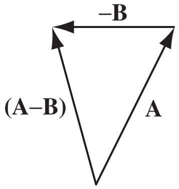


FIGURE 1.4

(i) Addition of two vectors. Place the tail of **B** at the head of **A**; the sum, $\mathbf{A}+\mathbf{B}$, is the vector from the tail of **A** to the head of **B** (Fig. 1.3). (This rule generalizes the obvious procedure for combining two displacements.) Addition is *commutative:*

$$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A};$$

3 miles east followed by 4 miles north gets you to the same place as 4 miles north followed by 3 miles east. Addition is also *associative:*

$$(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C}).$$

To subtract a vector, add its opposite (Fig. 1.4):

$$\mathbf{A}-\mathbf{B}=\mathbf{A}+(-\mathbf{B}).$$

(ii) Multiplication by a scalar. Multiplication of a vector by a positive scalar a multiplies the *magnitude* but leaves the direction unchanged (Fig. 1.5). (If a is negative, the direction is reversed.) Scalar multiplication is *distributive:*

$$a(\mathbf{A}+\mathbf{B})=a\mathbf{A}+a\mathbf{B}.$$

(iii) Dot product of two vectors. The dot product of two vectors is defined by

$$\mathbf{A}\cdot\mathbf{B}\equiv AB\cos\theta, \tag{1.1}$$

where θ is the angle they form when placed tail-to-tail (Fig. 1.6). Note that $\mathbf{A}\cdot\mathbf{B}$ is itself a *scalar* (hence the alternative name **scalar product**). The dot product is *commutative,*

$$\mathbf{A}\cdot\mathbf{B}=\mathbf{B}\cdot\mathbf{A},$$

and *distributive,*

$$\mathbf{A}\cdot(\mathbf{B}+\mathbf{C})=\mathbf{A}\cdot\mathbf{B}+\mathbf{A}\cdot\mathbf{C}. \tag{1.2}$$

Geometrically, $\mathbf{A}\cdot\mathbf{B}$ is the product of A times the projection of **B** along **A** (or the product of B times the projection of **A** along **B**). If the two vectors are parallel, then $\mathbf{A}\cdot\mathbf{B}=AB$. In particular, for any vector **A**,

$$\mathbf{A}\cdot\mathbf{A}=A^2. \tag{1.3}$$

If **A** and **B** are perpendicular, then $\mathbf{A}\cdot\mathbf{B}=0$.

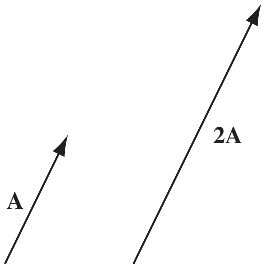


FIGURE 1.5

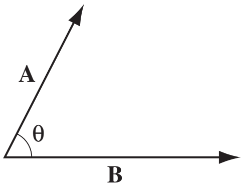


FIGURE 1.6

Example 1.1. Let $\mathbf{C} = \mathbf{A} - \mathbf{B}$ (Fig. 1.7), and calculate the dot product of $\mathbf{C}$ with itself.

Solution

$$\mathbf{C} \cdot \mathbf{C} = (\mathbf{A} - \mathbf{B}) \cdot (\mathbf{A} - \mathbf{B}) = \mathbf{A} \cdot \mathbf{A} - \mathbf{A} \cdot \mathbf{B} - \mathbf{B} \cdot \mathbf{A} + \mathbf{B} \cdot \mathbf{B},$$

or

$$C^2 = A^2 + B^2 - 2AB\cos\theta.$$

This is the **law of cosines**.

(iv) Cross product of two vectors. The cross product of two vectors is defined by

$$\mathbf{A} \times \mathbf{B} \equiv AB \sin\theta \, \hat{\mathbf{n}}, \tag{1.4}$$

where $\hat{\mathbf{n}}$ is a **unit vector** (vector of magnitude 1) pointing perpendicular to the plane of **A** and **B**. (I shall use a hat (ˆ) to denote unit vectors.) Of course, there are *two* directions perpendicular to any plane: "in" and "out." The ambiguity is resolved by the **right-hand rule**: let your fingers point in the direction of the first vector and curl around (via the smaller angle) toward the second; then your thumb indicates the direction of $\hat{\mathbf{n}}$. (In Fig. 1.8, $\mathbf{A} \times \mathbf{B}$ points *into* the page; $\mathbf{B} \times \mathbf{A}$ points *out* of the page.) Note that $\mathbf{A} \times \mathbf{B}$ is itself a *vector* (hence the alternative name **vector product**). The cross product is *distributive,*

$$\mathbf{A} \times (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) + (\mathbf{A} \times \mathbf{C}), \tag{1.5}$$

but *not commutative.* In fact,

$$(\mathbf{B} \times \mathbf{A}) = -(\mathbf{A} \times \mathbf{B}). \tag{1.6}$$

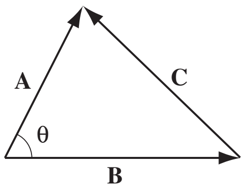


FIGURE 1.7

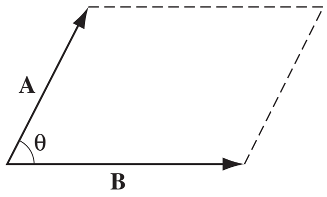


FIGURE 1.8

Geometrically, $|\mathbf{A} \times \mathbf{B}|$ is the area of the parallelogram generated by **A** and **B** (Fig. 1.8). If two vectors are parallel, their cross product is zero. In particular,

$$\mathbf{A} \times \mathbf{A} = \mathbf{0}$$

for any vector **A**. (Here **0** is the **zero vector**, with magnitude 0.)

Problem 1.1 Using the definitions in Eqs. 1.1 and 1.4, and appropriate diagrams, show that the dot product and cross product are distributive,

a) when the three vectors are coplanar;

! b) in the general case.

Problem 1.2 Is the cross product associative?

$$(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} \overset{?}{=} \mathbf{A} \times (\mathbf{B} \times \mathbf{C}).$$

If so, *prove* it; if not, provide a counterexample (the simpler the better).

1.1.2 ■ Vector Algebra: Component Form

In the previous section, I defined the four vector operations (addition, scalar multiplication, dot product, and cross product) in "abstract" form—that is, without reference to any particular coordinate system. In practice, it is often easier to set up Cartesian coordinates x, y, z and work with vector **components**. Let $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ be unit vectors parallel to the x, y, and z axes, respectively (Fig. 1.9(a)). An arbitrary vector **A** can be expanded in terms of these **basis vectors** (Fig. 1.9(b)):

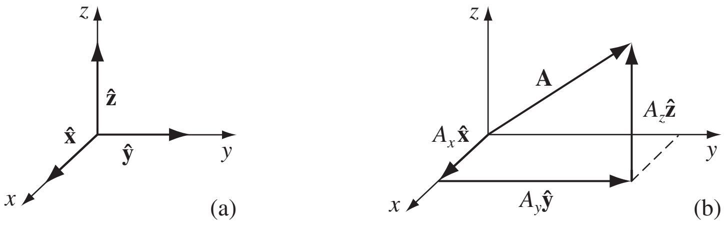


FIGURE 1.9

$$\mathbf{A} = A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}} + A_z\hat{\mathbf{z}}.$$

The numbers A_x, A_y, and A_z, are the "components" of **A**; geometrically, they are the projections of **A** along the three coordinate axes ($A_x = \mathbf{A}\cdot\hat{\mathbf{x}}$, $A_y = \mathbf{A}\cdot\hat{\mathbf{y}}$, $A_z = \mathbf{A}\cdot\hat{\mathbf{z}}$). We can now reformulate each of the four vector operations as a rule for manipulating components:

$$\begin{aligned}\mathbf{A}+\mathbf{B} &= (A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}} + A_z\hat{\mathbf{z}}) + (B_x\hat{\mathbf{x}} + B_y\hat{\mathbf{y}} + B_z\hat{\mathbf{z}}) \\ &= (A_x + B_x)\hat{\mathbf{x}} + (A_y + B_y)\hat{\mathbf{y}} + (A_z + B_z)\hat{\mathbf{z}}.\end{aligned} \tag{1.7}$$

Rule (i): *To add vectors, add like components.*

$$a\mathbf{A} = (aA_x)\hat{\mathbf{x}} + (aA_y)\hat{\mathbf{y}} + (aA_z)\hat{\mathbf{z}}. \tag{1.8}$$

Rule (ii): *To multiply by a scalar, multiply each component.*

Because $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are mutually perpendicular unit vectors,

$$\hat{\mathbf{x}}\cdot\hat{\mathbf{x}} = \hat{\mathbf{y}}\cdot\hat{\mathbf{y}} = \hat{\mathbf{z}}\cdot\hat{\mathbf{z}} = 1; \quad \hat{\mathbf{x}}\cdot\hat{\mathbf{y}} = \hat{\mathbf{x}}\cdot\hat{\mathbf{z}} = \hat{\mathbf{y}}\cdot\hat{\mathbf{z}} = 0. \tag{1.9}$$

Accordingly,

$$\begin{aligned}\mathbf{A}\cdot\mathbf{B} &= (A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}} + A_z\hat{\mathbf{z}})\cdot(B_x\hat{\mathbf{x}} + B_y\hat{\mathbf{y}} + B_z\hat{\mathbf{z}}) \\ &= A_xB_x + A_yB_y + A_zB_z.\end{aligned} \tag{1.10}$$

Rule (iii): *To calculate the dot product, multiply like components, and add.* In particular,

$$\mathbf{A}\cdot\mathbf{A} = A_x^2 + A_y^2 + A_z^2,$$

so

$$A = \sqrt{A_x^2 + A_y^2 + A_z^2}. \tag{1.11}$$

(This is, if you like, the three-dimensional generalization of the Pythagorean theorem.)

Similarly,[1]

$$\begin{aligned}\hat{\mathbf{x}}\times\hat{\mathbf{x}} &= \quad\hat{\mathbf{y}}\times\hat{\mathbf{y}} = \hat{\mathbf{z}}\times\hat{\mathbf{z}} = \mathbf{0}, \\ \hat{\mathbf{x}}\times\hat{\mathbf{y}} &= -\hat{\mathbf{y}}\times\hat{\mathbf{x}} = \hat{\mathbf{z}}, \\ \hat{\mathbf{y}}\times\hat{\mathbf{z}} &= -\hat{\mathbf{z}}\times\hat{\mathbf{y}} = \hat{\mathbf{x}}, \\ \hat{\mathbf{z}}\times\hat{\mathbf{x}} &= -\hat{\mathbf{x}}\times\hat{\mathbf{z}} = \hat{\mathbf{y}}.\end{aligned} \tag{1.12}$$

[1] These signs pertain to a *right-handed* coordinate system (x-axis out of the page, y-axis to the right, z-axis up, or any rotated version thereof). In a *left-handed* system (z-axis down), the signs would be reversed: $\hat{\mathbf{x}}\times\hat{\mathbf{y}} = -\hat{\mathbf{z}}$, and so on. We shall use right-handed systems exclusively.

Therefore,

$$\begin{aligned}\mathbf{A}\times\mathbf{B} &= (A_x\hat{\mathbf{x}}+A_y\hat{\mathbf{y}}+A_z\hat{\mathbf{z}})\times(B_x\hat{\mathbf{x}}+B_y\hat{\mathbf{y}}+B_z\hat{\mathbf{z}}) \\ &= (A_yB_z-A_zB_y)\hat{\mathbf{x}}+(A_zB_x-A_xB_z)\hat{\mathbf{y}}+(A_xB_y-A_yB_x)\hat{\mathbf{z}}.\end{aligned} \tag{1.13}$$

This cumbersome expression can be written more neatly as a determinant:

$$\mathbf{A}\times\mathbf{B} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}. \tag{1.14}$$

Rule (iv): *To calculate the cross product, form the determinant whose first row is* $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$, *whose second row is* **A** *(in component form), and whose third row is* **B**.

Example 1.2. Find the angle between the face diagonals of a cube.

Solution

We might as well use a cube of side 1, and place it as shown in Fig. 1.10, with one corner at the origin. The face diagonals **A** and **B** are

$$\mathbf{A} = 1\,\hat{\mathbf{x}}+0\,\hat{\mathbf{y}}+1\,\hat{\mathbf{z}}; \qquad \mathbf{B} = 0\,\hat{\mathbf{x}}+1\,\hat{\mathbf{y}}+1\,\hat{\mathbf{z}}.$$

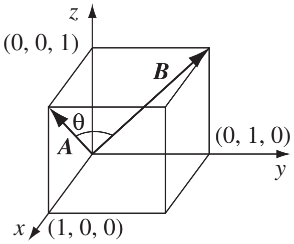


FIGURE 1.10

So, in component form,

$$\mathbf{A}\cdot\mathbf{B} = 1\cdot 0+0\cdot 1+1\cdot 1 = 1.$$

On the other hand, in “abstract” form,

$$\mathbf{A}\cdot\mathbf{B} = AB\cos\theta = \sqrt{2}\sqrt{2}\cos\theta = 2\cos\theta.$$

Therefore,

$$\cos\theta = 1/2, \quad \text{or} \quad \theta = 60^\circ.$$

Of course, you can get the answer more easily by drawing in a diagonal across the top of the cube, completing the equilateral triangle. But in cases where the geometry is not so simple, this device of comparing the abstract and component forms of the dot product can be a very efficient means of finding angles.

Problem 1.3 Find the angle between the body diagonals of a cube.

Problem 1.4 Use the cross product to find the components of the unit vector $\hat{\mathbf{n}}$ perpendicular to the shaded plane in Fig. 1.11.

1.1.3 ■ Triple Products

Since the cross product of two vectors is itself a vector, it can be dotted or crossed with a third vector to form a *triple* product.

(i) Scalar triple product: $\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})$. Geometrically, $|\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})|$ is the volume of the parallelepiped generated by $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, since $|\mathbf{B}\times\mathbf{C}|$ is the area of the base, and $|A\cos\theta|$ is the altitude (Fig. 1.12). Evidently,

$$\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = \mathbf{B}\cdot(\mathbf{C}\times\mathbf{A}) = \mathbf{C}\cdot(\mathbf{A}\times\mathbf{B}), \tag{1.15}$$

for they all correspond to the same figure. Note that "alphabetical" order is preserved—in view of Eq. 1.6, the "nonalphabetical" triple products,

$$\mathbf{A}\cdot(\mathbf{C}\times\mathbf{B}) = \mathbf{B}\cdot(\mathbf{A}\times\mathbf{C}) = \mathbf{C}\cdot(\mathbf{B}\times\mathbf{A}),$$

have the opposite sign. In component form,

$$\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}. \tag{1.16}$$

Note that the dot and cross can be interchanged:

$$\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = (\mathbf{A}\times\mathbf{B})\cdot\mathbf{C}$$

(this follows immediately from Eq. 1.15); however, the placement of the parentheses is critical: $(\mathbf{A}\cdot\mathbf{B})\times\mathbf{C}$ is a meaningless expression—you can't make a cross product from a *scalar* and a vector.

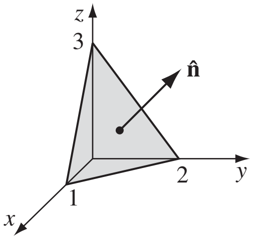


FIGURE 1.11

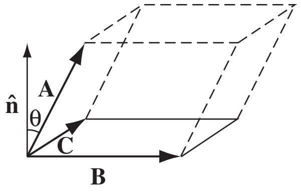


FIGURE 1.12

(ii) Vector triple product: $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$. The vector triple product can be simplified by the so-called **BAC-CAB** rule:

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B}). \tag{1.17}$$

Notice that

$$(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = -\mathbf{C} \times (\mathbf{A} \times \mathbf{B}) = -\mathbf{A}(\mathbf{B} \cdot \mathbf{C}) + \mathbf{B}(\mathbf{A} \cdot \mathbf{C})$$

is an entirely different vector (cross-products are not associative). All *higher* vector products can be similarly reduced, often by repeated application of Eq. 1.17, so it is never necessary for an expression to contain more than one cross product in any term. For instance,

$$\begin{aligned} (\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) &= (\mathbf{A} \cdot \mathbf{C})(\mathbf{B} \cdot \mathbf{D}) - (\mathbf{A} \cdot \mathbf{D})(\mathbf{B} \cdot \mathbf{C}); \\ \mathbf{A} \times [\mathbf{B} \times (\mathbf{C} \times \mathbf{D})] &= \mathbf{B}[\mathbf{A} \cdot (\mathbf{C} \times \mathbf{D})] - (\mathbf{A} \cdot \mathbf{B})(\mathbf{C} \times \mathbf{D}). \end{aligned} \tag{1.18}$$

Problem 1.5 Prove the **BAC-CAB** rule by writing out both sides in component form.

Problem 1.6 Prove that

$$[\mathbf{A} \times (\mathbf{B} \times \mathbf{C})] + [\mathbf{B} \times (\mathbf{C} \times \mathbf{A})] + [\mathbf{C} \times (\mathbf{A} \times \mathbf{B})] = \mathbf{0}.$$

Under what conditions does $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$?

1.1.4 ■ Position, Displacement, and Separation Vectors

The location of a point in three dimensions can be described by listing its Cartesian coordinates (x, y, z). The vector to that point from the origin ($\mathcal{O}$) is called the **position vector** (Fig. 1.13):

$$\mathbf{r} \equiv x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}. \tag{1.19}$$

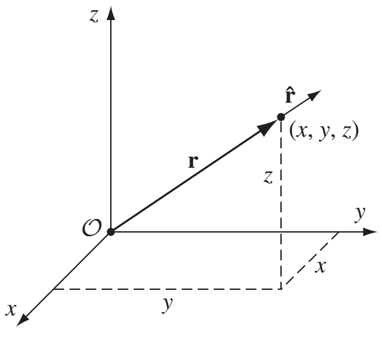


FIGURE 1.13

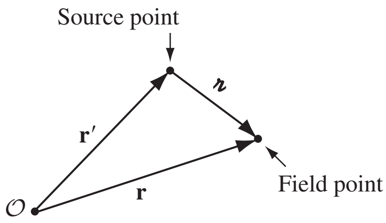


FIGURE 1.14

I will reserve the letter $\mathbf{r}$ for this purpose, throughout the book. Its magnitude,

$$r = \sqrt{x^2 + y^2 + z^2}, \tag{1.20}$$

is the distance from the origin, and

$$\hat{\mathbf{r}} = \frac{\mathbf{r}}{r} = \frac{x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}}{\sqrt{x^2 + y^2 + z^2}} \tag{1.21}$$

is a unit vector pointing radially outward. The **infinitesimal displacement vector**, from (x, y, z) to $(x + dx, y + dy, z + dz)$, is

$$d\mathbf{l} = dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}. \tag{1.22}$$

(We could call this $d\mathbf{r}$, since that's what it *is*, but it is useful to have a special notation for infinitesimal displacements.)

In electrodynamics, one frequently encounters problems involving *two* points—typically, a **source point**, $\mathbf{r}'$, where an electric charge is located, and a **field point**, $\mathbf{r}$, at which you are calculating the electric or magnetic field (Fig. 1.14). It pays to adopt right from the start some short-hand notation for the **separation vector** from the source point to the field point. I shall use for this purpose the script letter $\boldsymbol{\mathscr{r}}$:

$$\boldsymbol{\mathscr{r}} \equiv \mathbf{r} - \mathbf{r}'. \tag{1.23}$$

Its magnitude is

$$\mathscr{r} = |\mathbf{r} - \mathbf{r}'|, \tag{1.24}$$

and a unit vector in the direction from $\mathbf{r}'$ to $\mathbf{r}$ is

$$\hat{\boldsymbol{\mathscr{r}}} = \frac{\boldsymbol{\mathscr{r}}}{\mathscr{r}} = \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|}. \tag{1.25}$$

In Cartesian coordinates,

$$\boldsymbol{\mathscr{r}} = (x - x')\hat{\mathbf{x}} + (y - y')\hat{\mathbf{y}} + (z - z')\hat{\mathbf{z}}, \tag{1.26}$$

$$\mathscr{r} = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}, \tag{1.27}$$

$$\hat{\boldsymbol{\mathscr{r}}} = \frac{(x - x')\hat{\mathbf{x}} + (y - y')\hat{\mathbf{y}} + (z - z')\hat{\mathbf{z}}}{\sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}} \tag{1.28}$$

(from which you can appreciate the economy of the script-$\mathscr{r}$ notation).

Problem 1.7 Find the separation vector $\boldsymbol{\mathscr{r}}$ from the source point (2,8,7) to the field point (4,6,8). Determine its magnitude ($\mathscr{r}$), and construct the unit vector $\hat{\boldsymbol{\mathscr{r}}}$.

1.1.5 ■ How Vectors Transform[2]

The definition of a vector as "a quantity with a magnitude and direction" is not altogether satisfactory: What precisely does "direction" *mean*? This may seem a pedantic question, but we shall soon encounter a species of derivative that *looks* rather like a vector, and we'll want to know for sure whether it *is* one.

You might be inclined to say that a vector is anything that has three components that combine properly under addition. Well, how about this: We have a barrel of fruit that contains N_x pears, N_y apples, and N_z bananas. Is $\mathbf{N} = N_x\hat{\mathbf{x}} + N_y\hat{\mathbf{y}} + N_z\hat{\mathbf{z}}$ a vector? It has three components, and when you add another barrel with M_x pears, M_y apples, and M_z bananas the result is $(N_x + M_x)$ pears, $(N_y + M_y)$ apples, $(N_z + M_z)$ bananas. So it does *add* like a vector. Yet it's obviously *not* a vector, in the physicist's sense of the word, because it doesn't really have a direction. What exactly is wrong with it?

The answer is that **N** *does not transform properly when you change coordinates.* The coordinate frame we use to describe positions in space is of course entirely arbitrary, but there is a specific geometrical transformation law for converting vector components from one frame to another. Suppose, for instance, the $\overline{x}, \overline{y}, \overline{z}$ system is rotated by angle ϕ, relative to x, y, z, about the common $x = \overline{x}$ axes. From Fig. 1.15,

$$A_y = A\cos\theta, \qquad A_z = A\sin\theta,$$

while

$$\begin{aligned}\overline{A}_y &= A\cos\overline{\theta} = A\cos(\theta - \phi) = A(\cos\theta\cos\phi + \sin\theta\sin\phi)\\ &= \cos\phi A_y + \sin\phi A_z,\\ \overline{A}_z &= A\sin\overline{\theta} = A\sin(\theta - \phi) = A(\sin\theta\cos\phi - \cos\theta\sin\phi)\\ &= -\sin\phi A_y + \cos\phi A_z.\end{aligned}$$

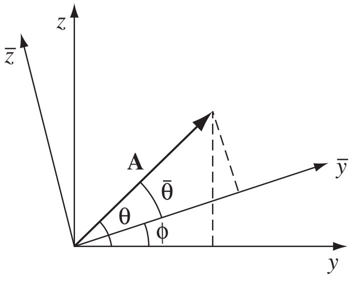


FIGURE 1.15

[2] This section can be skipped without loss of continuity.

We might express this conclusion in matrix notation:

$$\begin{pmatrix} \overline{A}_y \\ \overline{A}_z \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} A_y \\ A_z \end{pmatrix}. \tag{1.29}$$

More generally, for rotation about an *arbitrary* axis in three dimensions, the transformation law takes the form

$$\begin{pmatrix} \overline{A}_x \\ \overline{A}_y \\ \overline{A}_z \end{pmatrix} = \begin{pmatrix} R_{xx} & R_{xy} & R_{xz} \\ R_{yx} & R_{yy} & R_{yz} \\ R_{zx} & R_{zy} & R_{zz} \end{pmatrix} \begin{pmatrix} A_x \\ A_y \\ A_z \end{pmatrix}, \tag{1.30}$$

or, more compactly,

$$\overline{A}_i = \sum_{j=1}^{3} R_{ij} A_j, \tag{1.31}$$

where the index 1 stands for x, 2 for y, and 3 for z. The elements of the matrix R can be ascertained, for a given rotation, by the same sort of trigonometric arguments as we used for a rotation about the x axis.

Now: *Do* the components of **N** transform in this way? Of *course* not—it doesn't matter what coordinates you use to represent positions in space; there are still just as many apples in the barrel. You can't convert a pear into a banana by choosing a different set of axes, but you *can* turn A_x into $\overline{A}_y$. Formally, then, a *vector is any set of three components that transforms in the same manner as a displacement when you change coordinates.* As always, displacement is the *model* for the behavior of all vectors.[3]

By the way, a (second-rank) **tensor** is a quantity with *nine* components, T_{xx}, T_{xy}, T_{xz}, T_{yx}, $\ldots$, T_{zz}, which transform with *two* factors of R:

$$\begin{aligned} \overline{T}_{xx} = {} & R_{xx}(R_{xx}T_{xx} + R_{xy}T_{xy} + R_{xz}T_{xz}) \\ & + R_{xy}(R_{xx}T_{yx} + R_{xy}T_{yy} + R_{xz}T_{yz}) \\ & + R_{xz}(R_{xx}T_{zx} + R_{xy}T_{zy} + R_{xz}T_{zz}), \ldots \end{aligned}$$

or, more compactly,

$$\overline{T}_{ij} = \sum_{k=1}^{3} \sum_{l=1}^{3} R_{ik} R_{jl} T_{kl}. \tag{1.32}$$

[3] If you're a mathematician you might want to contemplate generalized vector spaces in which the "axes" have nothing to do with direction and the basis vectors are no longer $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ (indeed, there may be more than three dimensions). This is the subject of **linear algebra**. But for our purposes all vectors live in ordinary 3-space (or, in Chapter 12, in 4-dimensional space-time.)

In general, an nth-rank tensor has n indices and 3^n components, and transforms with n factors of R. In this hierarchy, a vector is a tensor of rank 1, and a scalar is a tensor of rank zero.[4]

Problem 1.8

(a) Prove that the two-dimensional rotation matrix (Eq. 1.29) preserves dot products. (That is, show that $\overline{A}_y \overline{B}_y + \overline{A}_z \overline{B}_z = A_y B_y + A_z B_z$.)

(b) What constraints must the elements (R_{ij}) of the three-dimensional rotation matrix (Eq. 1.30) satisfy, in order to preserve the length of **A** (for all vectors **A**)?

Problem 1.9 Find the transformation matrix R that describes a rotation by 120° about an axis from the origin through the point (1, 1, 1). The rotation is clockwise as you look down the axis toward the origin.

Problem 1.10

(a) How do the components of a vector[5] transform under a **translation** of coordinates ($\overline{x} = x$, $\overline{y} = y - a$, $\overline{z} = z$, Fig. 1.16a)?

(b) How do the components of a vector transform under an **inversion** of coordinates ($\overline{x} = -x$, $\overline{y} = -y$, $\overline{z} = -z$, Fig. 1.16b)?

(c) How do the components of a cross product (Eq. 1.13) transform under inversion? [The cross-product of two vectors is properly called a **pseudovector** because of this "anomalous" behavior.] Is the cross product of two pseudovectors a vector, or a pseudovector? Name two pseudovector quantities in classical mechanics.

(d) How does the scalar triple product of three vectors transform under inversions? (Such an object is called a **pseudoscalar.**)

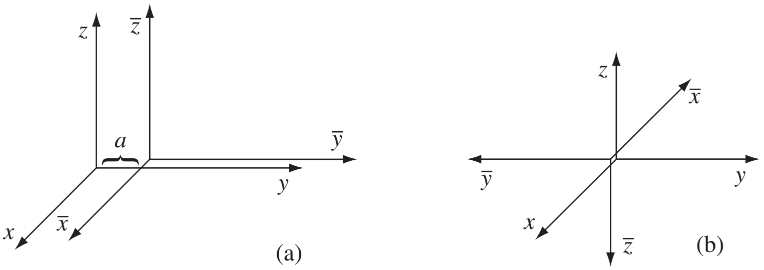


FIGURE 1.16

[4] A scalar does not change when you change coordinates. In particular, the components of a vector are *not* scalars, but the magnitude is.

[5] *Beware:* The vector **r** (Eq. 1.19) goes from a specific point in space (the origin, $\mathcal{O}$) to the point $P = (x, y, z)$. Under translations the *new* origin ($\bar{\mathcal{O}}$) is at a different location, and the arrow from $\bar{\mathcal{O}}$ to P is a completely different vector. The original vector **r** still goes from $\mathcal{O}$ to P, regardless of the coordinates used to label these points.

1.2 ■ DIFFERENTIAL CALCULUS

1.2.1 ■ "Ordinary" Derivatives

Suppose we have a function of one variable: $f(x)$. *Question:* What does the derivative, df/dx, do for us? *Answer:* It tells us how rapidly the function $f(x)$ varies when we change the argument x by a tiny amount, dx:

$$df = \left(\frac{df}{dx}\right) dx. \tag{1.33}$$

In words: If we increment x by an infinitesimal amount dx, then f changes by an amount df; the derivative is the proportionality factor. For example, in Fig. 1.17(a), the function varies slowly with x, and the derivative is correspondingly small. In Fig. 1.17(b), f increases rapidly with x, and the derivative is large, as you move away from $x = 0$.

Geometrical Interpretation: The derivative df/dx is the *slope* of the graph of f versus x.

1.2.2 ■ Gradient

Suppose, now, that we have a function of *three* variables—say, the temperature $T(x, y, z)$ in this room. (Start out in one corner, and set up a system of axes; then for each point (x, y, z) in the room, T gives the temperature at that spot.) We want to generalize the notion of "derivative" to functions like T, which depend not on *one* but on *three* variables.

A derivative is supposed to tell us how fast the function varies, if we move a little distance. But this time the situation is more complicated, because it depends on what *direction* we move: If we go straight up, then the temperature will probably increase fairly rapidly, but if we move horizontally, it may not change much at all. In fact, the question "How fast does T vary?" has an infinite number of answers, one for each direction we might choose to explore.

Fortunately, the problem is not as bad as it looks. A theorem on partial derivatives states that

$$dT = \left(\frac{\partial T}{\partial x}\right) dx + \left(\frac{\partial T}{\partial y}\right) dy + \left(\frac{\partial T}{\partial z}\right) dz. \tag{1.34}$$

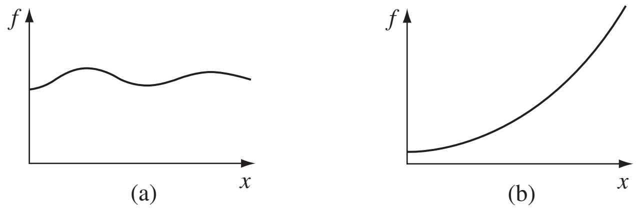


FIGURE 1.17

This tells us how T changes when we alter all three variables by the infinitesimal amounts dx, dy, dz. Notice that we do *not* require an infinite number of derivatives—*three* will suffice: the *partial* derivatives along each of the three coordinate directions.

Equation 1.34 is reminiscent of a dot product:

$$\begin{aligned} dT &= \left(\frac{\partial T}{\partial x}\hat{\mathbf{x}} + \frac{\partial T}{\partial y}\hat{\mathbf{y}} + \frac{\partial T}{\partial z}\hat{\mathbf{z}}\right) \cdot (dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}) \\ &= (\boldsymbol{\nabla} T) \cdot (d\mathbf{l}), \end{aligned} \tag{1.35}$$

where

$$\boldsymbol{\nabla} T \equiv \frac{\partial T}{\partial x}\hat{\mathbf{x}} + \frac{\partial T}{\partial y}\hat{\mathbf{y}} + \frac{\partial T}{\partial z}\hat{\mathbf{z}} \tag{1.36}$$

is the **gradient** of T. Note that $\boldsymbol{\nabla} T$ is a *vector* quantity, with three components; it is the generalized derivative we have been looking for. Equation 1.35 is the three-dimensional version of Eq. 1.33.

Geometrical Interpretation of the Gradient: Like any vector, the gradient has *magnitude* and *direction*. To determine its geometrical meaning, let's rewrite the dot product (Eq. 1.35) using Eq. 1.1:

$$dT = \boldsymbol{\nabla} T \cdot d\mathbf{l} = |\boldsymbol{\nabla} T||d\mathbf{l}|\cos\theta, \tag{1.37}$$

where θ is the angle between $\boldsymbol{\nabla} T$ and $d\mathbf{l}$. Now, if we *fix* the *magnitude* $|d\mathbf{l}|$ and search around in various *directions* (that is, vary θ), the *maximum* change in T evidentally occurs when $\theta = 0$ (for then $\cos\theta = 1$). That is, for a fixed distance $|d\mathbf{l}|$, dT is greatest when I move in the *same direction* as $\boldsymbol{\nabla} T$. Thus:

> *The gradient $\boldsymbol{\nabla} T$ points in the direction of maximum increase of the function T.*

Moreover:

> *The magnitude $|\boldsymbol{\nabla} T|$ gives the slope (rate of increase) along this maximal direction.*

Imagine you are standing on a hillside. Look all around you, and find the direction of steepest ascent. That is the *direction* of the gradient. Now measure the *slope* in that direction (rise over run). That is the *magnitude* of the gradient. (Here the function we're talking about is the height of the hill, and the coordinates it depends on are positions—latitude and longitude, say. This function depends on only *two* variables, not *three*, but the geometrical meaning of the gradient is easier to grasp in two dimensions.) Notice from Eq. 1.37 that the direction of maximum *descent* is opposite to the direction of maximum *ascent*, while at right angles ($\theta = 90°$) the slope is zero (the gradient is perpendicular to the contour lines). You can conceive of surfaces that do not have these properties, but they always have "kinks" in them, and correspond to nondifferentiable functions.

What would it mean for the gradient to vanish? If $\boldsymbol{\nabla} T = \mathbf{0}$ at (x, y, z), then $dT = 0$ for small displacements about the point (x, y, z). This is, then, a **stationary point** of the function $T(x, y, z)$. It could be a maximum (a summit), a minimum (a valley), a saddle point (a pass), or a “shoulder.” This is analogous to the situation for functions of *one* variable, where a vanishing derivative signals a maximum, a minimum, or an inflection. In particular, if you want to locate the extrema of a function of three variables, set its gradient equal to zero.

Example 1.3. Find the gradient of $r = \sqrt{x^2 + y^2 + z^2}$ (the magnitude of the position vector).

Solution

$$\begin{aligned}\nabla r &= \frac{\partial r}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial r}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial r}{\partial z}\,\hat{\mathbf{z}} \\ &= \frac{1}{2}\frac{2x}{\sqrt{x^2+y^2+z^2}}\,\hat{\mathbf{x}} + \frac{1}{2}\frac{2y}{\sqrt{x^2+y^2+z^2}}\,\hat{\mathbf{y}} + \frac{1}{2}\frac{2z}{\sqrt{x^2+y^2+z^2}}\,\hat{\mathbf{z}} \\ &= \frac{x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}}{\sqrt{x^2+y^2+z^2}} = \frac{\mathbf{r}}{r} = \hat{\mathbf{r}}.\end{aligned}$$

Does this make sense? Well, it says that the distance from the origin increases most rapidly in the radial direction, and that its *rate* of increase in that direction is 1…just what you’d expect.

Problem 1.11 Find the gradients of the following functions:

(a) $f(x, y, z) = x^2 + y^3 + z^4$.

(b) $f(x, y, z) = x^2 y^3 z^4$.

(c) $f(x, y, z) = e^x \sin(y) \ln(z)$.

Problem 1.12 The height of a certain hill (in feet) is given by

$$h(x, y) = 10(2xy - 3x^2 - 4y^2 - 18x + 28y + 12),$$

where y is the distance (in miles) north, x the distance east of South Hadley.

(a) Where is the top of the hill located?

(b) How high is the hill?

(c) How steep is the slope (in feet per mile) at a point 1 mile north and one mile east of South Hadley? In what direction is the slope steepest, at that point?

• **Problem 1.13** Let $\boldsymbol{\mathscr{r}}$ be the separation vector from a fixed point (x', y', z') to the point (x, y, z), and let $\mathscr{r}$ be its length. Show that

(a) $\nabla(\mathscr{r}^2) = 2\boldsymbol{\mathscr{r}}$.

(b) $\nabla(1/\mathscr{r}) = -\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$.

(c) What is the *general* formula for $\nabla(\mathscr{r}^n)$?

! **Problem 1.14** Suppose that f is a function of two variables (y and z) only. Show that the gradient $\nabla f = (\partial f/\partial y)\hat{\mathbf{y}} + (\partial f/\partial z)\hat{\mathbf{z}}$ transforms as a vector under rotations, Eq. 1.29. [*Hint:* $(\partial f/\partial \overline{y}) = (\partial f/\partial y)(\partial y/\partial \overline{y}) + (\partial f/\partial z)(\partial z/\partial \overline{y})$, and the analogous formula for $\partial f/\partial \overline{z}$. We know that $\overline{y} = y\cos\phi + z\sin\phi$ and $\overline{z} = -y\sin\phi + z\cos\phi$; "solve" these equations for y and z (as functions of $\overline{y}$ and $\overline{z}$), and compute the needed derivatives $\partial y/\partial \overline{y}$, $\partial z/\partial \overline{y}$, etc.]

1.2.3 ■ The Del Operator

The gradient has the formal appearance of a vector, ∇, "multiplying" a scalar T:

$$\nabla T = \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right)T. \tag{1.38}$$

(For once, I write the unit vectors to the *left,* just so no one will think this means $\partial\hat{\mathbf{x}}/\partial x$, and so on—which would be zero, since $\hat{\mathbf{x}}$ is constant.) The term in parentheses is called **del**:

$$\boxed{\nabla = \hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}.} \tag{1.39}$$

Of course, del is *not* a vector, in the usual sense. Indeed, it doesn't mean much until we provide it with a function to act upon. Furthermore, it does not "multiply" T; rather, it is an instruction to *differentiate* what follows. To be precise, then, we say that ∇ is a **vector operator** that *acts upon* T, not a vector that multiplies T.

With this qualification, though, ∇ mimics the behavior of an ordinary vector in virtually every way; almost anything that can be done with other vectors can also be done with ∇, if we merely translate "multiply" by "act upon." So by all means take the vector appearance of ∇ seriously: it is a marvelous piece of notational simplification, as you will appreciate if you ever consult Maxwell's original work on electromagnetism, written without the benefit of ∇.

Now, an ordinary vector $\mathbf{A}$ can multiply in three ways:

1. By a scalar $a : \mathbf{A}a$;
2. By a vector $\mathbf{B}$, via the dot product: $\mathbf{A}\cdot\mathbf{B}$;
3. By a vector $\mathbf{B}$ via the cross product: $\mathbf{A}\times\mathbf{B}$.

Correspondingly, there are three ways the operator ∇ can act:

1. On a scalar function $T : \nabla T$ (the gradient);
2. On a vector function $\mathbf{v}$, via the dot product: $\nabla\cdot\mathbf{v}$ (the **divergence**);
3. On a vector function $\mathbf{v}$, via the cross product: $\nabla\times\mathbf{v}$ (the **curl**).

We have already discussed the gradient. In the following sections we examine the other two vector derivatives: divergence and curl.

1.2.4 ■ The Divergence

From the definition of ∇ we construct the divergence:

$$\begin{aligned}\nabla \cdot \mathbf{v} &= \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right) \cdot (v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}} + v_z\hat{\mathbf{z}}) \\ &= \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}. \qquad (1.40)\end{aligned}$$

Observe that the divergence of a vector function[6] $\mathbf{v}$ is itself a *scalar* $\nabla \cdot \mathbf{v}$.

Geometrical Interpretation: The name **divergence** is well chosen, for $\nabla \cdot \mathbf{v}$ is a measure of how much the vector $\mathbf{v}$ spreads out (diverges) from the point in question. For example, the vector function in Fig. 1.18a has a large (positive) divergence (if the arrows pointed *in*, it would be a *negative* divergence), the function in Fig. 1.18b has zero divergence, and the function in Fig. 1.18c again has a positive divergence. (Please understand that $\mathbf{v}$ here is a *function*—there's a different vector associated with every point in space. In the diagrams, of course, I can only draw the arrows at a few representative locations.)

Imagine standing at the edge of a pond. Sprinkle some sawdust or pine needles on the surface. If the material spreads out, then you dropped it at a point of positive divergence; if it collects together, you dropped it at a point of negative divergence. (The vector function $\mathbf{v}$ in this model is the velocity of the water at the surface—this is a *two*-dimensional example, but it helps give one a "feel" for what the divergence means. A point of positive divergence is a source, or "faucet"; a point of negative divergence is a sink, or "drain.")

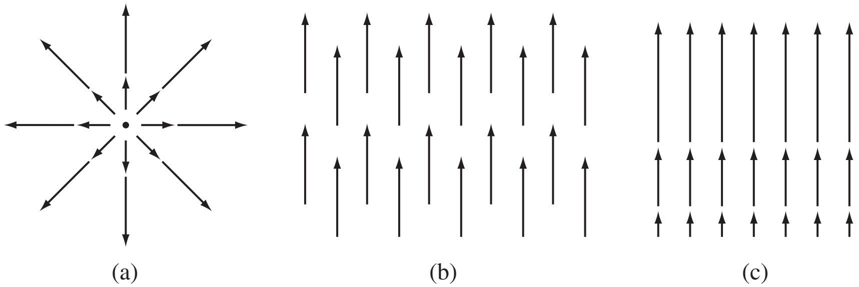


FIGURE 1.18

[6] A vector function $\mathbf{v}(x, y, z) = v_x(x, y, z)\,\hat{\mathbf{x}} + v_y(x, y, z)\,\hat{\mathbf{y}} + v_z(x, y, z)\,\hat{\mathbf{z}}$ is really *three* functions—one for each component. There's no such thing as the divergence of a scalar.

Example 1.4. Suppose the functions in Fig. 1.18 are $\mathbf{v}_a = \mathbf{r} = x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$, $\mathbf{v}_b = \hat{\mathbf{z}}$, and $\mathbf{v}_c = z\,\hat{\mathbf{z}}$. Calculate their divergences.

Solution

$$\nabla \cdot \mathbf{v}_a = \frac{\partial}{\partial x}(x) + \frac{\partial}{\partial y}(y) + \frac{\partial}{\partial z}(z) = 1 + 1 + 1 = 3.$$

As anticipated, this function has a positive divergence.

$$\nabla \cdot \mathbf{v}_b = \frac{\partial}{\partial x}(0) + \frac{\partial}{\partial y}(0) + \frac{\partial}{\partial z}(1) = 0 + 0 + 0 = 0,$$

as expected.

$$\nabla \cdot \mathbf{v}_c = \frac{\partial}{\partial x}(0) + \frac{\partial}{\partial y}(0) + \frac{\partial}{\partial z}(z) = 0 + 0 + 1 = 1.$$

Problem 1.15 Calculate the divergence of the following vector functions:

(a) $\mathbf{v}_a = x^2\,\hat{\mathbf{x}} + 3xz^2\,\hat{\mathbf{y}} - 2xz\,\hat{\mathbf{z}}$.

(b) $\mathbf{v}_b = xy\,\hat{\mathbf{x}} + 2yz\,\hat{\mathbf{y}} + 3zx\,\hat{\mathbf{z}}$.

(c) $\mathbf{v}_c = y^2\,\hat{\mathbf{x}} + (2xy + z^2)\,\hat{\mathbf{y}} + 2yz\,\hat{\mathbf{z}}$.

• **Problem 1.16** Sketch the vector function

$$\mathbf{v} = \frac{\hat{\mathbf{r}}}{r^2},$$

and compute its divergence. The answer may surprise you. . . can you explain it?

! **Problem 1.17** In two dimensions, show that the divergence transforms as a scalar under rotations. [*Hint:* Use Eq. 1.29 to determine $\overline{v}_y$ and $\overline{v}_z$, and the method of Prob. 1.14 to calculate the derivatives. Your aim is to show that $\partial\overline{v}_y/\partial\overline{y} + \partial\overline{v}_z/\partial\overline{z} = \partial v_y/\partial y + \partial v_z/\partial z$.]

1.2.5 ■ The Curl

From the definition of ∇ we construct the curl:

$$\nabla \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ v_x & v_y & v_z \end{vmatrix}$$

$$= \hat{\mathbf{x}}\left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right) + \hat{\mathbf{y}}\left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x}\right) + \hat{\mathbf{z}}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right). \quad (1.41)$$

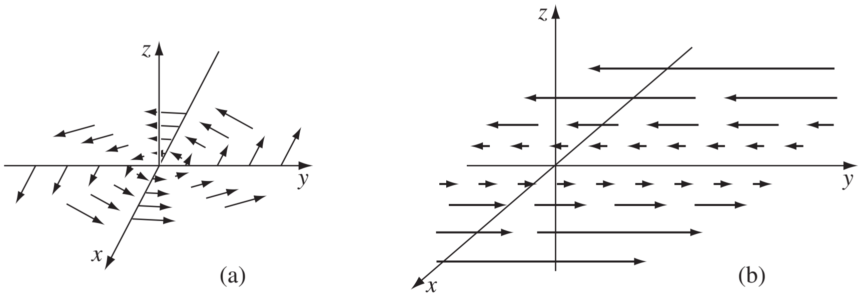


FIGURE 1.19

Notice that the curl of a vector function[7] **v** is, like any cross product, a *vector*.

Geometrical Interpretation: The name **curl** is also well chosen, for $\nabla \times \mathbf{v}$ is a measure of how much the vector **v** swirls around the point in question. Thus the three functions in Fig. 1.18 all have zero curl (as you can easily check for yourself), whereas the functions in Fig. 1.19 have a substantial curl, pointing in the z direction, as the natural right-hand rule would suggest. Imagine (again) you are standing at the edge of a pond. Float a small paddlewheel (a cork with toothpicks pointing out radially would do); if it starts to rotate, then you placed it at a point of nonzero *curl.* A whirlpool would be a region of large curl.

Example 1.5. Suppose the function sketched in Fig. 1.19a is $\mathbf{v}_a = -y\hat{\mathbf{x}} + x\hat{\mathbf{y}}$, and that in Fig. 1.19b is $\mathbf{v}_b = x\hat{\mathbf{y}}$. Calculate their curls.

Solution

$$\nabla \times \mathbf{v}_a = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ -y & x & 0 \end{vmatrix} = 2\hat{\mathbf{z}},$$

and

$$\nabla \times \mathbf{v}_b = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ 0 & x & 0 \end{vmatrix} = \hat{\mathbf{z}}.$$

As expected, these curls point in the $+z$ direction. (Incidentally, they both have zero divergence, as you might guess from the pictures: nothing is “spreading out”... it just “swirls around.”)

---

[7] There’s no such thing as the curl of a scalar.

Problem 1.18 Calculate the curls of the vector functions in Prob. 1.15.

Problem 1.19 Draw a circle in the xy plane. At a few representative points draw the vector $\mathbf{v}$ tangent to the circle, pointing in the clockwise direction. By comparing adjacent vectors, determine the *sign* of $\partial v_x/\partial y$ and $\partial v_y/\partial x$. According to Eq. 1.41, then, what is the direction of $\nabla \times \mathbf{v}$? Explain how this example illustrates the geometrical interpretation of the curl.

Problem 1.20 Construct a vector function that has zero divergence and zero curl everywhere. (A *constant* will do the job, of course, but make it something a little more interesting than that!)

1.2.6 ■ Product Rules

The calculation of ordinary derivatives is facilitated by a number of rules, such as the sum rule:

$$\frac{d}{dx}(f+g) = \frac{df}{dx} + \frac{dg}{dx},$$

the rule for multiplying by a constant:

$$\frac{d}{dx}(kf) = k\frac{df}{dx},$$

the product rule:

$$\frac{d}{dx}(fg) = f\frac{dg}{dx} + g\frac{df}{dx},$$

and the quotient rule:

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{g\dfrac{df}{dx} - f\dfrac{dg}{dx}}{g^2}.$$

Similar relations hold for the vector derivatives. Thus,

$$\boldsymbol{\nabla}(f+g) = \boldsymbol{\nabla} f + \boldsymbol{\nabla} g, \qquad \boldsymbol{\nabla}\cdot(\mathbf{A}+\mathbf{B}) = (\boldsymbol{\nabla}\cdot\mathbf{A}) + (\boldsymbol{\nabla}\cdot\mathbf{B}),$$

$$\boldsymbol{\nabla}\times(\mathbf{A}+\mathbf{B}) = (\boldsymbol{\nabla}\times\mathbf{A}) + (\boldsymbol{\nabla}\times\mathbf{B}),$$

and

$$\boldsymbol{\nabla}(kf) = k\boldsymbol{\nabla} f, \qquad \boldsymbol{\nabla}\cdot(k\mathbf{A}) = k(\boldsymbol{\nabla}\cdot\mathbf{A}), \qquad \boldsymbol{\nabla}\times(k\mathbf{A}) = k(\boldsymbol{\nabla}\times\mathbf{A}),$$

as you can check for yourself. The product rules are not quite so simple. There are two ways to construct a scalar as the product of two functions:

$$\begin{aligned} fg \quad & \text{(product of two scalar functions),} \\ \mathbf{A}\cdot\mathbf{B} \quad & \text{(dot product of two vector functions),} \end{aligned}$$

and two ways to make a vector:

$$\begin{aligned} f\mathbf{A} &\quad \text{(scalar times vector)},\\ \mathbf{A}\times\mathbf{B} &\quad \text{(cross product of two vectors)}. \end{aligned}$$

Accordingly, there are *six* product rules, two for gradients:

(i) $$\boldsymbol{\nabla}(fg) = f\boldsymbol{\nabla}g + g\boldsymbol{\nabla}f,$$

(ii) $$\boldsymbol{\nabla}(\mathbf{A}\cdot\mathbf{B}) = \mathbf{A}\times(\boldsymbol{\nabla}\times\mathbf{B}) + \mathbf{B}\times(\boldsymbol{\nabla}\times\mathbf{A}) + (\mathbf{A}\cdot\boldsymbol{\nabla})\mathbf{B} + (\mathbf{B}\cdot\boldsymbol{\nabla})\mathbf{A},$$

two for divergences:

(iii) $$\boldsymbol{\nabla}\cdot(f\mathbf{A}) = f(\boldsymbol{\nabla}\cdot\mathbf{A}) + \mathbf{A}\cdot(\boldsymbol{\nabla}f),$$

(iv) $$\boldsymbol{\nabla}\cdot(\mathbf{A}\times\mathbf{B}) = \mathbf{B}\cdot(\boldsymbol{\nabla}\times\mathbf{A}) - \mathbf{A}\cdot(\boldsymbol{\nabla}\times\mathbf{B}),$$

and two for curls:

(v) $$\boldsymbol{\nabla}\times(f\mathbf{A}) = f(\boldsymbol{\nabla}\times\mathbf{A}) - \mathbf{A}\times(\boldsymbol{\nabla}f),$$

(vi) $$\boldsymbol{\nabla}\times(\mathbf{A}\times\mathbf{B}) = (\mathbf{B}\cdot\boldsymbol{\nabla})\mathbf{A} - (\mathbf{A}\cdot\boldsymbol{\nabla})\mathbf{B} + \mathbf{A}(\boldsymbol{\nabla}\cdot\mathbf{B}) - \mathbf{B}(\boldsymbol{\nabla}\cdot\mathbf{A}).$$

You will be using these product rules so frequently that I have put them inside the front cover for easy reference. The proofs come straight from the product rule for ordinary derivatives. For instance,

$$\begin{aligned} \boldsymbol{\nabla}\cdot(f\mathbf{A}) &= \frac{\partial}{\partial x}(fA_x) + \frac{\partial}{\partial y}(fA_y) + \frac{\partial}{\partial z}(fA_z)\\ &= \left(\frac{\partial f}{\partial x}A_x + f\frac{\partial A_x}{\partial x}\right) + \left(\frac{\partial f}{\partial y}A_y + f\frac{\partial A_y}{\partial y}\right) + \left(\frac{\partial f}{\partial z}A_z + f\frac{\partial A_z}{\partial z}\right)\\ &= (\boldsymbol{\nabla}f)\cdot\mathbf{A} + f(\boldsymbol{\nabla}\cdot\mathbf{A}). \end{aligned}$$

It is also possible to formulate three quotient rules:

$$\begin{aligned} \boldsymbol{\nabla}\left(\frac{f}{g}\right) &= \frac{g\boldsymbol{\nabla}f - f\boldsymbol{\nabla}g}{g^2},\\ \boldsymbol{\nabla}\cdot\left(\frac{\mathbf{A}}{g}\right) &= \frac{g(\boldsymbol{\nabla}\cdot\mathbf{A}) - \mathbf{A}\cdot(\boldsymbol{\nabla}g)}{g^2},\\ \boldsymbol{\nabla}\times\left(\frac{\mathbf{A}}{g}\right) &= \frac{g(\boldsymbol{\nabla}\times\mathbf{A}) + \mathbf{A}\times(\boldsymbol{\nabla}g)}{g^2}. \end{aligned}$$

However, since these can be obtained quickly from the corresponding product rules, there is no point in listing them separately.

Problem 1.21 Prove product rules (i), (iv), and (v).

Problem 1.22

(a) If $\mathbf{A}$ and $\mathbf{B}$ are two vector functions, what does the expression $(\mathbf{A}\cdot\nabla)\mathbf{B}$ mean? (That is, what are its x, y, and z components, in terms of the Cartesian components of $\mathbf{A}$, $\mathbf{B}$, and ∇?)

(b) Compute $(\hat{\mathbf{r}}\cdot\nabla)\hat{\mathbf{r}}$, where $\hat{\mathbf{r}}$ is the unit vector defined in Eq. 1.21.

(c) For the functions in Prob. 1.15, evaluate $(\mathbf{v}_a\cdot\nabla)\mathbf{v}_b$.

Problem 1.23 (For masochists only.) Prove product rules (ii) and (vi). Refer to Prob. 1.22 for the definition of $(\mathbf{A}\cdot\nabla)\mathbf{B}$.

Problem 1.24 Derive the three quotient rules.

Problem 1.25

(a) Check product rule (iv) (by calculating each term separately) for the functions

$$\mathbf{A} = x\,\hat{\mathbf{x}} + 2y\,\hat{\mathbf{y}} + 3z\,\hat{\mathbf{z}}; \qquad \mathbf{B} = 3y\,\hat{\mathbf{x}} - 2x\,\hat{\mathbf{y}}.$$

(b) Do the same for product rule (ii).

(c) Do the same for rule (vi).

1.2.7 ■ Second Derivatives

The gradient, the divergence, and the curl are the only first derivatives we can make with ∇; by applying ∇ *twice*, we can construct five species of *second* derivatives. The gradient ∇T is a *vector*, so we can take the *divergence* and *curl* of it:

(1) Divergence of gradient: $\nabla\cdot(\nabla T)$.

(2) Curl of gradient: $\nabla\times(\nabla T)$.

The divergence $\nabla\cdot\mathbf{v}$ is a *scalar*—all we can do is take its *gradient:*

(3) Gradient of divergence: $\nabla(\nabla\cdot\mathbf{v})$.

The curl $\nabla\times\mathbf{v}$ is a *vector*, so we can take its *divergence* and *curl:*

(4) Divergence of curl: $\nabla\cdot(\nabla\times\mathbf{v})$.

(5) Curl of curl: $\nabla\times(\nabla\times\mathbf{v})$.

This exhausts the possibilities, and in fact not all of them give anything new. Let's consider them one at a time:

$$(1) \qquad \nabla\cdot(\nabla T) = \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right)\cdot\left(\frac{\partial T}{\partial x}\hat{\mathbf{x}} + \frac{\partial T}{\partial y}\hat{\mathbf{y}} + \frac{\partial T}{\partial z}\hat{\mathbf{z}}\right)$$
$$= \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}. \qquad (1.42)$$

This object, which we write as $\nabla^2 T$ for short, is called the **Laplacian** of T; we shall be studying it in great detail later on. Notice that the Laplacian of a *scalar* T is a *scalar*. Occasionally, we shall speak of the Laplacian of a *vector,* $\nabla^2\mathbf{v}$. By this we mean a *vector* quantity whose x-component is the Laplacian of v_x, and so on:[8]

$$\nabla^2\mathbf{v} \equiv (\nabla^2 v_x)\hat{\mathbf{x}} + (\nabla^2 v_y)\hat{\mathbf{y}} + (\nabla^2 v_z)\hat{\mathbf{z}}. \tag{1.43}$$

This is nothing more than a convenient extension of the meaning of ∇^2.

(2) The curl of a gradient is always zero:

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} T) = \mathbf{0}. \tag{1.44}$$

This is an important fact, which we shall use repeatedly; you can easily prove it from the definition of $\boldsymbol{\nabla}$, Eq. 1.39. *Beware*: You might think Eq. 1.44 is "obviously" true—isn't it just $(\boldsymbol{\nabla} \times \boldsymbol{\nabla})T$, and isn't the cross product of *any* vector (in this case, $\boldsymbol{\nabla}$) with itself always zero? This reasoning is suggestive, but not quite conclusive, since $\boldsymbol{\nabla}$ is an *operator* and does not "multiply" in the usual way. The proof of Eq. 1.44, in fact, hinges on the equality of cross derivatives:

$$\frac{\partial}{\partial x}\left(\frac{\partial T}{\partial y}\right) = \frac{\partial}{\partial y}\left(\frac{\partial T}{\partial x}\right). \tag{1.45}$$

If you think I'm being fussy, test your intuition on this one:

$$(\boldsymbol{\nabla} T) \times (\boldsymbol{\nabla} S).$$

Is *that* always zero? (It *would* be, of course, if you replaced the $\boldsymbol{\nabla}$'s by an ordinary vector.)

(3) $\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{v})$ seldom occurs in physical applications, and it has not been given any special name of its own—it's just **the gradient of the divergence.** Notice that $\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{v})$ is *not* the same as the Laplacian of a vector: $\nabla^2\mathbf{v} = (\boldsymbol{\nabla} \cdot \boldsymbol{\nabla})\mathbf{v} \neq \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{v})$.

(4) The divergence of a curl, like the curl of a gradient, is always zero:

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \mathbf{v}) = 0. \tag{1.46}$$

You can prove this for yourself. (Again, there is a fraudulent short-cut proof, using the vector identity $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}$.)

(5) As you can check from the definition of $\boldsymbol{\nabla}$:

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \mathbf{v}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{v}) - \nabla^2\mathbf{v}. \tag{1.47}$$

So curl-of-curl gives nothing new; the first term is just number (3), and the second is the Laplacian (of a vector). (In fact, Eq. 1.47 is often used to *define* the

[8] In curvilinear coordinates, where the unit vectors themselves depend on position, they too must be differentiated (see Sect. 1.4.1).

Laplacian of a vector, in preference to Eq. 1.43, which makes explicit reference to Cartesian coordinates.)

Really, then, there are just two kinds of second derivatives: the Laplacian (which is of fundamental importance) and the gradient-of-divergence (which we seldom encounter). We could go through a similar ritual to work out *third* derivatives, but fortunately second derivatives suffice for practically all physical applications.

A final word on vector differential calculus: It *all* flows from the operator $\boldsymbol{\nabla}$, and from taking seriously its vectorial character. Even if you remembered *only* the definition of $\boldsymbol{\nabla}$, you could easily reconstruct all the rest.

Problem 1.26 Calculate the Laplacian of the following functions:

(a) $T_a = x^2 + 2xy + 3z + 4$.

(b) $T_b = \sin x \sin y \sin z$.

(c) $T_c = e^{-5x} \sin 4y \cos 3z$.

(d) $\mathbf{v} = x^2\,\hat{\mathbf{x}} + 3xz^2\,\hat{\mathbf{y}} - 2xz\,\hat{\mathbf{z}}$.

Problem 1.27 Prove that the divergence of a curl is always zero. *Check* it for function $\mathbf{v}_a$ in Prob. 1.15.

Problem 1.28 Prove that the curl of a gradient is always zero. *Check* it for function (b) in Prob. 1.11.

1.3 ■ INTEGRAL CALCULUS

1.3.1 ■ Line, Surface, and Volume Integrals

In electrodynamics, we encounter several different kinds of integrals, among which the most important are **line** (or **path**) **integrals**, **surface integrals** (or **flux**), and **volume integrals**.

(a) **Line Integrals.** A line integral is an expression of the form

$$\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{v} \cdot d\mathbf{l}, \tag{1.48}$$

where $\mathbf{v}$ is a vector function, $d\mathbf{l}$ is the infinitesimal displacement vector (Eq. 1.22), and the integral is to be carried out along a prescribed path $\mathcal{P}$ from point $\mathbf{a}$ to point $\mathbf{b}$ (Fig. 1.20). If the path in question forms a closed loop (that is, if $\mathbf{b} = \mathbf{a}$), I shall put a circle on the integral sign:

$$\oint \mathbf{v} \cdot d\mathbf{l}. \tag{1.49}$$

At each point on the path, we take the dot product of $\mathbf{v}$ (evaluated at that point) with the displacement $d\mathbf{l}$ to the next point on the path. To a physicist, the most familiar example of a line integral is the work done by a force $\mathbf{F}$: $W = \int \mathbf{F} \cdot d\mathbf{l}$.

Ordinarily, the value of a line integral depends critically on the path taken from $\mathbf{a}$ to $\mathbf{b}$, but there is an important special class of vector functions for which the line

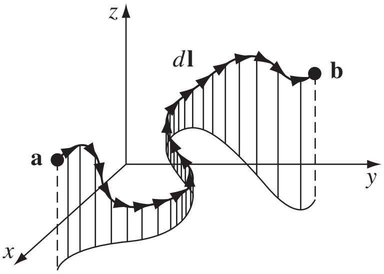


FIGURE 1.20

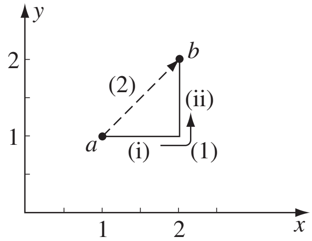


FIGURE 1.21

integral is *independent* of path and is determined entirely by the end points. It will be our business in due course to characterize this special class of vectors. (A *force* that has this property is called **conservative**.)

Example 1.6. Calculate the line integral of the function $\mathbf{v} = y^2\,\hat{\mathbf{x}} + 2x(y+1)\,\hat{\mathbf{y}}$ from the point $\mathbf{a} = (1, 1, 0)$ to the point $\mathbf{b} = (2, 2, 0)$, along the paths (1) and (2) in Fig. 1.21. What is $\oint \mathbf{v}\cdot d\mathbf{l}$ for the loop that goes from **a** to **b** along (1) and returns to **a** along (2)?

Solution

As always, $d\mathbf{l} = dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}$. Path (1) consists of two parts. Along the "horizontal" segment, $dy = dz = 0$, so

(i) $d\mathbf{l} = dx\,\hat{\mathbf{x}},\ y = 1,\ \mathbf{v}\cdot d\mathbf{l} = y^2\,dx = dx$, so $\int \mathbf{v}\cdot d\mathbf{l} = \int_1^2 dx = 1$.

On the "vertical" stretch, $dx = dz = 0$, so

(ii) $d\mathbf{l} = dy\,\hat{\mathbf{y}},\ x = 2,\ \mathbf{v}\cdot d\mathbf{l} = 2x(y+1)\,dy = 4(y+1)\,dy$, so

$$\int \mathbf{v}\cdot d\mathbf{l} = 4\int_1^2 (y+1)\,dy = 10.$$

By path (1), then,

$$\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{v}\cdot d\mathbf{l} = 1 + 10 = 11.$$

Meanwhile, on path (2) $x = y,\ dx = dy$, and $dz = 0$, so
$d\mathbf{l} = dx\,\hat{\mathbf{x}} + dx\,\hat{\mathbf{y}},\ \mathbf{v}\cdot d\mathbf{l} = x^2\,dx + 2x(x+1)\,dx = (3x^2 + 2x)\,dx$,
and

$$\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{v}\cdot d\mathbf{l} = \int_1^2 (3x^2 + 2x)\,dx = (x^3 + x^2)\Big|_1^2 = 10.$$

(The strategy here is to get everything in terms of one variable; I could just as well have eliminated x in favor of y.)

For the loop that goes *out* (1) and *back* (2), then,

$$\oint \mathbf{v}\cdot d\mathbf{l} = 11 - 10 = 1.$$

(b) **Surface Integrals.** A surface integral is an expression of the form

$$\int_{\mathcal{S}} \mathbf{v}\cdot d\mathbf{a}, \tag{1.50}$$

where **v** is again some vector function, and the integral is over a specified surface $\mathcal{S}$. Here $d\mathbf{a}$ is an infinitesimal patch of area, with direction perpendicular to the surface (Fig. 1.22). There are, of course, *two* directions perpendicular to any surface, so the *sign* of a surface integral is intrinsically ambiguous. If the surface is *closed* (forming a "balloon"), in which case I shall again put a circle on the integral sign

$$\oint \mathbf{v}\cdot d\mathbf{a},$$

then tradition dictates that "outward" is positive, but for open surfaces it's arbitrary. If **v** describes the flow of a fluid (mass per unit area per unit time), then $\int \mathbf{v}\cdot d\mathbf{a}$ represents the total mass per unit time passing through the surface—hence the alternative name, "flux."

Ordinarily, the value of a surface integral depends on the particular surface chosen, but there is a special class of vector functions for which it is *independent* of the surface and is determined entirely by the boundary line. An important task will be to characterize this special class of functions.

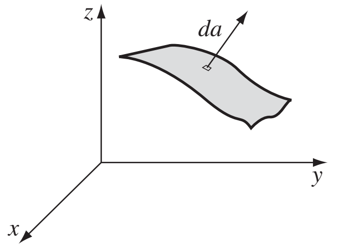


FIGURE 1.22

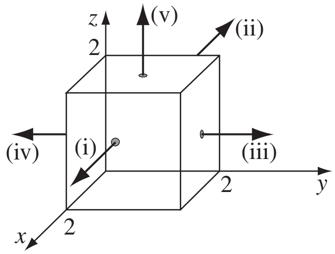


FIGURE 1.23

Example 1.7. Calculate the surface integral of $\mathbf{v} = 2xz\,\hat{\mathbf{x}} + (x+2)\,\hat{\mathbf{y}} + y(z^2-3)\,\hat{\mathbf{z}}$ over five sides (excluding the bottom) of the cubical box (side 2) in Fig. 1.23. Let "upward and outward" be the positive direction, as indicated by the arrows.

Solution

Taking the sides one at a time:

(i) $x = 2,\ d\mathbf{a} = dy\,dz\,\hat{\mathbf{x}},\ \mathbf{v}\cdot d\mathbf{a} = 2xz\,dy\,dz = 4z\,dy\,dz$, so

$$\int \mathbf{v}\cdot d\mathbf{a} = 4\int_0^2 dy\int_0^2 z\,dz = 16.$$

(ii) $x = 0,\ d\mathbf{a} = -dy\,dz\,\hat{\mathbf{x}},\ \mathbf{v}\cdot d\mathbf{a} = -2xz\,dy\,dz = 0$, so

$$\int \mathbf{v}\cdot d\mathbf{a} = 0.$$

(iii) $y = 2,\ d\mathbf{a} = dx\,dz\,\hat{\mathbf{y}},\ \mathbf{v}\cdot d\mathbf{a} = (x+2)\,dx\,dz$, so

$$\int \mathbf{v}\cdot d\mathbf{a} = \int_0^2 (x+2)\,dx\int_0^2 dz = 12.$$

(iv) $y = 0,\ d\mathbf{a} = -dx\,dz\,\hat{\mathbf{y}},\ \mathbf{v}\cdot d\mathbf{a} = -(x+2)\,dx\,dz$, so

$$\int \mathbf{v}\cdot d\mathbf{a} = -\int_0^2 (x+2)\,dx\int_0^2 dz = -12.$$

(v) $z = 2,\ d\mathbf{a} = dx\,dy\,\hat{\mathbf{z}},\ \mathbf{v}\cdot d\mathbf{a} = y(z^2-3)\,dx\,dy = y\,dx\,dy$, so

$$\int \mathbf{v}\cdot d\mathbf{a} = \int_0^2 dx\int_0^2 y\,dy = 4.$$

The *total* flux is

$$\int_{\text{surface}} \mathbf{v}\cdot d\mathbf{a} = 16 + 0 + 12 - 12 + 4 = 20.$$

(c) **Volume Integrals.** A volume integral is an expression of the form

$$\int_{\mathcal{V}} T\,d\tau, \tag{1.51}$$

where T is a scalar function and $d\tau$ is an infinitesimal volume element. In Cartesian coordinates,

$$d\tau = dx\,dy\,dz. \tag{1.52}$$

For example, if T is the density of a substance (which might vary from point to point), then the volume integral would give the total mass. Occasionally we shall encounter volume integrals of *vector* functions:

$$\int \mathbf{v}\,d\tau = \int (v_x\,\hat{\mathbf{x}} + v_y\,\hat{\mathbf{y}} + v_z\,\hat{\mathbf{z}})d\tau = \hat{\mathbf{x}}\int v_x d\tau + \hat{\mathbf{y}}\int v_y d\tau + \hat{\mathbf{z}}\int v_z d\tau; \tag{1.53}$$

because the unit vectors ($\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$) are constants, they come outside the integral.

Example 1.8. Calculate the volume integral of $T = xyz^2$ over the prism in Fig. 1.24.

Solution

You can do the three integrals in any order. Let's do x first: it runs from 0 to $(1 - y)$, then y (it goes from 0 to 1), and finally z (0 to 3):

$$\int T\,d\tau = \int_0^3 z^2 \left\{\int_0^1 y \left[\int_0^{1-y} x\,dx\right] dy\right\} dz$$

$$= \frac{1}{2}\int_0^3 z^2\,dz \int_0^1 (1-y)^2 y\,dy = \frac{1}{2}(9)\left(\frac{1}{12}\right) = \frac{3}{8}.$$

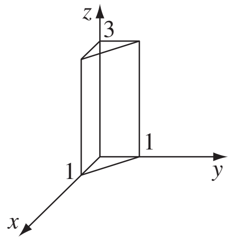


FIGURE 1.24

Problem 1.29 Calculate the line integral of the function $\mathbf{v} = x^2\,\hat{\mathbf{x}} + 2yz\,\hat{\mathbf{y}} + y^2\,\hat{\mathbf{z}}$ from the origin to the point (1,1,1) by three different routes:

(a) (0, 0, 0) → (1, 0, 0) → (1, 1, 0) → (1, 1, 1).

(b) (0, 0, 0) → (0, 0, 1) → (0, 1, 1) → (1, 1, 1).

(c) The direct straight line.

(d) What is the line integral around the closed loop that goes *out* along path (a) and *back* along path (b)?

Problem 1.30 Calculate the surface integral of the function in Ex. 1.7, over the *bottom* of the box. For consistency, let "upward" be the positive direction. Does the surface integral depend only on the boundary line for this function? What is the total flux over the *closed* surface of the box (*including* the bottom)? [*Note:* For the *closed* surface, the positive direction is "outward," and hence "down," for the bottom face.]

Problem 1.31 Calculate the volume integral of the function $T = z^2$ over the tetrahedron with corners at (0,0,0), (1,0,0), (0,1,0), and (0,0,1).

1.3.2 ■ The Fundamental Theorem of Calculus

Suppose $f(x)$ is a function of one variable. The **fundamental theorem of calculus** says:

$$\int_a^b \left(\frac{df}{dx}\right) dx = f(b) - f(a). \tag{1.54}$$

In case this doesn't look familiar, I'll write it another way:

$$\int_a^b F(x)\, dx = f(b) - f(a),$$

where $df/dx = F(x)$. The fundamental theorem tells you how to integrate $F(x)$: you think up a function $f(x)$ whose *derivative* is equal to F.

Geometrical Interpretation: According to Eq. 1.33, $df = (df/dx)dx$ is the infinitesimal change in f when you go from (x) to $(x + dx)$. The fundamental theorem (Eq. 1.54) says that if you chop the interval from a to b (Fig. 1.25) into many tiny pieces, dx, and add up the increments df from each little piece, the result is (not surprisingly) equal to the total change in f: $f(b) - f(a)$. In other words, there are two ways to determine the total change in the function: *either* subtract the values at the ends *or* go step-by-step, adding up all the tiny increments as you go. You'll get the same answer either way.

Notice the basic format of the fundamental theorem: the *integral* of a *derivative* over some *region* is given by the *value of the function* at the end points *(boundaries)*. In vector calculus there are three species of derivative (gradient, divergence, and curl), and each has its own "fundamental theorem," with essentially the same format. I don't plan to prove these theorems here; rather, I will explain what they *mean*, and try to make them *plausible*. Proofs are given in Appendix A.

1.3.3 ■ The Fundamental Theorem for Gradients

Suppose we have a scalar function of three variables $T(x, y, z)$. Starting at point $\mathbf{a}$, we move a small distance $d\mathbf{l}_1$ (Fig. 1.26). According to Eq. 1.37, the function T will change by an amount

$$dT = (\nabla T) \cdot d\mathbf{l}_1.$$

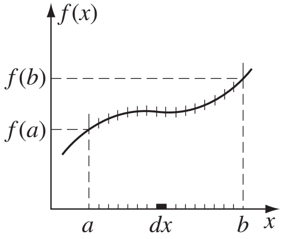


FIGURE 1.25

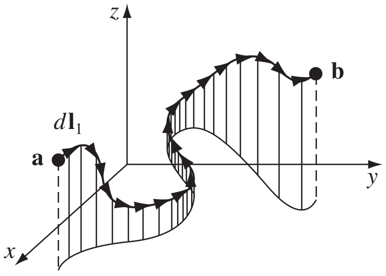


FIGURE 1.26

Now we move a little further, by an additional small displacement $d\mathbf{l}_2$; the incremental change in T will be $(\nabla T)\cdot d\mathbf{l}_2$. In this manner, proceeding by infinitesimal steps, we make the journey to point **b**. At each step we compute the gradient of T (at that point) and dot it into the displacement $d\mathbf{l}$... this gives us the change in T. Evidently the *total* change in T in going from **a** to **b** (along the path selected) is

$$\boxed{\int_{\mathbf{a}}^{\mathbf{b}} (\nabla T)\cdot d\mathbf{l} = T(\mathbf{b}) - T(\mathbf{a}).} \tag{1.55}$$

This is the **fundamental theorem for gradients**; like the "ordinary" fundamental theorem, it says that the integral (here a *line* integral) of a derivative (here the *gradient*) is given by the value of the function at the boundaries (**a** and **b**).

Geometrical Interpretation: Suppose you wanted to determine the height of the Eiffel Tower. You could climb the stairs, using a ruler to measure the rise at each step, and adding them all up (that's the left side of Eq. 1.55), or you could place altimeters at the top and the bottom, and subtract the two readings (that's the right side); you should get the same answer either way (that's the fundamental theorem).

Incidentally, as we found in Ex. 1.6, line integrals ordinarily depend on the *path* taken from **a** to **b**. But the *right* side of Eq. 1.55 makes no reference to the path—only to the end points. Evidently, *gradients* have the special property that their line integrals are path independent:

Corollary 1: $\int_{\mathbf{a}}^{\mathbf{b}}(\nabla T)\cdot d\mathbf{l}$ is independent of the path taken from **a** to **b**.

Corollary 2: $\oint(\nabla T)\cdot d\mathbf{l} = 0$, since the beginning and end points are identical, and hence $T(\mathbf{b}) - T(\mathbf{a}) = 0$.

Example 1.9. Let $T = xy^2$, and take point **a** to be the origin (0, 0, 0) and **b** the point (2, 1, 0). Check the fundamental theorem for gradients.

Solution

Although the integral is independent of path, we must *pick* a specific path in order to evaluate it. Let's go out along the x axis (step i) and then up (step ii) (Fig. 1.27). As always, $d\mathbf{l} = dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}$; $\nabla T = y^2\,\hat{\mathbf{x}} + 2xy\,\hat{\mathbf{y}}$.

(i) $y = 0;\ d\mathbf{l} = dx\,\hat{\mathbf{x}},\ \nabla T\cdot d\mathbf{l} = y^2\,dx = 0$, so

$$\int_{\text{i}} \nabla T\cdot d\mathbf{l} = 0.$$

(ii) $x = 2;\ d\mathbf{l} = dy\,\hat{\mathbf{y}},\ \nabla T\cdot d\mathbf{l} = 2xy\,dy = 4y\,dy$, so

$$\int_{\text{ii}} \nabla T\cdot d\mathbf{l} = \int_0^1 4y\,dy = 2y^2\Big|_0^1 = 2.$$

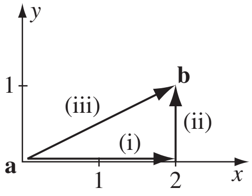


FIGURE 1.27

The total line integral is 2. Is this consistent with the fundamental theorem? Yes: $T(\mathbf{b}) - T(\mathbf{a}) = 2 - 0 = 2$.

Now, just to convince you that the answer is independent of path, let me calculate the same integral along path iii (the straight line from **a** to **b**):

(iii) $y = \frac{1}{2}x,\ dy = \frac{1}{2}\,dx,\ \nabla T \cdot d\mathbf{l} = y^2\,dx + 2xy\,dy = \frac{3}{4}x^2\,dx$, so

$$\int_{\text{iii}} \nabla T \cdot d\mathbf{l} = \int_0^2 \tfrac{3}{4}x^2\,dx = \tfrac{1}{4}x^3\Big|_0^2 = 2.$$

Problem 1.32 Check the fundamental theorem for gradients, using $T = x^2 + 4xy + 2yz^3$, the points $\mathbf{a} = (0, 0, 0)$, $\mathbf{b} = (1, 1, 1)$, and the three paths in Fig. 1.28:

(a) $(0, 0, 0) \to (1, 0, 0) \to (1, 1, 0) \to (1, 1, 1)$;

(b) $(0, 0, 0) \to (0, 0, 1) \to (0, 1, 1) \to (1, 1, 1)$;

(c) the parabolic path $z = x^2$; $y = x$.

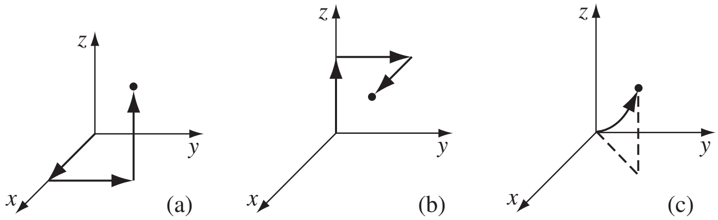


FIGURE 1.28

1.3.4 ■ The Fundamental Theorem for Divergences

The fundamental theorem for divergences states that:

$$\boxed{\int_{\mathcal{V}} (\nabla \cdot \mathbf{v})\,d\tau = \oint_{\mathcal{S}} \mathbf{v} \cdot d\mathbf{a}.} \tag{1.56}$$

In honor, I suppose, of its great importance, this theorem has at least three special names: **Gauss's theorem**, **Green's theorem**, or simply the **divergence theorem**. Like the other "fundamental theorems," it says that the *integral* of a *derivative* (in this case the *divergence*) over a *region* (in this case a *volume*, $\mathcal{V}$) is equal to the value of the function at the *boundary* (in this case the *surface* $\mathcal{S}$ that bounds the volume). Notice that the boundary term is itself an integral (specifically, a surface integral). This is reasonable: the "boundary" of a *line* is just two end points, but the boundary of a *volume* is a (closed) surface.

Geometrical Interpretation: If **v** represents the flow of an incompressible fluid, then the *flux* of **v** (the right side of Eq. 1.56) is the total amount of fluid passing out through the surface, per unit time. Now, the divergence measures the "spreading out" of the vectors from a point—a place of high divergence is like a "faucet," pouring out liquid. If we have a bunch of faucets in a region filled with incompressible fluid, an equal amount of liquid will be forced out through the boundaries of the region. In fact, there are *two* ways we could determine how much is being produced: (a) we could count up all the faucets, recording how much each puts out, or (b) we could go around the boundary, measuring the flow at each point, and add it all up. You get the same answer either way:

$$\int (\text{faucets within the volume}) = \oint (\text{flow out through the surface}).$$

This, in essence, is what the divergence theorem says.

Example 1.10. Check the divergence theorem using the function

$$\mathbf{v} = y^2\,\hat{\mathbf{x}} + (2xy + z^2)\,\hat{\mathbf{y}} + (2yz)\,\hat{\mathbf{z}}$$

and a unit cube at the origin (Fig. 1.29).

Solution

In this case

$$\nabla \cdot \mathbf{v} = 2(x + y),$$

and

$$\int_{\mathcal{V}} 2(x + y)\,d\tau = 2\int_0^1\int_0^1\int_0^1 (x + y)\,dx\,dy\,dz,$$

$$\int_0^1 (x + y)\,dx = \tfrac{1}{2} + y, \qquad \int_0^1 (\tfrac{1}{2} + y)\,dy = 1, \qquad \int_0^1 1\,dz = 1.$$

Thus,

$$\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\,d\tau = 2.$$

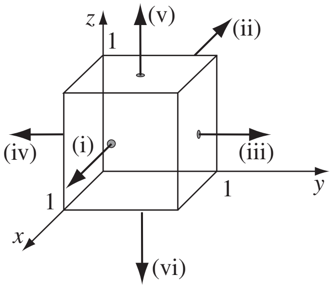


FIGURE 1.29

So much for the left side of the divergence theorem. To evaluate the surface integral we must consider separately the six faces of the cube:

(i) $$\int \mathbf{v}\cdot d\mathbf{a} = \int_0^1\int_0^1 y^2 dy\,dz = \tfrac{1}{3}.$$

(ii) $$\int \mathbf{v}\cdot d\mathbf{a} = -\int_0^1\int_0^1 y^2\,dy\,dz = -\tfrac{1}{3}.$$

(iii) $$\int \mathbf{v}\cdot d\mathbf{a} = \int_0^1\int_0^1 (2x+z^2)\,dx\,dz = \tfrac{4}{3}.$$

(iv) $$\int \mathbf{v}\cdot d\mathbf{a} = -\int_0^1\int_0^1 z^2\,dx\,dz = -\tfrac{1}{3}.$$

(v) $$\int \mathbf{v}\cdot d\mathbf{a} = \int_0^1\int_0^1 2y\,dx\,dy = 1.$$

(vi) $$\int \mathbf{v}\cdot d\mathbf{a} = -\int_0^1\int_0^1 0\,dx\,dy = 0.$$

So the total flux is:

$$\oint_{\mathcal{S}} \mathbf{v}\cdot d\mathbf{a} = \tfrac{1}{3} - \tfrac{1}{3} + \tfrac{4}{3} - \tfrac{1}{3} + 1 + 0 = 2,$$

as expected.

Problem 1.33 Test the divergence theorem for the function $\mathbf{v} = (xy)\,\hat{\mathbf{x}} + (2yz)\,\hat{\mathbf{y}} + (3zx)\,\hat{\mathbf{z}}$. Take as your volume the cube shown in Fig. 1.30, with sides of length 2.

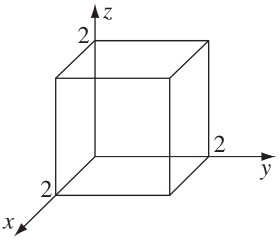


FIGURE 1.30

1.3.5 ■ The Fundamental Theorem for Curls

The fundamental theorem for curls, which goes by the special name of **Stokes' theorem**, states that

$$\boxed{\int_{\mathcal{S}} (\nabla \times \mathbf{v}) \cdot d\mathbf{a} = \oint_{\mathcal{P}} \mathbf{v} \cdot d\mathbf{l}.} \tag{1.57}$$

As always, the *integral* of a *derivative* (here, the *curl*) over a *region* (here, a patch of *surface*, $\mathcal{S}$) is equal to the value of the function at the *boundary* (here, the perimeter of the patch, $\mathcal{P}$). As in the case of the divergence theorem, the boundary term is itself an integral—specifically, a closed line integral.

Geometrical Interpretation: Recall that the curl measures the "twist" of the vectors **v**; a region of high curl is a whirlpool—if you put a tiny paddle wheel there, it will rotate. Now, the integral of the curl over some surface (or, more precisely, the *flux* of the curl *through* that surface) represents the "total amount of swirl," and we can determine that just as well by going around the edge and finding how much the flow is following the boundary (Fig. 1.31). Indeed, $\oint \mathbf{v} \cdot d\mathbf{l}$ is sometimes called the **circulation** of **v**.

You may have noticed an apparent ambiguity in Stokes' theorem: concerning the boundary line integral, which *way* are we supposed to go around (clockwise or counterclockwise)? If we go the "wrong" way, we'll pick up an overall sign error. The answer is that it doesn't matter which way you go as long as you are consistent, for there is a compensating sign ambiguity in the surface integral: Which way does $d\mathbf{a}$ point? For a *closed* surface (as in the divergence theorem), $d\mathbf{a}$ points in the direction of the *outward* normal; but for an *open* surface, which way is "out"? Consistency in Stokes' theorem (as in all such matters) is given by the right-hand rule: if your fingers point in the direction of the line integral, then your thumb fixes the direction of $d\mathbf{a}$ (Fig. 1.32).

Now, there are plenty of surfaces (infinitely many) that share any given boundary line. Twist a paper clip into a loop, and dip it in soapy water. The soap film constitutes a surface, with the wire loop as its boundary. If you blow on it, the soap film will expand, making a larger surface, with the same boundary. Ordinarily, a flux integral depends critically on what surface you integrate over, but evidently

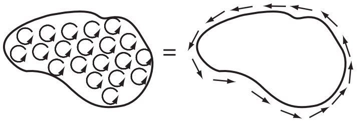


FIGURE 1.31

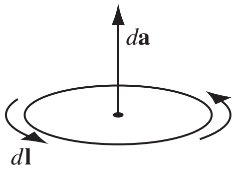


FIGURE 1.32

this is *not* the case with curls. For Stokes' theorem says that $\int(\nabla \times \mathbf{v}) \cdot d\mathbf{a}$ is equal to the line integral of $\mathbf{v}$ around the boundary, and the latter makes no reference to the specific surface you choose.

Corollary 1: $\int(\nabla \times \mathbf{v}) \cdot d\mathbf{a}$ depends only on the boundary line, not on the particular surface used.

Corollary 2: $\oint(\nabla \times \mathbf{v}) \cdot d\mathbf{a} = 0$ for any closed surface, since the boundary line, like the mouth of a balloon, shrinks down to a point, and hence the right side of Eq. 1.57 vanishes.

These corollaries are analogous to those for the gradient theorem. We will develop the parallel further in due course.

Example 1.11. Suppose $\mathbf{v} = (2xz + 3y^2)\hat{\mathbf{y}} + (4yz^2)\hat{\mathbf{z}}$. Check Stokes' theorem for the square surface shown in Fig. 1.33.

Solution
Here

$$\nabla \times \mathbf{v} = (4z^2 - 2x)\,\hat{\mathbf{x}} + 2z\,\hat{\mathbf{z}} \quad \text{and} \quad d\mathbf{a} = dy\,dz\,\hat{\mathbf{x}}.$$

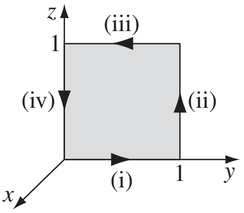


FIGURE 1.33

(In saying that $d\mathbf{a}$ points in the x direction, we are committing ourselves to a counterclockwise line integral. We could as well write $d\mathbf{a} = -dy\,dz\,\hat{\mathbf{x}}$, but then we would be obliged to go clockwise.) Since $x = 0$ for this surface,

$$\int(\nabla \times \mathbf{v}) \cdot d\mathbf{a} = \int_0^1 \int_0^1 4z^2\,dy\,dz = \frac{4}{3}.$$

Now, what about the line integral? We must break this up into four segments:

$$\begin{array}{lllll}
\text{(i)} & x=0, & z=0, & \mathbf{v}\cdot d\mathbf{l}=3y^2\,dy, & \int \mathbf{v}\cdot d\mathbf{l}=\int_0^1 3y^2\,dy=1,\\
\text{(ii)} & x=0, & y=1, & \mathbf{v}\cdot d\mathbf{l}=4z^2\,dz, & \int \mathbf{v}\cdot d\mathbf{l}=\int_0^1 4z^2\,dz=\dfrac{4}{3},\\
\text{(iii)} & x=0, & z=1, & \mathbf{v}\cdot d\mathbf{l}=3y^2\,dy, & \int \mathbf{v}\cdot d\mathbf{l}=\int_1^0 3y^2\,dy=-1,\\
\text{(iv)} & x=0, & y=0, & \mathbf{v}\cdot d\mathbf{l}=0, & \int \mathbf{v}\cdot d\mathbf{l}=\int_1^0 0\,dz=0.
\end{array}$$

So

$$\oint \mathbf{v}\cdot d\mathbf{l}=1+\frac{4}{3}-1+0=\frac{4}{3}.$$

It checks.

A point of strategy: notice how I handled step (iii). There is a temptation to write $d\mathbf{l}=-dy\,\hat{\mathbf{y}}$ here, since the path goes to the left. You can get away with this, if you absolutely insist, by running the integral from 0 → 1. But it is much safer to say $d\mathbf{l}=dx\,\hat{\mathbf{x}}+dy\,\hat{\mathbf{y}}+dz\,\hat{\mathbf{z}}$ *always* (never any minus signs) and let the limits of the integral take care of the direction.

Problem 1.34 Test Stokes' theorem for the function $\mathbf{v}=(xy)\,\hat{\mathbf{x}}+(2yz)\,\hat{\mathbf{y}}+(3zx)\,\hat{\mathbf{z}}$, using the triangular shaded area of Fig. 1.34.

Problem 1.35 Check Corollary 1 by using the same function and boundary line as in Ex. 1.11, but integrating over the five faces of the cube in Fig. 1.35. The back of the cube is open.

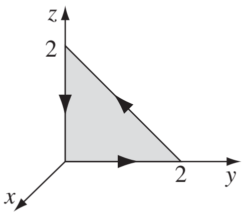


FIGURE 1.34

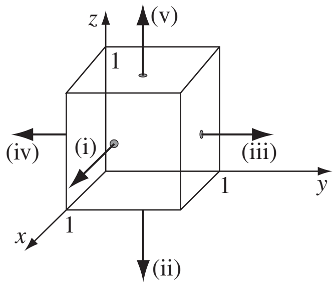


FIGURE 1.35

1.3.6 ■ Integration by Parts

The technique known (awkwardly) as **integration by parts** exploits the product rule for derivatives:

$$\frac{d}{dx}(fg)=f\left(\frac{dg}{dx}\right)+g\left(\frac{df}{dx}\right).$$

Integrating both sides, and invoking the fundamental theorem:

$$\int_a^b \frac{d}{dx}(fg)\,dx = fg\Big|_a^b = \int_a^b f\left(\frac{dg}{dx}\right)dx + \int_a^b g\left(\frac{df}{dx}\right)dx,$$

or

$$\int_a^b f\left(\frac{dg}{dx}\right)dx = -\int_a^b g\left(\frac{df}{dx}\right)dx + fg\Big|_a^b. \tag{1.58}$$

That's integration by parts. It applies to the situation in which you are called upon to integrate the product of one function (f) and the *derivative* of another (g); it says you can *transfer the derivative from g to f*, at the cost of a minus sign and a boundary term.

Example 1.12. Evaluate the integral

$$\int_0^\infty xe^{-x}\,dx.$$

Solution
The exponential can be expressed as a derivative:

$$e^{-x} = \frac{d}{dx}\left(-e^{-x}\right);$$

in this case, then, $f(x) = x$, $g(x) = -e^{-x}$, and $df/dx = 1$, so

$$\int_0^\infty xe^{-x}\,dx = \int_0^\infty e^{-x}\,dx - xe^{-x}\Big|_0^\infty = -e^{-x}\Big|_0^\infty = 1.$$

We can exploit the product rules of vector calculus, together with the appropriate fundamental theorems, in exactly the same way. For example, integrating

$$\nabla\cdot(f\mathbf{A}) = f(\nabla\cdot\mathbf{A}) + \mathbf{A}\cdot(\nabla f)$$

over a volume, and invoking the divergence theorem, yields

$$\int \nabla\cdot(f\mathbf{A})\,d\tau = \int f(\nabla\cdot\mathbf{A})\,d\tau + \int \mathbf{A}\cdot(\nabla f)\,d\tau = \oint f\mathbf{A}\cdot d\mathbf{a},$$

or

$$\int_{\mathcal{V}} f(\nabla\cdot\mathbf{A})\,d\tau = -\int_{\mathcal{V}} \mathbf{A}\cdot(\nabla f)\,d\tau + \oint_{\mathcal{S}} f\mathbf{A}\cdot d\mathbf{a}. \tag{1.59}$$

Here again the integrand is the product of one function (f) and the derivative (in this case the *divergence*) of another $(\mathbf{A})$, and integration by parts licenses us to transfer the derivative from **A** to f (where it becomes a *gradient*), at the cost of a minus sign and a boundary term (in this case a surface integral).

You might wonder how often one is likely to encounter an integral involving the product of one function and the derivative of another; the answer is *surprisingly* often, and integration by parts turns out to be one of the most powerful tools in vector calculus.

Problem 1.36

(a) Show that

$$\int_{\mathcal{S}} f(\nabla\times\mathbf{A})\cdot d\mathbf{a} = \int_{\mathcal{S}} [\mathbf{A}\times(\nabla f)]\cdot d\mathbf{a} + \oint_{\mathcal{P}} f\mathbf{A}\cdot d\mathbf{l}. \tag{1.60}$$

(b) Show that

$$\int_{\mathcal{V}} \mathbf{B}\cdot(\nabla\times\mathbf{A})\,d\tau = \int_{\mathcal{V}} \mathbf{A}\cdot(\nabla\times\mathbf{B})\,d\tau + \oint_{\mathcal{S}} (\mathbf{A}\times\mathbf{B})\cdot d\mathbf{a}. \tag{1.61}$$

1.4 ■ CURVILINEAR COORDINATES

1.4.1 ■ Spherical Coordinates

You can label a point P by its Cartesian coordinates (x, y, z), but sometimes it is more convenient to use **spherical** coordinates (r, θ, ϕ); r is the distance from the origin (the magnitude of the position vector **r**), θ (the angle down from the z axis) is called the **polar angle**, and ϕ (the angle around from the x axis) is the **azimuthal angle**. Their relation to Cartesian coordinates can be read from Fig. 1.36:

$$x = r\sin\theta\cos\phi, \qquad y = r\sin\theta\sin\phi, \qquad z = r\cos\theta. \tag{1.62}$$

Figure 1.36 also shows three unit vectors, $\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}$, pointing in the direction of increase of the corresponding coordinates. They constitute an orthogonal (mutually perpendicular) basis set (just like $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$), and any vector **A** can be expressed in terms of them, in the usual way:

$$\mathbf{A} = A_r\,\hat{\mathbf{r}} + A_\theta\,\hat{\boldsymbol{\theta}} + A_\phi\,\hat{\boldsymbol{\phi}}; \tag{1.63}$$

A_r, A_θ, and A_ϕ are the radial, polar, and azimuthal components of **A**. In terms of the Cartesian unit vectors,

$$\left.\begin{aligned} \hat{\mathbf{r}} &= \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}}, \\ \hat{\boldsymbol{\theta}} &= \cos\theta\cos\phi\,\hat{\mathbf{x}} + \cos\theta\sin\phi\,\hat{\mathbf{y}} - \sin\theta\,\hat{\mathbf{z}}, \\ \hat{\boldsymbol{\phi}} &= -\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}}, \end{aligned}\right\} \tag{1.64}$$

as you can check for yourself (Prob. 1.38). I have put these formulas inside the back cover, for easy reference.

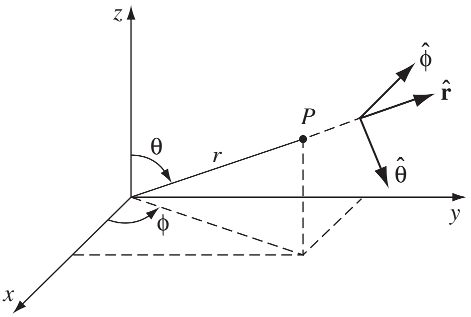


FIGURE 1.36

But there is a poisonous snake lurking here that I'd better warn you about: $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$ are associated with a *particular point P*, and they *change direction* as P moves around. For example, $\hat{\mathbf{r}}$ always points radially outward, but "radially outward" can be the x direction, the y direction, or any other direction, depending on where you are. In Fig. 1.37, $\mathbf{A} = \hat{\mathbf{y}}$ and $\mathbf{B} = -\hat{\mathbf{y}}$, and yet *both* of them would be written as $\hat{\mathbf{r}}$ in spherical coordinates. One could take account of this by explicitly indicating the point of reference: $\hat{\mathbf{r}}(\theta, \phi)$, $\hat{\boldsymbol{\theta}}(\theta, \phi)$, $\hat{\boldsymbol{\phi}}(\theta, \phi)$, but this would be cumbersome, and as long as you are alert to the problem, I don't think it will cause difficulties.[9] In particular, do not naïvely combine the spherical components of vectors associated with different points (in Fig. 1.37, $\mathbf{A} + \mathbf{B} = \mathbf{0}$, not $2\hat{\mathbf{r}}$, and $\mathbf{A} \cdot \mathbf{B} = -1$, not +1). Beware of differentiating a vector that is expressed in spherical coordinates, since the unit vectors themselves are functions of position ($\partial\hat{\mathbf{r}}/\partial\theta = \hat{\boldsymbol{\theta}}$, for example). And do not take $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$ outside an integral, as I did with $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ in Eq. 1.53. In general, if you're uncertain about the validity of an operation, rewrite the problem using Cartesian coordinates, for which this difficulty does not arise.

An infinitesimal displacement in the $\hat{\mathbf{r}}$ direction is simply dr (Fig. 1.38a), just as an infinitesimal element of length in the x direction is dx:

$$dl_r = dr. \tag{1.65}$$

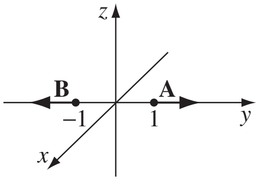


FIGURE 1.37

[9] I claimed back at the beginning that vectors have no location, and I'll stand by that. The vectors themselves live "out there," completely independent of our choice of coordinates. But the *notation* we use to represent them *does* depend on the point in question, in curvilinear coordinates.

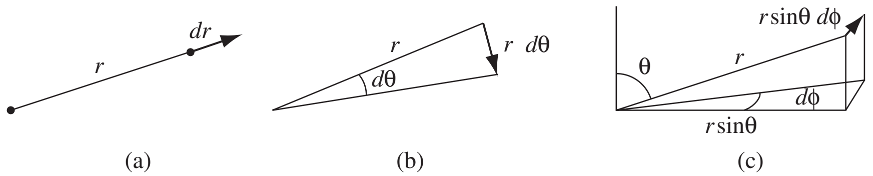


FIGURE 1.38

On the other hand, an infinitesimal element of length in the $\hat{\boldsymbol{\theta}}$ direction (Fig. 1.38b) is *not* just $d\theta$ (that's an *angle*—it doesn't even have the right *units* for a length); rather,

$$dl_\theta = r\,d\theta. \tag{1.66}$$

Similarly, an infinitesimal element of length in the $\hat{\boldsymbol{\phi}}$ direction (Fig. 1.38c) is

$$dl_\phi = r\sin\theta\,d\phi. \tag{1.67}$$

Thus the general infinitesimal displacement $d\mathbf{l}$ is

$$d\mathbf{l} = dr\,\hat{\mathbf{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\phi\,\hat{\boldsymbol{\phi}}. \tag{1.68}$$

This plays the role (in line integrals, for example) that $d\mathbf{l} = dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}$ played in Cartesian coordinates.

The infinitesimal volume element $d\tau$, in spherical coordinates, is the product of the three infinitesimal displacements:

$$d\tau = dl_r\,dl_\theta\,dl_\phi = r^2\sin\theta\,dr\,d\theta\,d\phi. \tag{1.69}$$

I cannot give you a general expression for *surface* elements $d\mathbf{a}$, since these depend on the orientation of the surface. You simply have to analyze the geometry for any given case (this goes for Cartesian and curvilinear coordinates alike). If you are integrating over the surface of a sphere, for instance, then r is constant, whereas θ and ϕ change (Fig. 1.39), so

$$d\mathbf{a}_1 = dl_\theta\,dl_\phi\,\hat{\mathbf{r}} = r^2\sin\theta\,d\theta\,d\phi\,\hat{\mathbf{r}}.$$

On the other hand, if the surface lies in the xy plane, say, so that θ is constant (to wit: $\pi/2$) while r and ϕ vary, then

$$d\mathbf{a}_2 = dl_r\,dl_\phi\,\hat{\boldsymbol{\theta}} = r\,dr\,d\phi\,\hat{\boldsymbol{\theta}}.$$

Notice, finally, that r ranges from 0 to ∞, ϕ from 0 to 2π, and θ from 0 to π (*not* 2π—that would count every point twice).[10]

[10] Alternatively, you could run ϕ from 0 to π (the "eastern hemisphere") and cover the "western hemisphere" by extending θ from π up to 2π. But this is very bad notation, since, among other things, $\sin\theta$ will then run negative, and you'll have to put absolute value signs around that term in volume and surface elements (area and volume being intrinsically positive quantities).

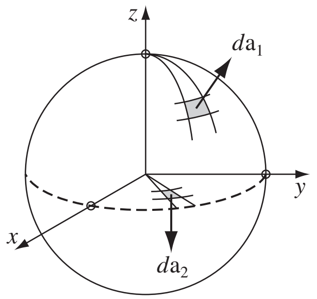


FIGURE 1.39

Example 1.13. Find the volume of a sphere of radius R.

Solution

$$V = \int d\tau = \int_{r=0}^{R} \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} r^2 \sin\theta \, dr \, d\theta \, d\phi$$
$$= \left(\int_0^R r^2 \, dr \right) \left(\int_0^\pi \sin\theta \, d\theta \right) \left(\int_0^{2\pi} d\phi \right)$$
$$= \left(\frac{R^3}{3} \right) (2)(2\pi) = \frac{4}{3}\pi R^3$$

(not a big surprise).

So far we have talked only about the *geometry* of spherical coordinates. Now I would like to "translate" the vector derivatives (gradient, divergence, curl, and Laplacian) into r, θ, ϕ notation. In principle, this is entirely straightforward: in the case of the gradient,

$$\nabla T = \frac{\partial T}{\partial x}\hat{\mathbf{x}} + \frac{\partial T}{\partial y}\hat{\mathbf{y}} + \frac{\partial T}{\partial z}\hat{\mathbf{z}},$$

for instance, we would first use the chain rule to expand the partials:

$$\frac{\partial T}{\partial x} = \frac{\partial T}{\partial r}\left(\frac{\partial r}{\partial x}\right) + \frac{\partial T}{\partial \theta}\left(\frac{\partial \theta}{\partial x}\right) + \frac{\partial T}{\partial \phi}\left(\frac{\partial \phi}{\partial x}\right).$$

The terms in parentheses could be worked out from Eq. 1.62—or rather, the *inverse* of those equations (Prob. 1.37). Then we'd do the same for $\partial T/\partial y$ and $\partial T/\partial z$. Finally, we'd substitute in the formulas for $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ in terms of $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$ (Prob. 1.38). It would take an hour to figure out the gradient in spherical coordinates by this brute-force method. I suppose this is how it was first done, but there is a much more efficient indirect approach, explained in Appendix A, which has the extra advantage of treating all coordinate systems at once. I described the "straightforward" method only to show you that there is nothing subtle or mysterious about transforming to spherical coordinates: you're expressing the *same quantity* (gradient, divergence, or whatever) in different notation, that's all.

Here, then, are the vector derivatives in spherical coordinates:

Gradient:

$$\nabla T = \frac{\partial T}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial T}{\partial \theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial T}{\partial \phi}\hat{\boldsymbol{\phi}}. \tag{1.70}$$

Divergence:

$$\nabla \cdot \mathbf{v} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 v_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, v_\theta) + \frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\phi}. \tag{1.71}$$

Curl:

$$\begin{aligned}\nabla \times \mathbf{v} = {} & \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta\, v_\phi) - \frac{\partial v_\theta}{\partial\phi}\right]\hat{\mathbf{r}} + \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial v_r}{\partial\phi} - \frac{\partial}{\partial r}(r v_\phi)\right]\hat{\boldsymbol{\theta}} \\ & + \frac{1}{r}\left[\frac{\partial}{\partial r}(r v_\theta) - \frac{\partial v_r}{\partial\theta}\right]\hat{\boldsymbol{\phi}}.\end{aligned} \tag{1.72}$$

Laplacian:

$$\nabla^2 T = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial T}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial T}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 T}{\partial\phi^2}. \tag{1.73}$$

For reference, these formulas are listed inside the front cover.

Problem 1.37 Find formulas for r, θ, ϕ in terms of x, y, z (the inverse, in other words, of Eq. 1.62).

• **Problem 1.38** Express the unit vectors $\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}$ in terms of $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ (that is, derive Eq. 1.64). Check your answers several ways ($\hat{\mathbf{r}}\cdot\hat{\mathbf{r}} \overset{?}{=} 1,\ \hat{\boldsymbol{\theta}}\cdot\hat{\boldsymbol{\phi}} \overset{?}{=} 0,\ \hat{\mathbf{r}}\times\hat{\boldsymbol{\theta}} \overset{?}{=} \hat{\boldsymbol{\phi}}, \ldots$). Also work out the inverse formulas, giving $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ in terms of $\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}$ (and θ, ϕ).

• **Problem 1.39**

(a) Check the divergence theorem for the function $\mathbf{v}_1 = r^2\hat{\mathbf{r}}$, using as your volume the sphere of radius R, centered at the origin.

(b) Do the same for $\mathbf{v}_2 = (1/r^2)\hat{\mathbf{r}}$. (If the answer surprises you, look back at Prob. 1.16.)

Problem 1.40 Compute the divergence of the function

$$\mathbf{v} = (r\cos\theta)\,\hat{\mathbf{r}} + (r\sin\theta)\,\hat{\boldsymbol{\theta}} + (r\sin\theta\cos\phi)\,\hat{\boldsymbol{\phi}}.$$

Check the divergence theorem for this function, using as your volume the inverted hemispherical bowl of radius R, resting on the xy plane and centered at the origin (Fig. 1.40).

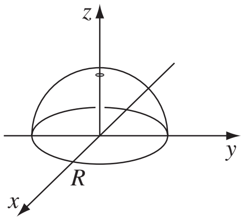


FIGURE 1.40

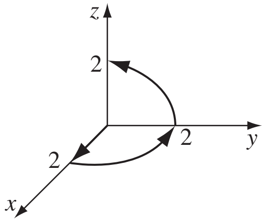


FIGURE 1.41

Problem 1.41 Compute the gradient and Laplacian of the function $T = r(\cos\theta + \sin\theta\cos\phi)$. Check the Laplacian by converting T to Cartesian coordinates and using Eq. 1.42. Test the gradient theorem for this function, using the path shown in Fig. 1.41, from (0, 0, 0) to (0, 0, 2).

1.4.2 ■ Cylindrical Coordinates

The cylindrical coordinates (s, ϕ, z) of a point P are defined in Fig. 1.42. Notice that ϕ has the same meaning as in spherical coordinates, and z is the same as Cartesian; s is the distance to P *from the* z *axis,* whereas the spherical coordinate r is the distance from the *origin*. The relation to Cartesian coordinates is

$$x = s\cos\phi, \qquad y = s\sin\phi, \qquad z = z. \tag{1.74}$$

The unit vectors (Prob. 1.42) are

$$\left.\begin{array}{rcl} \hat{\mathbf{s}} & = & \cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}, \\ \hat{\boldsymbol{\phi}} & = & -\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}}, \\ \hat{\mathbf{z}} & = & \hat{\mathbf{z}}. \end{array}\right\} \tag{1.75}$$

The infinitesimal displacements are

$$dl_s = ds, \qquad dl_\phi = s\,d\phi, \qquad dl_z = dz, \tag{1.76}$$

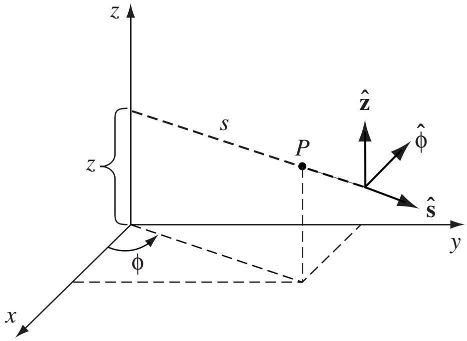


FIGURE 1.42

so

$$d\mathbf{l} = ds\,\hat{\mathbf{s}} + s\,d\phi\,\hat{\boldsymbol{\phi}} + dz\,\hat{\mathbf{z}}, \tag{1.77}$$

and the volume element is

$$d\tau = s\,ds\,d\phi\,dz. \tag{1.78}$$

The range of s is $0 \to \infty$, ϕ goes from $0 \to 2\pi$, and z from $-\infty$ to ∞.

The vector derivatives in cylindrical coordinates are:

Gradient:

$$\nabla T = \frac{\partial T}{\partial s}\,\hat{\mathbf{s}} + \frac{1}{s}\frac{\partial T}{\partial \phi}\,\hat{\boldsymbol{\phi}} + \frac{\partial T}{\partial z}\,\hat{\mathbf{z}}. \tag{1.79}$$

Divergence:

$$\nabla \cdot \mathbf{v} = \frac{1}{s}\frac{\partial}{\partial s}(s v_s) + \frac{1}{s}\frac{\partial v_\phi}{\partial \phi} + \frac{\partial v_z}{\partial z}. \tag{1.80}$$

Curl:

$$\nabla \times \mathbf{v} = \left(\frac{1}{s}\frac{\partial v_z}{\partial \phi} - \frac{\partial v_\phi}{\partial z}\right)\hat{\mathbf{s}} + \left(\frac{\partial v_s}{\partial z} - \frac{\partial v_z}{\partial s}\right)\hat{\boldsymbol{\phi}} + \frac{1}{s}\left[\frac{\partial}{\partial s}(s v_\phi) - \frac{\partial v_s}{\partial \phi}\right]\hat{\mathbf{z}}. \tag{1.81}$$

Laplacian:

$$\nabla^2 T = \frac{1}{s}\frac{\partial}{\partial s}\left(s\frac{\partial T}{\partial s}\right) + \frac{1}{s^2}\frac{\partial^2 T}{\partial \phi^2} + \frac{\partial^2 T}{\partial z^2}. \tag{1.82}$$

These formulas are also listed inside the front cover.

Problem 1.42 Express the cylindrical unit vectors $\hat{\mathbf{s}}$, $\hat{\boldsymbol{\phi}}$, $\hat{\mathbf{z}}$ in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ (that is, derive Eq. 1.75). "Invert" your formulas to get $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ in terms of $\hat{\mathbf{s}}$, $\hat{\boldsymbol{\phi}}$, $\hat{\mathbf{z}}$ (and ϕ).

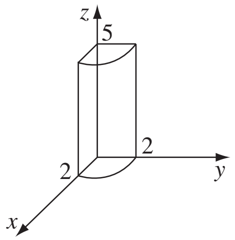


FIGURE 1.43

Problem 1.43

(a) Find the divergence of the function

$$\mathbf{v} = s(2 + \sin^2\phi)\,\hat{\mathbf{s}} + s\sin\phi\cos\phi\;\hat{\boldsymbol{\phi}} + 3z\;\hat{\mathbf{z}}.$$

(b) Test the divergence theorem for this function, using the quarter-cylinder (radius 2, height 5) shown in Fig. 1.43.

(c) Find the curl of **v**.

1.5 ■ THE DIRAC DELTA FUNCTION

1.5.1 ■ The Divergence of $\hat{\mathbf{r}}/r^2$

Consider the vector function

$$\mathbf{v} = \frac{1}{r^2}\,\hat{\mathbf{r}}. \tag{1.83}$$

At every location, **v** is directed radially outward (Fig. 1.44); if ever there was a function that ought to have a large positive divergence, this is it. And yet, when you actually *calculate* the divergence (using Eq. 1.71), you get precisely *zero*:

$$\nabla\cdot\mathbf{v} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{1}{r^2}\right) = \frac{1}{r^2}\frac{\partial}{\partial r}(1) = 0. \tag{1.84}$$

(You will have encountered this paradox already, if you worked Prob. 1.16.) The plot thickens when we apply the divergence theorem to this function. Suppose we integrate over a sphere of radius R, centered at the origin (Prob. 1.38b); the surface integral is

$$\begin{aligned}\oint \mathbf{v}\cdot d\mathbf{a} &= \int\left(\frac{1}{R^2}\hat{\mathbf{r}}\right)\cdot(R^2\sin\theta\,d\theta\,d\phi\,\hat{\mathbf{r}})\\ &= \left(\int_0^{\pi}\sin\theta\,d\theta\right)\left(\int_0^{2\pi}d\phi\right) = 4\pi.\end{aligned} \tag{1.85}$$

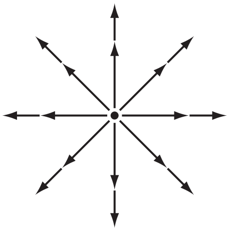


FIGURE 1.44

But the *volume* integral, $\int \nabla \cdot \mathbf{v}\, d\tau$, is *zero*, if we are really to believe Eq. 1.84. Does this mean that the divergence theorem is false? What's going on here?

The source of the problem is the point $r = 0$, where $\mathbf{v}$ blows up (and where, in Eq. 1.84, we have unwittingly divided by zero). It is quite true that $\nabla \cdot \mathbf{v} = 0$ everywhere *except* the origin, but right *at* the origin the situation is more complicated. Notice that the surface integral (Eq. 1.85) is *independent of R*; if the divergence theorem is right (and it *is*), we should get $\int(\nabla \cdot \mathbf{v})\, d\tau = 4\pi$ for *any* sphere centered at the origin, no matter how small. Evidently the entire contribution must be coming from the point $r = 0$! Thus, $\nabla \cdot \mathbf{v}$ has the bizarre property that it vanishes everywhere except at one point, and yet its *integral* (over any volume containing that point) is 4π. No ordinary function behaves like that. (On the other hand, a *physical* example *does* come to mind: the density (mass per unit volume) of a point particle. It's zero except at the exact location of the particle, and yet its *integral* is finite—namely, the mass of the particle.) What we have stumbled on is a mathematical object known to physicists as the **Dirac delta function**. It arises in many branches of theoretical physics. Moreover, the specific problem at hand (the divergence of the function $\hat{\mathbf{r}}/r^2$) is not just some arcane curiosity—it is, in fact, central to the whole theory of electrodynamics. So it is worthwhile to pause here and study the Dirac delta function with some care.

1.5.2 ■ The One-Dimensional Dirac Delta Function

The one-dimensional Dirac delta function, $\delta(x)$, can be pictured as an infinitely high, infinitesimally narrow "spike," with area 1 (Fig. 1.45). That is to say:

$$\delta(x) = \left\{ \begin{array}{ll} 0, & \text{if } x \neq 0 \\ \infty, & \text{if } x = 0 \end{array} \right\} \tag{1.86}$$

and[11]

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1. \tag{1.87}$$

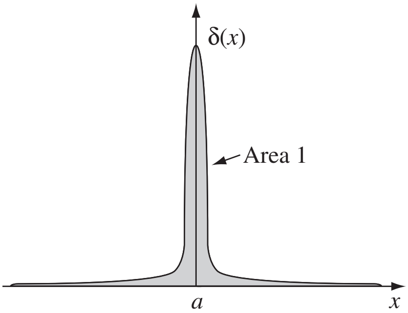
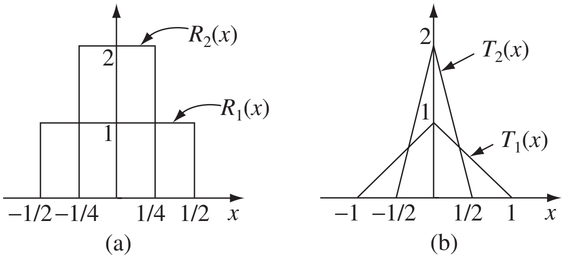


FIGURE 1.45

[11] Notice that the dimensions of $\delta(x)$ are one *over* the dimensions of its argument; if x is a length, $\delta(x)$ carries the units m^{-1}.



(a)

(b)

FIGURE 1.46

Technically, $\delta(x)$ is not a function at all, since its value is not finite at $x = 0$; in the mathematical literature it is known as a **generalized function**, or **distribution**. It is, if you like, the *limit* of a *sequence* of functions, such as rectangles $R_n(x)$, of height n and width $1/n$, or isosceles triangles $T_n(x)$, of height n and base $2/n$ (Fig. 1.46).

If $f(x)$ is some "ordinary" function (that is, *not* another delta function—in fact, just to be on the safe side, let's say that $f(x)$ is *continuous*), then the *product* $f(x)\delta(x)$ is zero everywhere except at $x = 0$. It follows that

$$f(x)\delta(x) = f(0)\delta(x). \tag{1.88}$$

(This is the most important fact about the delta function, so make sure you understand why it is true: since the product is zero anyway *except* at $x = 0$, we may as well replace $f(x)$ by the value it assumes at the origin.) In particular

$$\int_{-\infty}^{\infty} f(x)\delta(x)\,dx = f(0)\int_{-\infty}^{\infty} \delta(x)\,dx = f(0). \tag{1.89}$$

Under an integral, then, the delta function "picks out" the value of $f(x)$ at $x = 0$. (Here and below, the integral need not run from $-\infty$ to $+\infty$; it is sufficient that the domain extend across the delta function, and $-\epsilon$ to $+\epsilon$ would do as well.)

Of course, we can shift the spike from $x = 0$ to some other point, $x = a$ (Fig. 1.47):

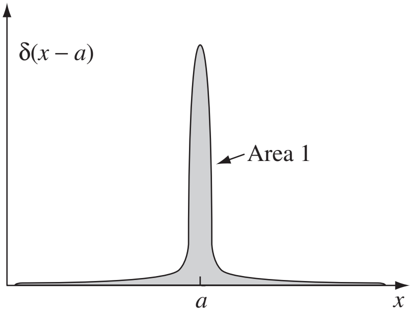


FIGURE 1.47

$$\delta(x-a) = \left\{ \begin{array}{ll} 0, & \text{if } x \neq a \\ \infty, & \text{if } x = a \end{array} \right\} \text{ with } \int_{-\infty}^{\infty} \delta(x-a)\, dx = 1. \tag{1.90}$$

Equation 1.88 becomes

$$f(x)\delta(x-a) = f(a)\delta(x-a), \tag{1.91}$$

and Eq. 1.89 generalizes to

$$\boxed{\int_{-\infty}^{\infty} f(x)\delta(x-a)\, dx = f(a).} \tag{1.92}$$

Example 1.14. Evaluate the integral

$$\int_0^3 x^3 \delta(x-2)\, dx.$$

Solution

The delta function picks out the value of x^3 at the point $x = 2$, so the integral is $2^3 = 8$. Notice, however, that if the upper limit had been 1 (instead of 3), the answer would be 0, because the spike would then be outside the domain of integration.

Although δ itself is not a legitimate function, *integrals* over δ are perfectly acceptable. In fact, it's best to think of the delta function as something that is *always intended for use under an integral sign.* In particular, two expressions involving delta functions (say, $D_1(x)$ and $D_2(x)$) are considered equal if [12]

$$\int_{-\infty}^{\infty} f(x) D_1(x)\, dx = \int_{-\infty}^{\infty} f(x) D_2(x)\, dx, \tag{1.93}$$

for all ("ordinary") functions $f(x)$.

Example 1.15. Show that

$$\delta(kx) = \frac{1}{|k|}\delta(x), \tag{1.94}$$

where k is any (nonzero) constant. (In particular, $\delta(-x) = \delta(x)$.)

[12] I emphasize that the integrals must be equal for *any* $f(x)$. Suppose $D_1(x)$ and $D_2(x)$ actually *differed,* say, in the neighborhood of the point $x = 17$. Then we could pick a function $f(x)$ that was sharply peaked about $x = 17$, and the integrals would not be equal.

Solution

For an arbitrary test function $f(x)$, consider the integral

$$\int_{-\infty}^{\infty} f(x)\delta(kx)\,dx.$$

Changing variables, we let $y \equiv kx$, so that $x = y/k$, and $dx = 1/k\,dy$. If k is positive, the integration still runs from $-\infty$ to $+\infty$, but if k is *negative,* then $x = \infty$ implies $y = -\infty$, and vice versa, so the order of the limits is reversed. Restoring the "proper" order costs a minus sign. Thus

$$\int_{-\infty}^{\infty} f(x)\delta(kx)\,dx = \pm\int_{-\infty}^{\infty} f(y/k)\delta(y)\frac{dy}{k} = \pm\frac{1}{k}f(0) = \frac{1}{|k|}f(0).$$

(The lower signs apply when k is negative, and we account for this neatly by putting absolute value bars around the final k, as indicated.) Under the integral sign, then, $\delta(kx)$ serves the same purpose as $(1/|k|)\delta(x)$:

$$\int_{-\infty}^{\infty} f(x)\delta(kx)\,dx = \int_{-\infty}^{\infty} f(x)\left[\frac{1}{|k|}\delta(x)\right]dx.$$

According to the criterion Eq. 1.93, therefore, $\delta(kx)$ and $(1/|k|)\delta(x)$ are equal.

Problem 1.44 Evaluate the following integrals:

(a) $\int_2^6 (3x^2 - 2x - 1)\,\delta(x-3)\,dx$.

(b) $\int_0^5 \cos x\,\delta(x-\pi)\,dx$.

(c) $\int_0^3 x^3\delta(x+1)\,dx$.

(d) $\int_{-\infty}^{\infty} \ln(x+3)\,\delta(x+2)\,dx$.

Problem 1.45 Evaluate the following integrals:

(a) $\int_{-2}^{2} (2x+3)\,\delta(3x)\,dx$.

(b) $\int_0^2 (x^3+3x+2)\,\delta(1-x)\,dx$.

(c) $\int_{-1}^{1} 9x^2\delta(3x+1)\,dx$.

(d) $\int_{-\infty}^{a} \delta(x-b)\,dx$.

Problem 1.46

(a) Show that

$$x\frac{d}{dx}(\delta(x)) = -\delta(x).$$

[*Hint:* Use integration by parts.]

(b) Let $\theta(x)$ be the **step function**:

$$\theta(x) \equiv \left\{ \begin{array}{ll} 1, & \text{if } x > 0 \\ & \\ 0, & \text{if } x \leq 0 \end{array} \right\}. \tag{1.95}$$

Show that $d\theta/dx = \delta(x)$.

1.5.3 ■ The Three-Dimensional Delta Function

It is easy to generalize the delta function to three dimensions:

$$\delta^3(\mathbf{r}) = \delta(x)\,\delta(y)\,\delta(z). \tag{1.96}$$

(As always, $\mathbf{r} \equiv x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$ is the position vector, extending from the origin to the point (x, y, z).) This three-dimensional delta function is zero everywhere except at (0, 0, 0), where it blows up. Its volume integral is 1:

$$\int_{\text{all space}} \delta^3(\mathbf{r})\,d\tau = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x)\,\delta(y)\,\delta(z)\,dx\,dy\,dz = 1. \tag{1.97}$$

And, generalizing Eq. 1.92,

$$\int_{\text{all space}} f(\mathbf{r})\delta^3(\mathbf{r} - \mathbf{a})\,d\tau = f(\mathbf{a}). \tag{1.98}$$

As in the one-dimensional case, integration with δ picks out the value of the function f at the location of the spike.

We are now in a position to resolve the paradox introduced in Sect. 1.5.1. As you will recall, we found that the divergence of $\hat{\mathbf{r}}/r^2$ is zero everywhere except at the origin, and yet its *integral* over any volume containing the origin is a constant (to wit: 4π). These are precisely the defining conditions for the Dirac delta function; evidently

$$\nabla \cdot \left(\frac{\hat{\mathbf{r}}}{r^2}\right) = 4\pi\delta^3(\mathbf{r}). \tag{1.99}$$

More generally,

$$\boxed{\nabla \cdot \left(\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\right) = 4\pi\delta^3(\boldsymbol{\mathscr{r}}),} \tag{1.100}$$

where, as always, $\boldsymbol{\mathscr{r}}$ is the separation vector: $\boldsymbol{\mathscr{r}} \equiv \mathbf{r} - \mathbf{r}'$. Note that differentiation here is with respect to $\mathbf{r}$, while $\mathbf{r}'$ is held constant. Incidentally, since

$$\nabla\left(\frac{1}{\mathscr{r}}\right) = -\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2} \tag{1.101}$$

(Prob. 1.13b), it follows that

$$\nabla^2 \frac{1}{\mathscr{r}} = -4\pi\delta^3(\boldsymbol{\mathscr{r}}). \tag{1.102}$$

Example 1.16. Evaluate the integral

$$J = \int_{\mathcal{V}} (r^2 + 2)\, \nabla \cdot \left(\frac{\hat{\mathbf{r}}}{r^2}\right) d\tau,$$

where $\mathcal{V}$ is a sphere[13] of radius R centered at the origin.

Solution 1
Use Eq. 1.99 to rewrite the divergence, and Eq. 1.98 to do the integral:

$$J = \int_{\mathcal{V}} (r^2 + 2) 4\pi\delta^3(\mathbf{r})\, d\tau = 4\pi(0 + 2) = 8\pi.$$

This one-line solution demonstrates something of the power and beauty of the delta function, but I would like to show you a second method, which is much more cumbersome but serves to illustrate the method of integration by parts (Sect. 1.3.6).

Solution 2
Using Eq. 1.59, we transfer the derivative from $\hat{\mathbf{r}}/r^2$ to $(r^2 + 2)$:

$$J = -\int_{\mathcal{V}} \frac{\hat{\mathbf{r}}}{r^2} \cdot [\nabla(r^2 + 2)]\, d\tau + \oint_{\mathcal{S}} (r^2 + 2) \frac{\hat{\mathbf{r}}}{r^2} \cdot d\mathbf{a}.$$

The gradient is

$$\nabla(r^2 + 2) = 2r\hat{\mathbf{r}},$$

so the volume integral becomes

$$\int \frac{2}{r}\, d\tau = \int \frac{2}{r} r^2 \sin\theta\, dr\, d\theta\, d\phi = 8\pi \int_0^R r\, dr = 4\pi R^2.$$

Meanwhile, on the boundary of the sphere (where $r = R$),

$$d\mathbf{a} = R^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{r}},$$

so the surface integral is

$$\int (R^2 + 2) \sin\theta\, d\theta\, d\phi = 4\pi(R^2 + 2).$$

[13] In proper mathematical jargon, "sphere" denotes the *surface*, and "ball" the volume it encloses. But physicists are (as usual) sloppy about this sort of thing, and I use the word "sphere" for both the surface and the volume. Where the meaning is not clear from the context, I will write "spherical surface" or "spherical volume." The language police tell me that the former is redundant and the latter an oxymoron, but a poll of my physics colleagues reveals that this is (for us) the standard usage.

Putting it all together,

$$J = -4\pi R^2 + 4\pi(R^2 + 2) = 8\pi,$$

as before.

Problem 1.47

(a) Write an expression for the volume charge density $\rho(\mathbf{r})$ of a point charge q at $\mathbf{r}'$. Make sure that the volume integral of ρ equals q.

(b) What is the volume charge density of an electric dipole, consisting of a point charge $-q$ at the origin and a point charge $+q$ at $\mathbf{a}$?

(c) What is the volume charge density (in spherical coordinates) of a uniform, infinitesimally thin spherical shell of radius R and total charge Q, centered at the origin? [*Beware:* the integral over all space must equal Q.]

Problem 1.48 Evaluate the following integrals:

(a) $\int(r^2 + \mathbf{r}\cdot\mathbf{a} + a^2)\delta^3(\mathbf{r}-\mathbf{a})\,d\tau$, where $\mathbf{a}$ is a fixed vector, a is its magnitude, and the integral is over all space.

(b) $\int_{\mathcal{V}} |\mathbf{r}-\mathbf{b}|^2\delta^3(5\mathbf{r})\,d\tau$, where $\mathcal{V}$ is a cube of side 2, centered on the origin, and $\mathbf{b} = 4\,\hat{\mathbf{y}} + 3\,\hat{\mathbf{z}}$.

(c) $\int_{\mathcal{V}} \left[r^4 + r^2(\mathbf{r}\cdot\mathbf{c}) + c^4\right]\delta^3(\mathbf{r}-\mathbf{c})\,d\tau$, where $\mathcal{V}$ is a sphere of radius 6 about the origin, $\mathbf{c} = 5\,\hat{\mathbf{x}} + 3\,\hat{\mathbf{y}} + 2\,\hat{\mathbf{z}}$, and c is its magnitude.

(d) $\int_{\mathcal{V}} \mathbf{r}\cdot(\mathbf{d}-\mathbf{r})\delta^3(\mathbf{e}-\mathbf{r})\,d\tau$, where $\mathbf{d} = (1, 2, 3)$, $\mathbf{e} = (3, 2, 1)$, and $\mathcal{V}$ is a sphere of radius 1.5 centered at (2, 2, 2).

Problem 1.49 Evaluate the integral

$$J = \int_{\mathcal{V}} e^{-r}\left(\nabla\cdot\frac{\hat{\mathbf{r}}}{r^2}\right)d\tau$$

(where $\mathcal{V}$ is a sphere of radius R, centered at the origin) by two different methods, as in Ex. 1.16.

1.6 ■ THE THEORY OF VECTOR FIELDS

1.6.1 ■ The Helmholtz Theorem

Ever since Faraday, the laws of electricity and magnetism have been expressed in terms of **electric** and **magnetic fields**, **E** and **B**. Like many physical laws, these are most compactly expressed as differential equations. Since **E** and **B** are *vectors*, the differential equations naturally involve vector derivatives: divergence and curl. Indeed, Maxwell reduced the entire theory to four equations, specifying respectively the divergence and the curl of **E** and **B**.

Maxwell's formulation raises an important mathematical question: To what extent is a vector function determined by its divergence and curl? In other words, if I tell you that the *divergence* of **F** (which stands for **E** or **B**, as the case may be) is a specified (scalar) function D,

$$\boldsymbol{\nabla} \cdot \mathbf{F} = D,$$

and the curl of **F** is a specified (vector) function **C**,

$$\boldsymbol{\nabla} \times \mathbf{F} = \mathbf{C},$$

(for consistency, **C** must be divergenceless,

$$\boldsymbol{\nabla} \cdot \mathbf{C} = 0,$$

because the divergence of a curl is always zero), can you then determine the function **F**?

Well... not quite. For example, as you may have discovered in Prob. 1.20, there are many functions whose divergence and curl are both zero everywhere—the trivial case $\mathbf{F} = \mathbf{0}$, of course, but also $\mathbf{F} = yz\,\hat{\mathbf{x}} + zx\,\hat{\mathbf{y}} + xy\,\hat{\mathbf{z}}$, $\mathbf{F} = \sin x \cosh y\,\hat{\mathbf{x}} - \cos x \sinh y\,\hat{\mathbf{y}}$, etc. To solve a differential equation you must also be supplied with appropriate **boundary conditions**. In electrodynamics we typically require that the fields go to zero "at infinity" (far away from all charges).[14] With that extra information, the **Helmholtz theorem** guarantees that the field is uniquely determined by its divergence and curl. (The Helmholtz theorem is discussed in Appendix B.)

1.6.2 ■ Potentials

If the curl of a vector field (**F**) vanishes (everywhere), then **F** can be written as the gradient of a **scalar potential** (V):

$$\boldsymbol{\nabla} \times \mathbf{F} = \mathbf{0} \Longleftrightarrow \mathbf{F} = -\boldsymbol{\nabla} V. \tag{1.103}$$

(The minus sign is purely conventional.) That's the essential burden of the following theorem:

Theorem 1
Curl-less (or "**irrotational**") **fields**. The following conditions are equivalent (that is, **F** satisfies one if and only if it satisfies all the others):

[14] In some textbook problems the charge itself extends to infinity (we speak, for instance, of the electric field of an infinite plane, or the magnetic field of an infinite wire). In such cases the normal boundary conditions do not apply, and one must invoke symmetry arguments to determine the fields uniquely.

(a) $\nabla \times \mathbf{F} = \mathbf{0}$ everywhere.

(b) $\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{F} \cdot d\mathbf{l}$ is independent of path, for any given end points.

(c) $\oint \mathbf{F} \cdot d\mathbf{l} = 0$ for any closed loop.

(d) $\mathbf{F}$ is the gradient of some scalar function: $\mathbf{F} = -\nabla V$.

The potential is not unique—any constant can be added to V with impunity, since this will not affect its gradient.

If the divergence of a vector field ($\mathbf{F}$) vanishes (everywhere), then $\mathbf{F}$ can be expressed as the curl of a **vector potential** ($\mathbf{A}$):

$$\nabla \cdot \mathbf{F} = 0 \Longleftrightarrow \mathbf{F} = \nabla \times \mathbf{A}. \tag{1.104}$$

That's the main conclusion of the following theorem:

Theorem 2

Divergence-less (or "**solenoidal**") **fields**. The following conditions are equivalent:

(a) $\nabla \cdot \mathbf{F} = 0$ everywhere.

(b) $\int \mathbf{F} \cdot d\mathbf{a}$ is independent of surface, for any given boundary line.

(c) $\oint \mathbf{F} \cdot d\mathbf{a} = 0$ for any closed surface.

(d) $\mathbf{F}$ is the curl of some vector function: $\mathbf{F} = \nabla \times \mathbf{A}$.

The vector potential is not unique—the gradient of any scalar function can be added to $\mathbf{A}$ without affecting the curl, since the curl of a gradient is zero.

You should by now be able to prove all the connections in these theorems, save for the ones that say (a), (b), or (c) implies (d). Those are more subtle, and will come later. Incidentally, in *all* cases (*whatever* its curl and divergence may be) a vector field $\mathbf{F}$ can be written as the gradient of a scalar plus the curl of a vector:[15]

$$\mathbf{F} = -\nabla V + \nabla \times \mathbf{A} \qquad \text{(always)}. \tag{1.105}$$

Problem 1.50

(a) Let $\mathbf{F}_1 = x^2\,\hat{\mathbf{z}}$ and $\mathbf{F}_2 = x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$. Calculate the divergence and curl of $\mathbf{F}_1$ and $\mathbf{F}_2$. Which one can be written as the gradient of a scalar? Find a scalar potential that does the job. Which one can be written as the curl of a vector? Find a suitable vector potential.

[15] In physics, the word **field** denotes generically any function of position (x, y, z) and time (t). But in electrodynamics two particular fields ($\mathbf{E}$ and $\mathbf{B}$) are of such paramount importance as to preempt the term. Thus technically the potentials are also "fields," but we never call them that.

(b) Show that $\mathbf{F}_3 = yz\,\hat{\mathbf{x}} + zx\,\hat{\mathbf{y}} + xy\,\hat{\mathbf{z}}$ can be written both as the gradient of a scalar and as the curl of a vector. Find scalar and vector potentials for this function.

Problem 1.51 For Theorem 1, show that (d) $\Rightarrow$ (a), (a) $\Rightarrow$ (c), (c) $\Rightarrow$ (b), (b) $\Rightarrow$ (c), and (c) $\Rightarrow$ (a).

Problem 1.52 For Theorem 2, show that (d) $\Rightarrow$ (a), (a) $\Rightarrow$ (c), (c) $\Rightarrow$ (b), (b) $\Rightarrow$ (c), and (c) $\Rightarrow$ (a).

Problem 1.53

(a) Which of the vectors in Problem 1.15 can be expressed as the gradient of a scalar? Find a scalar function that does the job.

(b) Which can be expressed as the curl of a vector? Find such a vector.

More Problems on Chapter 1

Problem 1.54 Check the divergence theorem for the function

$$\mathbf{v} = r^2 \cos\theta\,\hat{\mathbf{r}} + r^2\cos\phi\,\hat{\boldsymbol{\theta}} - r^2\cos\theta\sin\phi\,\hat{\boldsymbol{\phi}},$$

using as your volume one octant of the sphere of radius R (Fig. 1.48). Make sure you include the *entire* surface. [*Answer:* $\pi R^4/4$]

Problem 1.55 Check Stokes' theorem using the function $\mathbf{v} = ay\,\hat{\mathbf{x}} + bx\,\hat{\mathbf{y}}$ (a and b are constants) and the circular path of radius R, centered at the origin in the xy plane. [*Answer:* $\pi R^2(b-a)$]

Problem 1.56 Compute the line integral of

$$\mathbf{v} = 6\,\hat{\mathbf{x}} + yz^2\,\hat{\mathbf{y}} + (3y+z)\,\hat{\mathbf{z}}$$

along the triangular path shown in Fig. 1.49. Check your answer using Stokes' theorem. [*Answer:* 8/3]

Problem 1.57 Compute the line integral of

$$\mathbf{v} = (r\cos^2\theta)\,\hat{\mathbf{r}} - (r\cos\theta\sin\theta)\,\hat{\boldsymbol{\theta}} + 3r\,\hat{\boldsymbol{\phi}}$$

around the path shown in Fig. 1.50 (the points are labeled by their Cartesian coordinates). Do it either in cylindrical or in spherical coordinates. Check your answer, using Stokes' theorem. [*Answer:* $3\pi/2$]

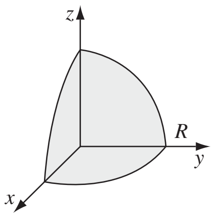


FIGURE 1.48

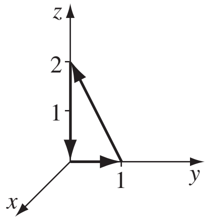


FIGURE 1.49

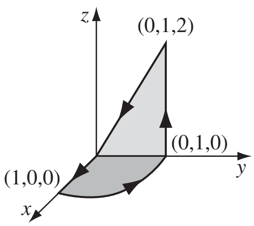


FIGURE 1.50

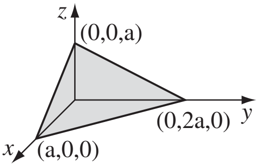


FIGURE 1.51

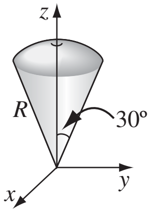


FIGURE 1.52

Problem 1.58 Check Stokes' theorem for the function $\mathbf{v} = y\,\hat{\mathbf{z}}$, using the triangular surface shown in Fig. 1.51. [*Answer:* a^2]

Problem 1.59 Check the divergence theorem for the function

$$\mathbf{v} = r^2 \sin\theta\,\hat{\mathbf{r}} + 4r^2\cos\theta\,\hat{\boldsymbol{\theta}} + r^2\tan\theta\,\hat{\boldsymbol{\phi}},$$

using the volume of the "ice-cream cone" shown in Fig. 1.52 (the top surface is spherical, with radius R and centered at the origin). [*Answer:* $(\pi R^4/12)(2\pi + 3\sqrt{3})$]

Problem 1.60 Here are two cute checks of the fundamental theorems:

(a) Combine Corollary 2 to the gradient theorem with Stokes' theorem ($\mathbf{v} = \boldsymbol{\nabla} T$, in this case). Show that the result is consistent with what you already knew about second derivatives.

(b) Combine Corollary 2 to Stokes' theorem with the divergence theorem. Show that the result is consistent with what you already knew.

• **Problem 1.61** Although the gradient, divergence, and curl theorems are the fundamental integral theorems of vector calculus, it is possible to derive a number of corollaries from them. Show that:

(a) $\int_{\mathcal{V}} (\boldsymbol{\nabla} T)\,d\tau = \oint_{\mathcal{S}} T\,d\mathbf{a}$. [*Hint:* Let $\mathbf{v} = \mathbf{c}T$, where $\mathbf{c}$ is a constant, in the divergence theorem; use the product rules.]

(b) $\int_{\mathcal{V}} (\boldsymbol{\nabla} \times \mathbf{v})\,d\tau = -\oint_{\mathcal{S}} \mathbf{v} \times\, d\mathbf{a}$. [*Hint:* Replace $\mathbf{v}$ by $(\mathbf{v} \times \mathbf{c})$ in the divergence theorem.]

(c) $\int_{\mathcal{V}} [T\nabla^2 U + (\boldsymbol{\nabla} T)\cdot(\boldsymbol{\nabla} U)]\,d\tau = \oint_{\mathcal{S}} (T\boldsymbol{\nabla} U)\cdot\, d\mathbf{a}$. [*Hint:* Let $\mathbf{v} = T\boldsymbol{\nabla} U$ in the divergence theorem.]

(d) $\int_{\mathcal{V}} (T\nabla^2 U - U\nabla^2 T)\,d\tau = \oint_{\mathcal{S}} (T\boldsymbol{\nabla} U - U\boldsymbol{\nabla} T)\cdot\, d\mathbf{a}$. [*Comment:* This is sometimes called **Green's second identity**; it follows from (c), which is known as **Green's identity**.]

(e) $\int_{\mathcal{S}} \boldsymbol{\nabla} T \times d\mathbf{a} = -\oint_{\mathcal{P}} T\,d\mathbf{l}$. [*Hint:* Let $\mathbf{v} = \mathbf{c}T$ in Stokes' theorem.]

• **Problem 1.62** The integral

$$\mathbf{a} \equiv \int_{\mathcal{S}} d\mathbf{a} \tag{1.106}$$

is sometimes called the **vector area** of the surface $\mathcal{S}$. If $\mathcal{S}$ happens to be *flat*, then $|\mathbf{a}|$ is the *ordinary* (scalar) area, obviously.

(a) Find the vector area of a hemispherical bowl of radius R.

(b) Show that $\mathbf{a} = \mathbf{0}$ for any *closed* surface. [*Hint:* Use Prob. 1.61a.]

(c) Show that $\mathbf{a}$ is the same for all surfaces sharing the same boundary.

(d) Show that

$$\mathbf{a} = \tfrac{1}{2} \oint \mathbf{r} \times d\mathbf{l}, \tag{1.107}$$

where the integral is around the boundary line. [*Hint:* One way to do it is to draw the cone subtended by the loop at the origin. Divide the conical surface up into infinitesimal triangular wedges, each with vertex at the origin and opposite side $d\mathbf{l}$, and exploit the geometrical interpretation of the cross product (Fig. 1.8).]

(e) Show that

$$\oint (\mathbf{c} \cdot \mathbf{r})\, d\mathbf{l} = \mathbf{a} \times \mathbf{c}, \tag{1.108}$$

for any constant vector $\mathbf{c}$. [*Hint:* Let $T = \mathbf{c} \cdot \mathbf{r}$ in Prob. 1.61e.]

• **Problem 1.63**

(a) Find the divergence of the function

$$\mathbf{v} = \frac{\hat{\mathbf{r}}}{r}.$$

First compute it directly, as in Eq. 1.84. Test your result using the divergence theorem, as in Eq. 1.85. Is there a delta function at the origin, as there was for $\hat{\mathbf{r}}/r^2$? What is the general formula for the divergence of $r^n\hat{\mathbf{r}}$? [*Answer:* $\nabla \cdot (r^n\hat{\mathbf{r}}) = (n+2)r^{n-1}$, unless $n = -2$, in which case it is $4\pi\delta^3(\mathbf{r})$; for $n < -2$, the divergence is ill-defined at the origin.]

(b) Find the *curl* of $r^n\hat{\mathbf{r}}$. Test your conclusion using Prob. 1.61b. [*Answer:* $\nabla \times (r^n\hat{\mathbf{r}}) = \mathbf{0}$]

Problem 1.64 In case you're not persuaded that $\nabla^2(1/r) = -4\pi\delta^3(\mathbf{r})$ (Eq. 1.102 with $\mathbf{r}' = \mathbf{0}$ for simplicity), try replacing r by $\sqrt{r^2+\epsilon^2}$, and watching what happens as $\epsilon \to 0$.[16] Specifically, let

$$D(r,\epsilon) \equiv -\frac{1}{4\pi}\nabla^2 \frac{1}{\sqrt{r^2+\epsilon^2}}.$$

[16] This problem was suggested by Frederick Strauch.

To demonstrate that this goes to $\delta^3(\mathbf{r})$ as $\epsilon \to 0$:

(a) Show that $D(r, \epsilon) = (3\epsilon^2/4\pi)(r^2 + \epsilon^2)^{-5/2}$.

(b) Check that $D(0, \epsilon) \to \infty$, as $\epsilon \to 0$.

(c) Check that $D(r, \epsilon) \to 0$, as $\epsilon \to 0$, for all $r \neq 0$.

(d) Check that the integral of $D(r, \epsilon)$ over all space is 1.

CHAPTER 7

Electrodynamics

7.1 ■ ELECTROMOTIVE FORCE

7.1.1 ■ Ohm's Law

To make a current flow, you have to *push* on the charges. How *fast* they move, in response to a given push, depends on the nature of the material. For most substances, the current density $\mathbf{J}$ is proportional to the *force per unit charge*, $\mathbf{f}$:

$$\mathbf{J} = \sigma \mathbf{f}. \tag{7.1}$$

The proportionality factor σ (not to be confused with surface charge) is an empirical constant that varies from one material to another; it's called the **conductivity** of the medium. Actually, the handbooks usually list the *reciprocal* of σ, called the **resistivity**: $\rho = 1/\sigma$ (not to be confused with charge density—I'm sorry, but we're running out of Greek letters, and this is the standard notation). Some typical values are listed in Table 7.1. Notice that even *insulators* conduct slightly, though the conductivity of a metal is astronomically greater; in fact, for most purposes metals can be regarded as **perfect conductors**, with $\sigma = \infty$, while for insulators we can pretend $\sigma = 0$.

In principle, the force that drives the charges to produce the current could be anything—chemical, gravitational, or trained ants with tiny harnesses. For *our* purposes, though, it's usually an electromagnetic force that does the job. In this case Eq. 7.1 becomes

$$\mathbf{J} = \sigma(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{7.2}$$

Ordinarily, the velocity of the charges is sufficiently small that the second term can be ignored:

$$\boxed{\mathbf{J} = \sigma \mathbf{E}.} \tag{7.3}$$

(However, in plasmas, for instance, the magnetic contribution to $\mathbf{f}$ can be significant.) Equation 7.3 is called **Ohm's law**, though the physics behind it is really contained in Eq. 7.1, of which 7.3 is just a special case.

I know: you're confused because I said $\mathbf{E} = \mathbf{0}$ inside a conductor (Sect. 2.5.1). But that's for *stationary* charges ($\mathbf{J} = \mathbf{0}$). Moreover, for *perfect* conductors

Material	Resistivity	Material	Resistivity
Conductors:		*Semiconductors:*	
Silver	1.59×10^{-8}	Sea water	0.2
Copper	1.68×10^{-8}	Germanium	0.46
Gold	2.21×10^{-8}	Diamond	2.7
Aluminum	2.65×10^{-8}	Silicon	2500
Iron	9.61×10^{-8}	*Insulators:*	
Mercury	9.61×10^{-7}	Water (pure)	8.3×10^{3}
Nichrome	1.08×10^{-6}	Glass	$10^{9} - 10^{14}$
Manganese	1.44×10^{-6}	Rubber	$10^{13} - 10^{15}$
Graphite	1.6×10^{-5}	Teflon	$10^{22} - 10^{24}$

TABLE 7.1 Resistivities, in ohm-meters (all values are for 1 atm, 20° C). *Data from Handbook of Chemistry and Physics,* 91st ed. (Boca Raton, Fla.: CRC Press, 2010) and other references.

$\mathbf{E} = \mathbf{J}/\sigma = \mathbf{0}$ even if current *is* flowing. In practice, metals are such good conductors that the electric field required to drive current in them is negligible. Thus we routinely treat the connecting wires in electric circuits (for example) as equipotentials. **Resistors**, by contrast, are made from *poorly* conducting materials.

Example 7.1. A cylindrical resistor of cross-sectional area A and length L is made from material with conductivity σ. (See Fig. 7.1; as indicated, the cross section need not be circular, but I *do* assume it is the same all the way down.) If we stipulate that the potential is constant over each end, and the potential difference between the ends is V, what current flows?

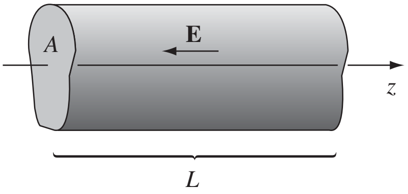


FIGURE 7.1

Solution

As it turns out, the electric field is *uniform* within the wire (I'll prove this in a moment). It follows from Eq. 7.3 that the current density is also uniform, so

$$I = JA = \sigma EA = \frac{\sigma A}{L} V.$$

Example 7.2. Two long coaxial metal cylinders (radii a and b) are separated by material of conductivity σ (Fig. 7.2). If they are maintained at a potential difference V, what current flows from one to the other, in a length L?

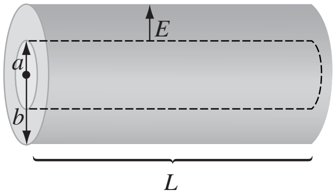


FIGURE 7.2

Solution

The field between the cylinders is

$$\mathbf{E} = \frac{\lambda}{2\pi\epsilon_0 s}\hat{\mathbf{s}},$$

where λ is the charge per unit length on the inner cylinder. The current is therefore

$$I = \int \mathbf{J}\cdot d\mathbf{a} = \sigma\int \mathbf{E}\cdot d\mathbf{a} = \frac{\sigma}{\epsilon_0}\lambda L.$$

(The integral is over any surface enclosing the inner cylinder.) Meanwhile, the potential difference between the cylinders is

$$V = -\int_b^a \mathbf{E}\cdot d\mathbf{l} = \frac{\lambda}{2\pi\epsilon_0}\ln\left(\frac{b}{a}\right),$$

so

$$I = \frac{2\pi\sigma L}{\ln(b/a)}V.$$

As these examples illustrate, the total current flowing from one **electrode** to the other is proportional to the potential difference between them:

$$\boxed{V = IR.} \qquad (7.4)$$

This, of course, is the more familiar version of Ohm's law. The constant of proportionality R is called the **resistance**; it's a function of the geometry of the arrangement and the conductivity of the medium between the electrodes. (In Ex. 7.1, $R = (L/\sigma A)$; in Ex. 7.2, $R = \ln(b/a)/2\pi\sigma L$.) Resistance is measured in **ohms** (Ω): an ohm is a volt per ampere. Notice that the proportionality between V and I

is a direct consequence of Eq. 7.3: if you want to double V, you simply double the charge on the electrodes—that doubles $\mathbf{E}$, which (for an ohmic material) doubles $\mathbf{J}$, which doubles I.

For *steady* currents and *uniform* conductivity,

$$\boldsymbol{\nabla} \cdot \mathbf{E} = \frac{1}{\sigma} \boldsymbol{\nabla} \cdot \mathbf{J} = 0, \tag{7.5}$$

(Eq. 5.33), and therefore the charge density is zero; any unbalanced charge resides on the surface. (We proved this long ago, for the case of *stationary* charges, using the fact that $\mathbf{E} = \mathbf{0}$; evidently, it is still true when the charges are allowed to move.) It follows, in particular, that Laplace's equation holds within a homogeneous ohmic material carrying a steady current, so all the tools and tricks of Chapter 3 are available for calculating the potential.

Example 7.3. I asserted that the field in Ex. 7.1 is *uniform*. Let's prove it.

Solution

Within the cylinder V obeys Laplace's equation. What are the boundary conditions? At the left end the potential is constant—we may as well set it equal to zero. At the right end the potential is likewise constant—call it V_0. On the cylindrical surface, $\mathbf{J} \cdot \hat{\mathbf{n}} = 0$, or else charge would be leaking out into the surrounding space (which we take to be nonconducting). Therefore $\mathbf{E} \cdot \hat{\mathbf{n}} = 0$, and hence $\partial V / \partial n = 0$. With V or its normal derivative specified on all surfaces, the potential is uniquely determined (Prob. 3.5). But it's easy to guess *one* potential that obeys Laplace's equation and fits these boundary conditions:

$$V(z) = \frac{V_0 z}{L},$$

where z is measured along the axis. The uniqueness theorem guarantees that this is *the* solution. The corresponding field is

$$\mathbf{E} = -\boldsymbol{\nabla} V = -\frac{V_0}{L} \hat{\mathbf{z}},$$

which is indeed uniform. □

Contrast the enormously more difficult problem that arises if the conducting material is removed, leaving only a metal plate at either end (Fig. 7.3). Evidently

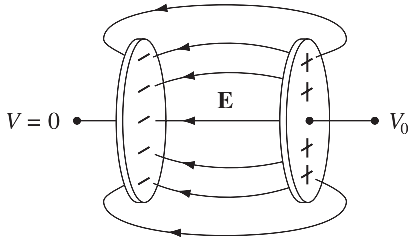


FIGURE 7.3

in the present case charge arranges itself over the surface of the wire in just such a way as to produce a nice uniform field within.[1]

I don't suppose there is any formula in physics more familiar than Ohm's law, and yet it's not really a true law, in the sense of Coulomb's or Ampère's; rather, it is a "rule of thumb" that applies pretty well to many substances. You're not going to win a Nobel prize for finding an exception. In fact, when you stop to think about it, it's a little surprising that Ohm's law *ever* holds. After all, a given field **E** produces a force $q\mathbf{E}$ (on a charge q), and according to Newton's second law, the charge will accelerate. But if the charges are *accelerating*, why doesn't the current *increase* with time, growing larger and larger the longer you leave the field on? Ohm's law implies, on the contrary, that a constant field produces a constant *current*, which suggests a constant *velocity*. Isn't that a contradiction to Newton's law?

No, for we are forgetting the frequent collisions electrons make as they pass down the wire. It's a little like this: Suppose you're driving down a street with a stop sign at every intersection, so that, although you accelerate constantly in between, you are obliged to start all over again with each new block. Your *average* speed is then a constant, in spite of the fact that (save for the periodic abrupt stops) you are always accelerating. If the length of a block is λ and your acceleration is a, the time it takes to go a block is

$$t = \sqrt{\frac{2\lambda}{a}},$$

and hence your average velocity is

$$v_{\text{ave}} = \frac{1}{2}at = \sqrt{\frac{\lambda a}{2}}.$$

But wait! That's no good *either!* It says that the velocity is proportional to the *square root* of the acceleration, and therefore that the current should be proportional to the square root of the field! There's another twist to the story: In practice, the charges are already moving very fast because of their thermal energy. But the thermal velocities have random directions, and average to zero. The **drift velocity** we are concerned with is a tiny extra bit (Prob. 5.20). So the time between collisions is actually much shorter than we supposed; if we assume for the sake of argument that all charges travel the same distance λ between collisions, then

$$t = \frac{\lambda}{v_{\text{thermal}}},$$

and therefore

$$v_{\text{ave}} = \frac{1}{2}at = \frac{a\lambda}{2v_{\text{thermal}}}.$$

[1] *Calculating* this surface charge is not easy. See, for example, J. D. Jackson, *Am. J. Phys.* **64**, 855 (1996). Nor is it a simple matter to determine the field *outside* the wire—see Prob. 7.43.

If there are n molecules per unit volume, and f free electrons per molecule, each with charge q and mass m, the current density is

$$\mathbf{J} = nfq\mathbf{v}_{\text{ave}} = \frac{nfq\lambda}{2v_{\text{thermal}}}\frac{\mathbf{F}}{m} = \left(\frac{nf\lambda q^2}{2mv_{\text{thermal}}}\right)\mathbf{E}. \tag{7.6}$$

I don't claim that the term in parentheses is an accurate formula for the conductivity,[2] but it does indicate the basic ingredients, and it correctly predicts that conductivity is proportional to the density of the moving charges and (ordinarily) decreases with increasing temperature.

As a result of all the collisions, the work done by the electrical force is converted into heat in the resistor. Since the work done per unit charge is V and the charge flowing per unit time is I, the power delivered is

$$\boxed{P = VI = I^2R.} \tag{7.7}$$

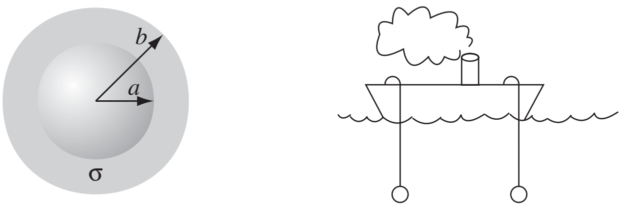
This is the **Joule heating law**. With I in amperes and R in ohms, P comes out in watts (joules per second).

Problem 7.1 Two concentric metal spherical shells, of radius a and b, respectively, are separated by weakly conducting material of conductivity σ (Fig. 7.4a).

(a) If they are maintained at a potential difference V, what current flows from one to the other?

(b) What is the resistance between the shells?

(c) Notice that if $b \gg a$ the outer radius (b) is irrelevant. How do you account for that? Exploit this observation to determine the current flowing between two metal spheres, each of radius a, immersed deep in the sea and held quite far apart (Fig. 7.4b), if the potential difference between them is V. (This arrangement can be used to measure the conductivity of sea water.)



(a) (b)

FIGURE 7.4

[2] This classical model (due to Drude) bears little resemblance to the modern quantum theory of conductivity. See, for instance, D. Park's *Introduction to the Quantum Theory*, 3rd ed., Chap. 15 (New York: McGraw-Hill, 1992).

Problem 7.2 A capacitor C has been charged up to potential V_0; at time $t = 0$, it is connected to a resistor R, and begins to discharge (Fig. 7.5a).

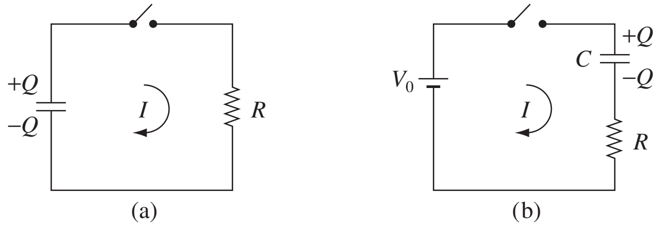


FIGURE 7.5

(a) Determine the charge on the capacitor as a function of time, $Q(t)$. What is the current through the resistor, $I(t)$?

(b) What was the original energy stored in the capacitor (Eq. 2.55)? By integrating Eq. 7.7, confirm that the heat delivered to the resistor is equal to the energy lost by the capacitor.

Now imagine *charging up* the capacitor, by connecting it (and the resistor) to a battery of voltage V_0, at time $t = 0$ (Fig. 7.5b).

(c) Again, determine $Q(t)$ and $I(t)$.

(d) Find the total energy output of the battery ($\int V_0 I\, dt$). Determine the heat delivered to the resistor. What is the final energy stored in the capacitor? What fraction of the work done by the battery shows up as energy in the capacitor? [Notice that the answer is independent of R!]

Problem 7.3

(a) Two metal objects are embedded in weakly conducting material of conductivity σ (Fig. 7.6). Show that the resistance between them is related to the capacitance of the arrangement by

$$R = \frac{\epsilon_0}{\sigma C}.$$

(b) Suppose you connected a battery between 1 and 2, and charged them up to a potential difference V_0. If you then disconnect the battery, the charge will gradually leak off. Show that $V(t) = V_0 e^{-t/\tau}$, and find the **time constant**, τ, in terms of ϵ_0 and σ.

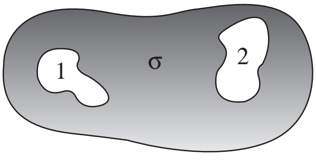


FIGURE 7.6

Problem 7.4 Suppose the conductivity of the material separating the cylinders in Ex. 7.2 is not uniform; specifically, $\sigma(s) = k/s$, for some constant k. Find the resistance between the cylinders. [*Hint:* Because σ is a function of position, Eq. 7.5 does not hold, the charge density is not zero in the resistive medium, and **E** does not go like $1/s$. But we *do* know that for steady currents I is the same across each cylindrical surface. Take it from there.]

7.1.2 ■ Electromotive Force

If you think about a typical electric circuit—a battery hooked up to a light bulb, say (Fig. 7.7)—a perplexing question arises: In practice, the *current is the same all the way around the loop*; why is this the case, when the only obvious driving force is inside the battery? Off hand, you might expect a large current in the battery and none at all in the lamp. Who's doing the pushing, in the rest of the circuit, and how does it happen that this push is exactly right to produce the same current in each segment? What's more, given that the charges in a typical wire move (literally) at a *snail's* pace (see Prob. 5.20), why doesn't it take half an hour for the current to reach the light bulb? How do all the charges know to start moving at the same instant?

Answer: If the current were *not* the same all the way around (for instance, during the first split second after the switch is closed), then charge would be piling up somewhere, and—here's the crucial point—the electric field of this accumulating charge is in such a direction as to even out the flow. Suppose, for instance, that the current *into* the bend in Fig. 7.8 is greater than the current *out*. Then charge piles up at the "knee," and this produces a field aiming *away* from the kink.[3] This field *opposes* the current flowing in (slowing it down) and *promotes* the current flowing out (speeding it up) until these currents are equal, at which point there is no further accumulation of charge, and equilibrium is established. It's a beautiful system, automatically self-correcting to keep the current uniform, and it does it all so quickly that, in practice, you can safely assume the current is the same all around the circuit, even in systems that oscillate at radio frequencies.

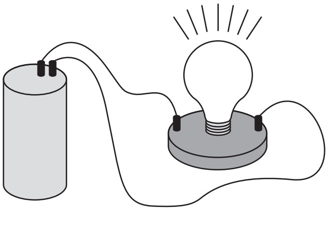


FIGURE 7.7

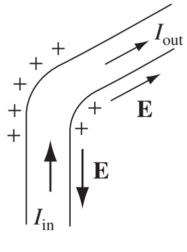


FIGURE 7.8

[3] The amount of charge involved is surprisingly small—see W. G. V. Rosser, *Am. J. Phys.* **38**, 265 (1970); nevertheless, the resulting field can be detected experimentally—see R. Jacobs, A. de Salazar, and A. Nassar, *Am. J. Phys.* **78**, 1432 (2010).

There are really *two* forces involved in driving current around a circuit: the *source*, $\mathbf{f}_s$, which is ordinarily confined to one portion of the loop (a battery, say), and an *electrostatic* force, which serves to smooth out the flow and communicate the influence of the source to distant parts of the circuit:

$$\mathbf{f} = \mathbf{f}_s + \mathbf{E}. \tag{7.8}$$

The physical agency responsible for $\mathbf{f}_s$ can be many different things: in a battery it's a chemical force; in a piezoelectric crystal mechanical pressure is converted into an electrical impulse; in a thermocouple it's a temperature gradient that does the job; in a photoelectric cell it's light; and in a Van de Graaff generator the electrons are literally loaded onto a conveyer belt and swept along. Whatever the *mechanism*, its net effect is determined by the line integral of **f** around the circuit:

$$\boxed{\mathcal{E} \equiv \oint \mathbf{f} \cdot d\mathbf{l} = \oint \mathbf{f}_s \cdot d\mathbf{l}.} \tag{7.9}$$

(Because $\oint \mathbf{E} \cdot d\mathbf{l} = 0$ for electrostatic fields, it doesn't matter whether you use **f** or $\mathbf{f}_s$.) $\mathcal{E}$ is called the **electromotive force**, or **emf**, of the circuit. It's a lousy term, since this is not a *force* at all—it's the *integral* of a *force per unit charge*. Some people prefer the word **electromotance**, but emf is so established that I think we'd better stick with it.

Within an ideal source of emf (a resistanceless battery,[4] for instance), the *net* force on the charges is *zero* (Eq. 7.1 with $\sigma = \infty$), so $\mathbf{E} = -\mathbf{f}_s$. The potential difference between the terminals (*a* and *b*) is therefore

$$V = -\int_a^b \mathbf{E} \cdot d\mathbf{l} = \int_a^b \mathbf{f}_s \cdot d\mathbf{l} = \oint \mathbf{f}_s \cdot d\mathbf{l} = \mathcal{E} \tag{7.10}$$

(we can extend the integral to the entire loop because $\mathbf{f}_s = \mathbf{0}$ outside the source). The function of a battery, then, is to establish and maintain a voltage difference equal to the electromotive force (a 6 V battery, for example, holds the positive terminal 6 V above the negative terminal). The resulting electrostatic field drives current around the rest of the circuit (notice, however, that *inside* the battery $\mathbf{f}_s$ drives current in the direction *opposite* to **E**).[5]

Because it's the line integral of $\mathbf{f}_s$, $\mathcal{E}$ can be interpreted as the *work done per unit charge*, by the source—indeed, in some books electromotive force is *defined* this way. However, as you'll see in the next section, there is some subtlety involved in this interpretation, so I prefer Eq. 7.9.

[4] *Real* batteries have a certain **internal resistance**, r, and the potential difference between their terminals is $\mathcal{E} - Ir$, when a current I is flowing. For an illuminating discussion of how batteries work, see D. Roberts, *Am. J. Phys.* **51**, 829 (1983).

[5] Current in an electric circuit is somewhat analogous to the flow of water in a closed system of pipes, with gravity playing the role of the electrostatic field, and a pump (lifting the water up *against* gravity) in the role of the battery. In this story *height* is analogous to voltage.

Problem 7.5 A battery of emf $\mathcal{E}$ and internal resistance r is hooked up to a variable "load" resistance, R. If you want to deliver the maximum possible power to the load, what resistance R should you choose? (You can't change $\mathcal{E}$ and r, of course.)

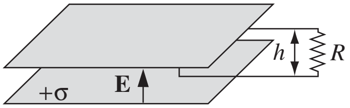


FIGURE 7.9

Problem 7.6 A rectangular loop of wire is situated so that one end (height h) is between the plates of a parallel-plate capacitor (Fig. 7.9), oriented parallel to the field $\mathbf{E}$. The other end is way outside, where the field is essentially zero. What is the emf in this loop? If the total resistance is R, what current flows? Explain. [*Warning:* This is a trick question, so be careful; if you have invented a perpetual motion machine, there's probably something wrong with it.]

7.1.3 ■ Motional emf

In the last section, I listed several possible sources of electromotive force, batteries being the most familiar. But I did not mention the commonest one of all: the **generator**. Generators exploit **motional emfs**, which arise when you *move a wire through a magnetic field*. Figure 7.10 suggests a primitive model for a generator. In the shaded region there is a uniform magnetic field $\mathbf{B}$, pointing into the page, and the resistor R represents whatever it is (maybe a light bulb or a toaster) we're trying to drive current through. If the entire loop is pulled to the right with speed v, the charges in segment ab experience a magnetic force whose vertical component qvB drives current around the loop, in the clockwise direction. The emf is

$$\mathcal{E} = \oint \mathbf{f}_{\text{mag}} \cdot d\mathbf{l} = vBh, \tag{7.11}$$

where h is the width of the loop. (The horizontal segments bc and ad contribute nothing, since the force there is perpendicular to the wire.)

Notice that the integral you perform to calculate $\mathcal{E}$ (Eq. 7.9 or 7.11) is carried out at *one instant of time*—take a "snapshot" of the loop, if you like, and work

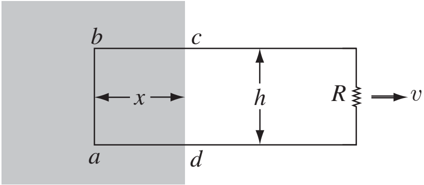


FIGURE 7.10

from that. Thus $d\mathbf{l}$, for the segment *ab* in Fig. 7.10, points straight up, even though the loop is moving to the right. You can't quarrel with this—it's simply the way emf is *defined*—but it is important to be clear about it.

In particular, although the magnetic force is responsible for establishing the emf, it is *not* doing any work—magnetic forces *never* do work. Who, then, *is* supplying the energy that heats the resistor? *Answer:* The person who's pulling on the loop. With the current flowing, the free charges in segment *ab* have a vertical velocity (call it $\mathbf{u}$) in addition to the horizontal velocity $\mathbf{v}$ they inherit from the motion of the loop. Accordingly, the magnetic force has a component quB to the left. To counteract this, the person pulling on the wire must exert a force per unit charge

$$f_{\text{pull}} = uB$$

to the *right* (Fig. 7.11). This force is transmitted to the charge by the structure of the wire.

Meanwhile, the particle is actually *moving* in the direction of the resultant velocity $\mathbf{w}$, and the distance it goes is $(h/\cos\theta)$. The work done per unit charge is therefore

$$\int \mathbf{f}_{\text{pull}} \cdot d\mathbf{l} = (uB)\left(\frac{h}{\cos\theta}\right)\sin\theta = vBh = \mathcal{E}$$

($\sin\theta$ coming from the dot product). As it turns out, then, the *work done per unit charge is exactly equal to the emf*, though the integrals are taken along entirely different paths (Fig. 7.12), and completely different forces are involved. To calculate the emf, you integrate around the loop at *one instant*, but to calculate the work done you follow a charge in its journey around the loop; $\mathbf{f}_{\text{pull}}$ contributes nothing to the emf, because it is perpendicular to the wire, whereas $\mathbf{f}_{\text{mag}}$ contributes nothing to work because it is perpendicular to the motion of the charge.[6]

There is a particularly nice way of expressing the emf generated in a moving loop. Let Φ be the flux of $\mathbf{B}$ through the loop:

$$\Phi \equiv \int \mathbf{B} \cdot d\mathbf{a}. \tag{7.12}$$

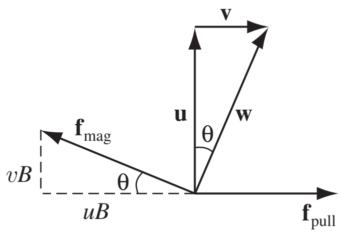
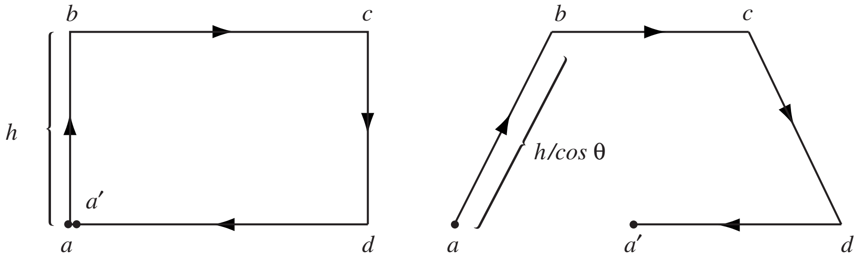


FIGURE 7.11

[6] For further discussion, see E. P. Mosca, *Am. J. Phys.* **42**, 295 (1974).



(a) Integration path for computing $\mathcal{E}$ (follow the wire at one instant of time).

(b) Integration path for calculating work done (follow the charge around the loop).

FIGURE 7.12

For the rectangular loop in Fig. 7.10,

$$\Phi = Bhx.$$

As the loop moves, the flux decreases:

$$\frac{d\Phi}{dt} = Bh\frac{dx}{dt} = -Bhv.$$

(The minus sign accounts for the fact that dx/dt is negative.) But this is precisely the emf (Eq. 7.11); evidently the emf generated in the loop is minus the rate of change of flux through the loop:

$$\boxed{\mathcal{E} = -\frac{d\Phi}{dt}.} \tag{7.13}$$

This is the **flux rule** for motional emf.

Apart from its delightful simplicity, the flux rule has the virtue of applying to *non*rectangular loops moving in *arbitrary* directions through *non*uniform magnetic fields; in fact, the loop need not even maintain a fixed shape.

Proof. Figure 7.13 shows a loop of wire at time t, and also a short time dt later. Suppose we compute the flux at time t, using surface $\mathcal{S}$, and the flux at time $t + dt$, using the surface consisting of $\mathcal{S}$ plus the "ribbon" that connects the new position of the loop to the old. The *change* in flux, then, is

$$d\Phi = \Phi(t + dt) - \Phi(t) = \Phi_{\text{ribbon}} = \int_{\text{ribbon}} \mathbf{B} \cdot d\mathbf{a}.$$

Focus your attention on point P: in time dt, it moves to P'. Let $\mathbf{v}$ be the velocity of the *wire*, and $\mathbf{u}$ the velocity of a charge *down* the wire; $\mathbf{w} = \mathbf{v} + \mathbf{u}$ is the resultant

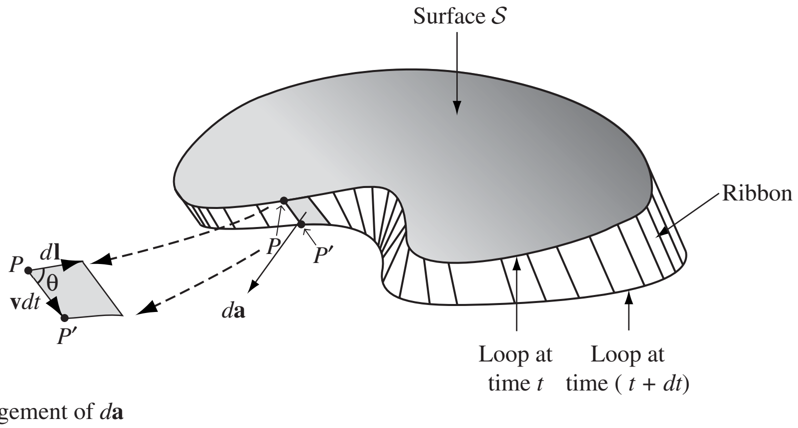


FIGURE 7.13

velocity of a charge at P. The infinitesimal element of area on the ribbon can be written as

$$d\mathbf{a} = (\mathbf{v} \times d\mathbf{l})\, dt$$

(see inset in Fig. 7.13). Therefore

$$\frac{d\Phi}{dt} = \oint \mathbf{B} \cdot (\mathbf{v} \times d\mathbf{l}).$$

Since $\mathbf{w} = (\mathbf{v} + \mathbf{u})$ and $\mathbf{u}$ is parallel to $d\mathbf{l}$, we can just as well write this as

$$\frac{d\Phi}{dt} = \oint \mathbf{B} \cdot (\mathbf{w} \times d\mathbf{l}).$$

Now, the scalar triple-product can be rewritten:

$$\mathbf{B} \cdot (\mathbf{w} \times d\mathbf{l}) = -(\mathbf{w} \times \mathbf{B}) \cdot d\mathbf{l},$$

so

$$\frac{d\Phi}{dt} = -\oint (\mathbf{w} \times \mathbf{B}) \cdot d\mathbf{l}.$$

But $(\mathbf{w} \times \mathbf{B})$ is the magnetic force per unit charge, $\mathbf{f}_{\text{mag}}$, so

$$\frac{d\Phi}{dt} = -\oint \mathbf{f}_{\text{mag}} \cdot d\mathbf{l},$$

and the integral of $\mathbf{f}_{\text{mag}}$ is the emf:

$$\mathcal{E} = -\frac{d\Phi}{dt}. \qquad \square$$

There is a sign ambiguity in the definition of emf (Eq. 7.9): Which *way* around the loop are you supposed to integrate? There is a compensatory ambiguity in the definition of *flux* (Eq. 7.12): Which is the positive direction for $d\mathbf{a}$? In applying

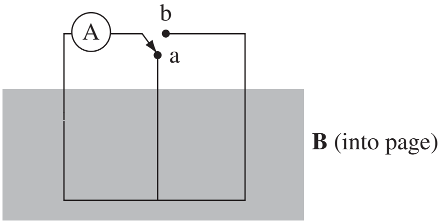


FIGURE 7.14

the flux rule, sign consistency is governed (as always) by your right hand: If your fingers define the positive direction around the loop, then your thumb indicates the direction of $d\mathbf{a}$. Should the emf come out negative, it means the current will flow in the negative direction around the circuit.

The flux rule is a nifty short-cut for calculating motional emfs. It does not contain any new physics—just the Lorentz force law. But it can lead to error or ambiguity if you're not careful. The flux rule assumes you have a single wire loop—it can move, rotate, stretch, or distort (continuously), but beware of switches, sliding contacts, or extended conductors allowing a variety of current paths. A standard "flux rule paradox" involves the circuit in Figure 7.14. When the switch is thrown (from a to b) the flux through the circuit doubles, but there's no motional emf (no conductor moving through a magnetic field), and the ammeter (A) records no current.

Example 7.4. A metal disk of radius a rotates with angular velocity ω about a vertical axis, through a uniform field $\mathbf{B}$, pointing up. A circuit is made by connecting one end of a resistor to the axle and the other end to a sliding contact, which touches the outer edge of the disk (Fig. 7.15). Find the current in the resistor.

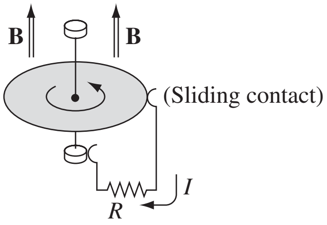


FIGURE 7.15

Solution

The speed of a point on the disk at a distance s from the axis is $v = \omega s$, so the force per unit charge is $\mathbf{f}_{\text{mag}} = \mathbf{v} \times \mathbf{B} = \omega s B \hat{\mathbf{s}}$. The emf is therefore

$$\mathcal{E} = \int_0^a f_{\text{mag}}\, ds = \omega B \int_0^a s\, ds = \frac{\omega B a^2}{2},$$

and the current is

$$I = \frac{\mathcal{E}}{R} = \frac{\omega B a^2}{2R}.$$

Example 7.4 (the **Faraday disk**, or **Faraday dynamo**) involves a motional emf that you can't calculate (at least, not directly) from the flux rule. The flux rule assumes the current flows along a well-defined path, whereas in this example the current spreads out over the whole disk. It's not even clear what the "flux through the circuit" would *mean* in this context.

Even more tricky is the case of **eddy currents**. Take a chunk of aluminum (say), and shake it around in a nonuniform magnetic field. Currents will be generated in the material, and you will feel a kind of "viscous drag"—as though you were pulling the block through molasses (this is the force I called $\mathbf{f}_{\text{pull}}$ in the discussion of motional emf). Eddy currents are notoriously difficult to calculate,[7] but easy and dramatic to demonstrate. You may have witnessed the classic experiment in which an aluminum disk mounted as a pendulum on a horizontal axis swings down and passes between the poles of a magnet (Fig. 7.16a). When it enters the field region it suddenly slows way down. To confirm that eddy currents are responsible, one repeats the demonstration using a disk that has many slots cut in it, to prevent the flow of large-scale currents (Fig. 7.16b). This time the disk swings freely, unimpeded by the field.

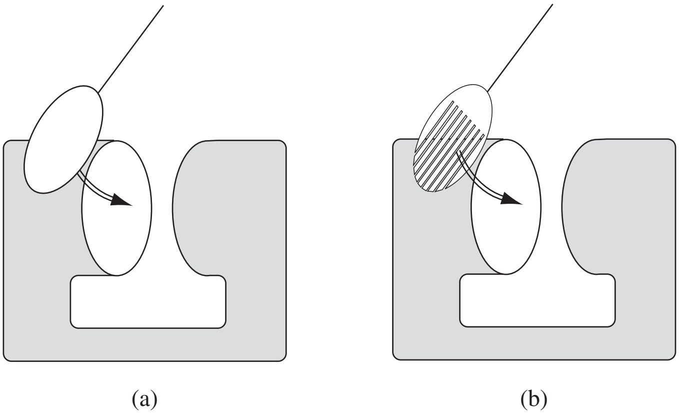


FIGURE 7.16

Problem 7.7 A metal bar of mass m slides frictionlessly on two parallel conducting rails a distance l apart (Fig. 7.17). A resistor R is connected across the rails, and a uniform magnetic field $\mathbf{B}$, pointing into the page, fills the entire region.

[7] See, for example, W. M. Saslow, *Am. J. Phys.*, **60**, 693 (1992).

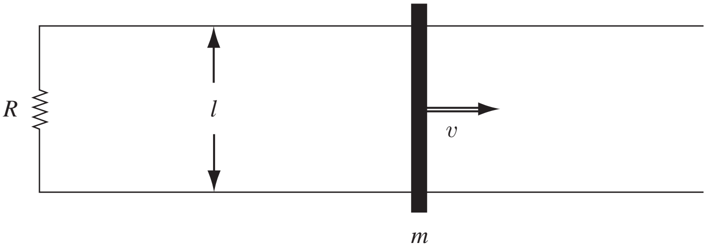


FIGURE 7.17

(a) If the bar moves to the right at speed v, what is the current in the resistor? In what direction does it flow?

(b) What is the magnetic force on the bar? In what direction?

(c) If the bar starts out with speed v_0 at time $t = 0$, and is left to slide, what is its speed at a later time t?

(d) The initial kinetic energy of the bar was, of course, $\frac{1}{2}mv_0{}^2$. Check that the energy delivered to the resistor is exactly $\frac{1}{2}mv_0{}^2$.

Problem 7.8 A square loop of wire (side a) lies on a table, a distance s from a very long straight wire, which carries a current I, as shown in Fig. 7.18.

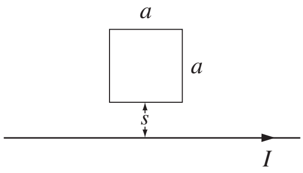


FIGURE 7.18

(a) Find the flux of **B** through the loop.

(b) If someone now pulls the loop directly away from the wire, at speed v, what emf is generated? In what direction (clockwise or counterclockwise) does the current flow?

(c) What if the loop is pulled to the *right* at speed v?

Problem 7.9 An infinite number of different surfaces can be fit to a given boundary line, and yet, in defining the magnetic flux through a loop, $\Phi = \int \mathbf{B} \cdot d\mathbf{a}$, I never specified the particular surface to be used. Justify this apparent oversight.

Problem 7.10 A square loop (side a) is mounted on a vertical shaft and rotated at angular velocity ω (Fig. 7.19). A uniform magnetic field **B** points to the right. Find the $\mathcal{E}(t)$ for this **alternating current** generator.

Problem 7.11 A square loop is cut out of a thick sheet of aluminum. It is then placed so that the top portion is in a uniform magnetic field **B**, and is allowed to fall under gravity (Fig. 7.20). (In the diagram, shading indicates the field region; **B** points into the page.) If the magnetic field is 1 T (a pretty standard laboratory field), find the terminal velocity of the loop (in m/s). Find the velocity of the loop as a function of time. How long does it take (in seconds) to reach, say, 90% of the terminal velocity? What would happen if you cut a tiny slit in the ring, breaking the circuit? [*Note:* The dimensions of the loop cancel out; determine the actual *numbers*, in the units indicated.]

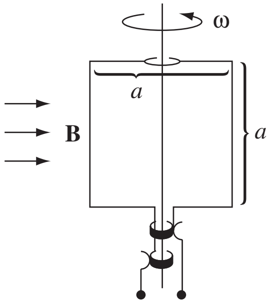


FIGURE 7.19

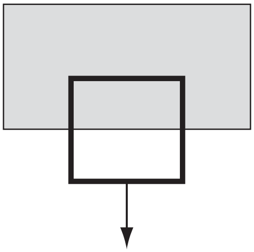


FIGURE 7.20

7.2 ■ ELECTROMAGNETIC INDUCTION

7.2.1 ■ Faraday's Law

In 1831 Michael Faraday reported on a series of experiments, including three that (with some violence to history) can be characterized as follows:

Experiment 1. He pulled a loop of wire to the right through a magnetic field (Fig. 7.21a). A current flowed in the loop.

Experiment 2. He moved the *magnet* to the *left*, holding the loop still (Fig. 7.21b). Again, a current flowed in the loop.

Experiment 3. With both the loop and the magnet at rest (Fig. 7.21c), he changed the *strength* of the field (he used an electromagnet, and varied the current in the coil). Once again, current flowed in the loop.

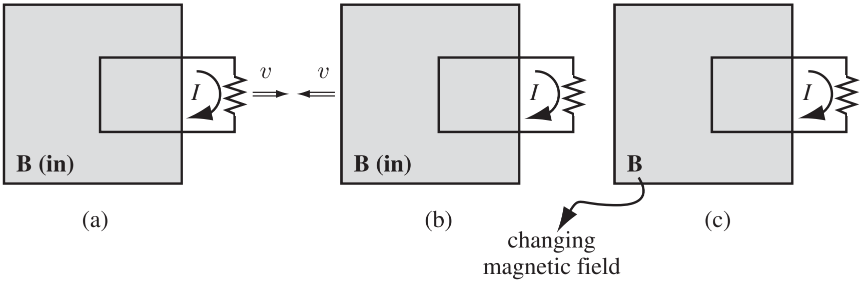


FIGURE 7.21

The first experiment, of course, is a straightforward case of motional emf; according to the flux rule:

$$\mathcal{E} = -\frac{d\Phi}{dt}.$$

I don't think it will surprise you to learn that exactly the same emf arises in Experiment 2—all that really matters is the *relative* motion of the magnet and the loop. Indeed, in the light of special relativity it *has* to be so. But Faraday knew nothing of relativity, and in classical electrodynamics this simple reciprocity is a remarkable coincidence. For if the *loop* moves, it's a *magnetic* force that sets up the emf, but if the loop is *stationary*, the force *cannot* be magnetic—stationary charges experience no magnetic forces. In that case, what *is* responsible? What sort of field exerts a force on charges at rest? Well, *electric* fields do, of course, but in this case there doesn't seem to be any electric field in sight.

Faraday had an ingenious inspiration:

A changing magnetic field induces an electric field.

It is this induced[8] electric field that accounts for the emf in Experiment 2.[9] Indeed, if (as Faraday found empirically) the emf is again equal to the rate of change of the flux,

$$\mathcal{E} = \oint \mathbf{E} \cdot d\mathbf{l} = -\frac{d\Phi}{dt}, \tag{7.14}$$

then **E** is related to the change in **B** by the equation

$$\oint \mathbf{E} \cdot d\mathbf{l} = -\int \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a}. \tag{7.15}$$

This is **Faraday's law**, in integral form. We can convert it to differential form by applying Stokes' theorem:

$$\boxed{\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}.} \tag{7.16}$$

[8] "Induce" is a subtle and slippery verb. It carries a faint odor of *causation* ("*pro*duce" would make this explicit) without quite committing itself. There is a sterile ongoing debate in the literature as to whether a changing magnetic field should be regarded as an independent "source" of electric fields (along with electric charge)—after all, the magnetic field *itself* is due to electric currents. It's like asking whether the postman is the "source" of my mail. Well, sure—he delivered it to my door. On the other hand, Grandma wrote the letter. Ultimately, ρ and **J** are the sources of *all* electromagnetic fields, and a changing magnetic field merely delivers electromagnetic news from currents elsewhere. But it is often convenient to think of a changing magnetic field "producing" an electric field, and it won't hurt you as long as you understand that this is the condensed version of a more complicated story. For a nice discussion, see S. E. Hill, *Phys. Teach*. **48**, 410 (2010).

[9] You might argue that the magnetic field in Experiment 2 is not really *changing*—just *moving*. What I mean is that if you sit at a *fixed location*, the field you experience changes as the magnet passes by.

Note that Faraday's law reduces to the old rule $\oint \mathbf{E} \cdot d\mathbf{l} = 0$ (or, in differential form, $\boldsymbol{\nabla} \times \mathbf{E} = \mathbf{0}$) in the static case (constant $\mathbf{B}$) as, of course, it should.

In Experiment 3, the magnetic field changes for entirely different reasons, but according to Faraday's law an electric field will again be induced, giving rise to an emf $-d\Phi/dt$. Indeed, one can subsume all three cases (and for that matter any combination of them) into a kind of **universal flux rule**:

> **Whenever (and for whatever reason) the magnetic flux through a loop changes, an emf**
>
> $$\mathcal{E} = -\frac{d\Phi}{dt} \tag{7.17}$$
>
> **will appear in the loop**.

Many people call *this* "Faraday's law." Maybe I'm overly fastidious, but I find this confusing. There are really *two* totally different mechanisms underlying Eq. 7.17, and to identify them both as "Faraday's law" is a little like saying that because identical twins look alike we ought to call them by the same name. In Faraday's first experiment it's the Lorentz force law at work; the emf is *magnetic*. But in the other two it's an *electric* field (induced by the changing magnetic field) that does the job. Viewed in this light, it is quite astonishing that all three processes yield the same formula for the emf. In fact, it was precisely this "coincidence" that led Einstein to the special theory of relativity—he sought a deeper understanding of what is, in classical electrodynamics, a peculiar accident. But that's a story for Chapter 12. In the meantime, I shall reserve the term "Faraday's law" for electric fields induced by changing magnetic fields, and I do *not* regard Experiment 1 as an instance of Faraday's law.

Example 7.5. A long cylindrical magnet of length L and radius a carries a uniform magnetization $\mathbf{M}$ parallel to its axis. It passes at constant velocity v through a circular wire ring of slightly larger diameter (Fig. 7.22). Graph the emf induced in the ring, as a function of time.

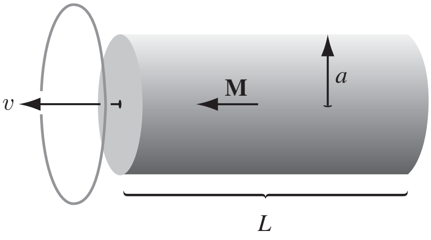


FIGURE 7.22

Solution

The magnetic field is the same as that of a long solenoid with surface current $\mathbf{K}_b = M\,\hat{\boldsymbol{\phi}}$. So the field inside is $\mathbf{B} = \mu_0\mathbf{M}$, except near the ends, where it starts to spread out. The flux through the ring is zero when the magnet is far away; it builds up to a maximum of $\mu_0 M\pi a^2$ as the leading end passes through; and it drops back to zero as the trailing end emerges (Fig. 7.23a). The emf is (minus) the derivative of Φ with respect to time, so it consists of two spikes, as shown in Fig. 7.23b.

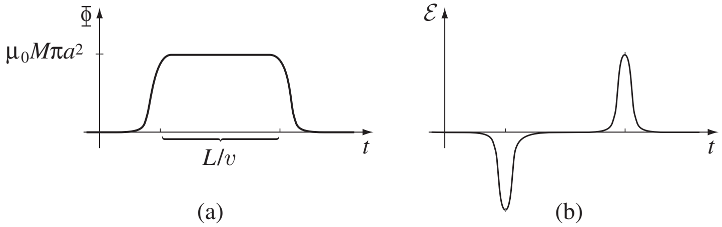


FIGURE 7.23

Keeping track of the *signs* in Faraday's law can be a real headache. For instance, in Ex. 7.5 we would like to know which *way* around the ring the induced current flows. In principle, the right-hand rule does the job (we called Φ positive to the left, in Fig. 7.22, so the positive direction for current in the ring is counterclockwise, as viewed from the left; since the first spike in Fig. 7.23b is *negative*, the first current pulse flows *clockwise*, and the second counterclockwise). But there's a handy rule, called **Lenz's law**, whose sole purpose is to help you get the directions right:[10]

Nature abhors a change in flux.

The induced current will flow in such a direction that the flux *it* produces tends to cancel the change. (As the front end of the magnet in Ex. 7.5 enters the ring, the flux increases, so the current in the ring must generate a field to the *right*—it therefore flows *clockwise*.) Notice that it is the *change* in flux, not the flux itself, that nature abhors (when the tail end of the magnet exits the ring, the flux *drops*, so the induced current flows *counterclockwise*, in an effort to restore it). Faraday induction is a kind of "inertial" phenomenon: A conducting loop "likes" to maintain a constant flux through it; if you try to *change* the flux, the loop responds by sending a current around in such a direction as to frustrate your efforts. (It doesn't *succeed* completely; the flux produced by the induced current is typically only a tiny fraction of the original. All Lenz's law tells you is the *direction* of the flow.)

[10] Lenz's law applies to *motional* emfs, too, but for them it is usually easier to get the direction of the current from the Lorentz force law.

Example 7.6. The "jumping ring" demonstration. If you wind a solenoidal coil around an iron core (the iron is there to beef up the magnetic field), place a metal ring on top, and plug it in, the ring will jump several feet in the air (Fig. 7.24). Why?

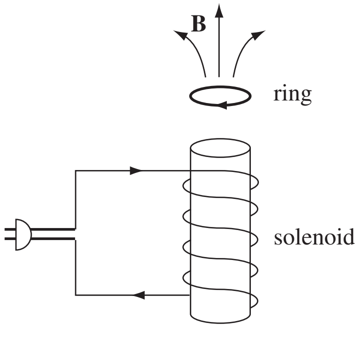


FIGURE 7.24

Solution

Before you turned on the current, the flux through the ring was *zero*. *Afterward* a flux appeared (upward, in the diagram), and the emf generated in the ring led to a current (in the ring) which, according to Lenz's law, was in such a direction that *its* field tended to cancel this new flux. This means that the current in the loop is *opposite* to the current in the solenoid. And opposite currents repel, so the ring flies off.[11]

Problem 7.12 A long solenoid, of radius a, is driven by an alternating current, so that the field inside is sinusoidal: $\mathbf{B}(t) = B_0 \cos(\omega t)\,\hat{\mathbf{z}}$. A circular loop of wire, of radius $a/2$ and resistance R, is placed inside the solenoid, and coaxial with it. Find the current induced in the loop, as a function of time.

Problem 7.13 A square loop of wire, with sides of length a, lies in the first quadrant of the xy plane, with one corner at the origin. In this region, there is a nonuniform time-dependent magnetic field $\mathbf{B}(y, t) = ky^3t^2\,\hat{\mathbf{z}}$ (where k is a constant). Find the emf induced in the loop.

Problem 7.14 As a lecture demonstration a short cylindrical bar magnet is dropped down a vertical aluminum pipe of slightly larger diameter, about 2 meters long. It takes several seconds to emerge at the bottom, whereas an otherwise identical piece of *unmagnetized* iron makes the trip in a fraction of a second. Explain why the magnet falls more slowly.[12]

[11] For further discussion of the jumping ring (and the related "floating ring"), see C. S. Schneider and J. P. Ertel, *Am. J. Phys.* **66**, 686 (1998); P. J. H. Tjossem and E. C. Brost, *Am. J. Phys.* **79**, 353 (2011).

[12] For a discussion of this amazing demonstration see K. D. Hahn et al., *Am. J. Phys.* **66**, 1066 (1998) and G. Donoso, C. L. Ladera, and P. Martin, *Am. J. Phys.* **79**, 193 (2011).

7.2.2 ■ The Induced Electric Field

Faraday's law generalizes the electrostatic rule $\boldsymbol{\nabla} \times \mathbf{E} = \mathbf{0}$ to the time-dependent régime. The *divergence* of $\mathbf{E}$ is still given by Gauss's law ($\boldsymbol{\nabla} \cdot \mathbf{E} = \frac{1}{\epsilon_0}\rho$). If $\mathbf{E}$ is a *pure* Faraday field (due exclusively to a changing $\mathbf{B}$, with $\rho = 0$), then

$$\boldsymbol{\nabla} \cdot \mathbf{E} = 0, \quad \boldsymbol{\nabla} \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}.$$

This is mathematically identical to magnetostatics,

$$\boldsymbol{\nabla} \cdot \mathbf{B} = 0, \quad \boldsymbol{\nabla} \times \mathbf{B} = \mu_0 \mathbf{J}.$$

Conclusion: Faraday-induced electric fields are determined by $-(\partial \mathbf{B}/\partial t)$ in exactly the same way as magnetostatic fields are determined by $\mu_0 \mathbf{J}$. The analog to Biot-Savart is[13] is

$$\mathbf{E} = -\frac{1}{4\pi}\int \frac{(\partial \mathbf{B}/\partial t) \times \hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\,d\tau = -\frac{1}{4\pi}\frac{\partial}{\partial t}\int \frac{\mathbf{B} \times \hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\,d\tau, \tag{7.18}$$

and if symmetry permits, we can use all the tricks associated with Ampère's law in integral form ($\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 I_{\text{enc}}$), only now it's *Faraday's* law in integral form:

$$\oint \mathbf{E} \cdot d\mathbf{l} = -\frac{d\Phi}{dt}. \tag{7.19}$$

The rate of change of (magnetic) flux through the Amperian loop plays the role formerly assigned to $\mu_0 I_{\text{enc}}$.

Example 7.7. A uniform magnetic field $\mathbf{B}(t)$, pointing straight up, fills the shaded circular region of Fig. 7.25. If $\mathbf{B}$ is changing with time, what is the induced electric field?

Solution

$\mathbf{E}$ points in the circumferential direction, just like the *magnetic* field inside a long straight wire carrying a uniform *current* density. Draw an Amperian loop of radius s, and apply Faraday's law:

$$\oint \mathbf{E} \cdot d\mathbf{l} = E(2\pi s) = -\frac{d\Phi}{dt} = -\frac{d}{dt}\left(\pi s^2 B(t)\right) = -\pi s^2 \frac{dB}{dt}.$$

Therefore

$$\mathbf{E} = -\frac{s}{2}\frac{dB}{dt}\,\hat{\boldsymbol{\phi}}.$$

If $\mathbf{B}$ is *increasing*, $\mathbf{E}$ runs *clockwise*, as viewed from above.

[13] Magnetostatics holds only for time-independent currents, but there is no such restriction on $\partial \mathbf{B}/\partial t$.

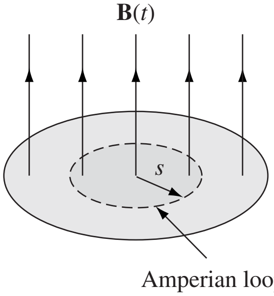


FIGURE 7.25

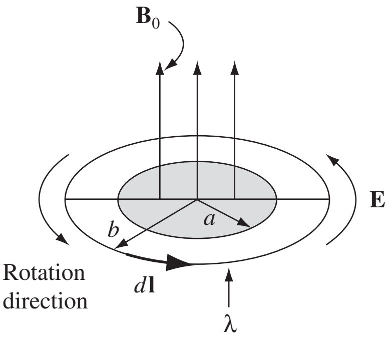


FIGURE 7.26

Example 7.8. A line charge λ is glued onto the rim of a wheel of radius b, which is then suspended horizontally, as shown in Fig. 7.26, so that it is free to rotate (the spokes are made of some nonconducting material—wood, maybe). In the central region, out to radius a, there is a uniform magnetic field $\mathbf{B}_0$, pointing up. Now someone turns the field off. What happens?

Solution

The changing magnetic field will induce an electric field, curling around the axis of the wheel. This electric field exerts a force on the charges at the rim, and the wheel starts to turn. According to Lenz's law, it will rotate in such a direction that *its* field tends to restore the upward flux. The motion, then, is counterclockwise, as viewed from above.

Faraday's law, applied to the loop at radius b, says

$$\oint \mathbf{E}\cdot d\mathbf{l} = E(2\pi b) = -\frac{d\Phi}{dt} = -\pi a^2\frac{dB}{dt}, \quad \text{or} \quad \mathbf{E} = -\frac{a^2}{2b}\frac{dB}{dt}\,\hat{\boldsymbol{\phi}}.$$

The torque on a segment of length $d\mathbf{l}$ is $(\mathbf{r}\times\mathbf{F})$, or $b\lambda E\,dl$. The total torque on the wheel is therefore

$$N = b\lambda\left(-\frac{a^2}{2b}\frac{dB}{dt}\right)\oint dl = -b\lambda\pi a^2\frac{dB}{dt},$$

and the angular momentum imparted to the wheel is

$$\int N dt = -\lambda\pi a^2 b\int_{B_0}^{0} dB = \lambda\pi a^2 b B_0.$$

It doesn't matter how quickly or slowly you turn off the field; the resulting angular velocity of the wheel is the same regardless. (If you find yourself wondering where the angular momentum *came* from, you're getting ahead of the story! Wait for the next chapter.)

Note that it's the *electric* field that did the rotating. To convince you of this, I deliberately set things up so that the *magnetic* field is *zero* at the location of the charge. The experimenter may tell you she never put in any electric field—all she did was switch off the magnetic field. But when she did that, an electric field automatically appeared, and it's this electric field that turned the wheel.

I must warn you, now, of a small fraud that tarnishes many applications of Faraday's law: Electromagnetic induction, of course, occurs only when the magnetic fields are *changing*, and yet we would like to use the apparatus of magneto*statics* (Ampère's law, the Biot-Savart law, and the rest) to *calculate* those magnetic fields. Technically, any result derived in this way is only approximately correct. But in practice the error is usually negligible, unless the field fluctuates extremely rapidly, or you are interested in points very far from the source. Even the case of a wire snipped by a pair of scissors (Prob. 7.18) is *static enough* for Ampère's law to apply. This régime, in which magnetostatic rules can be used to calculate the magnetic field on the right hand side of Faraday's law, is called **quasistatic**. Generally speaking, it is only when we come to electromagnetic waves and radiation that we must worry seriously about the breakdown of magnetostatics itself.

Example 7.9. An infinitely long straight wire carries a slowly varying current $I(t)$. Determine the induced electric field, as a function of the distance s from the wire.[14]

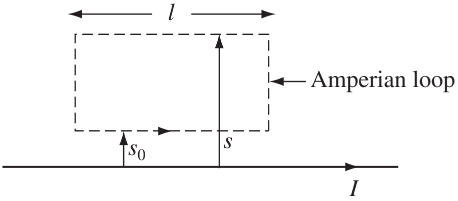


FIGURE 7.27

Solution

In the quasistatic approximation, the magnetic field is $(\mu_0 I/2\pi s)$, and it circles around the wire. Like the **B**-field of a solenoid, **E** here runs parallel to the axis. For the rectangular "Amperian loop" in Fig. 7.27, Faraday's law gives:

$$\begin{aligned}\oint \mathbf{E}\cdot d\mathbf{l} &= E(s_0)l - E(s)l = -\frac{d}{dt}\int \mathbf{B}\cdot d\mathbf{a} \\ &= -\frac{\mu_0 l}{2\pi}\frac{dI}{dt}\int_{s_0}^{s}\frac{1}{s'}\,ds' = -\frac{\mu_0 l}{2\pi}\frac{dI}{dt}(\ln s - \ln s_0).\end{aligned}$$

[14] This example is artificial, and not just in the obvious sense of involving infinite wires, but in a more subtle respect. It assumes that the current is the same (at any given instant) all the way down the line. This is a safe assumption for the *short* wires in typical electric circuits, but not for *long* wires (**transmission lines**), unless you supply a distributed and synchronized driving mechanism. But never mind—the problem doesn't inquire how you would *produce* such a current; it only asks what *fields* would result if you *did*. Variations on this problem are discussed by M. A. Heald, *Am. J. Phys.* **54**, 1142 (1986).

Thus

$$\mathbf{E}(s) = \left[\frac{\mu_0}{2\pi}\frac{dI}{dt}\ln s + K\right]\hat{\mathbf{z}}, \tag{7.20}$$

where K is a constant (that is to say, it is independent of s—it might still be a function of t). The actual *value* of K depends on the whole history of the function $I(t)$—we'll see some examples in Chapter 10.

Equation 7.20 has the peculiar implication that E blows up as s goes to infinity. *That* can't be true … What's gone wrong? *Answer:* We have overstepped the limits of the quasistatic approximation. As we shall see in Chapter 9, electromagnetic "news" travels at the speed of light, and at large distances $\mathbf{B}$ depends not on the current *now*, but on the current *as it was* at some earlier time (indeed, a whole *range* of earlier times, since different points on the wire are different distances away). If τ is the time it takes I to change substantially, then the quasistatic approximation should hold only for

$$s \ll c\tau, \tag{7.21}$$

and hence Eq. 7.20 simply does not apply, at extremely large s.

Problem 7.15 A long solenoid with radius a and n turns per unit length carries a time-dependent current $I(t)$ in the $\hat{\boldsymbol{\phi}}$ direction. Find the electric field (magnitude and direction) at a distance s from the axis (both inside and outside the solenoid), in the quasistatic approximation.

Problem 7.16 An alternating current $I = I_0\cos(\omega t)$ flows down a long straight wire, and returns along a coaxial conducting tube of radius a.

(a) In what *direction* does the induced electric field point (radial, circumferential, or longitudinal)?

(b) Assuming that the field goes to zero as $s \to \infty$, find $\mathbf{E}(s, t)$.[15]

Problem 7.17 A long solenoid of radius a, carrying n turns per unit length, is looped by a wire with resistance R, as shown in Fig. 7.28.

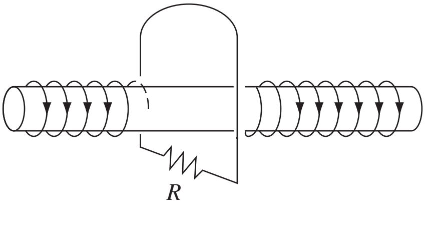


FIGURE 7.28

[15] This is not at all the way electric fields *actually* behave in coaxial cables, for reasons suggested in the previous footnote. See Sect. 9.5.3, or J. G. Cherveniak, *Am. J. Phys.*, **54**, 946 (1986), for a more realistic treatment.

(a) If the current in the solenoid is increasing at a constant rate ($dI/dt = k$), what current flows in the loop, and which way (left or right) does it pass through the resistor?

(b) If the current I in the solenoid is constant but the solenoid is pulled out of the loop (toward the left, to a place far from the loop), what total charge passes through the resistor?

Problem 7.18 A square loop, side a, resistance R, lies a distance s from an infinite straight wire that carries current I (Fig. 7.29). Now someone cuts the wire, so I drops to zero. In what direction does the induced current in the square loop flow, and what total charge passes a given point in the loop during the time this current flows? If you don't like the scissors model, turn the current down *gradually*:

$$I(t) = \begin{cases} (1-\alpha t)I, & \text{for } 0 \le t \le 1/\alpha, \\ 0, & \text{for } t > 1/\alpha. \end{cases}$$

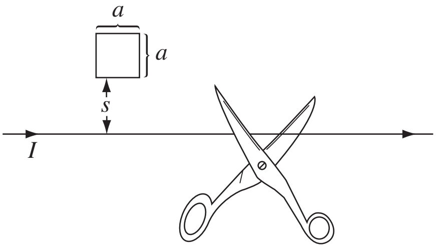


FIGURE 7.29

Problem 7.19 A toroidal coil has a rectangular cross section, with inner radius a, outer radius $a + w$, and height h. It carries a total of N tightly wound turns, and the current is increasing at a constant rate ($dI/dt = k$). If w and h are both much less than a, find the electric field at a point z above the center of the toroid. [*Hint:* Exploit the analogy between Faraday fields and magnetostatic fields, and refer to Ex. 5.6.]

Problem 7.20 Where is $\partial \mathbf{B}/\partial t$ nonzero, in Figure 7.21(b)? Exploit the analogy between Faraday's law and Ampère's law to sketch (qualitatively) the electric field.

Problem 7.21 Imagine a uniform magnetic field, pointing in the z direction and filling all space ($\mathbf{B} = B_0\,\hat{\mathbf{z}}$). A positive charge is at rest, at the origin. Now somebody turns off the magnetic field, thereby inducing an electric field. In what direction does the charge move?[16]

7.2.3 ■ Inductance

Suppose you have two loops of wire, at rest (Fig. 7.30). If you run a steady current I_1 around loop 1, it produces a magnetic field $\mathbf{B}_1$. Some of the field lines pass

[16] This paradox was suggested by Tom Colbert. Refer to Problem 2.55.

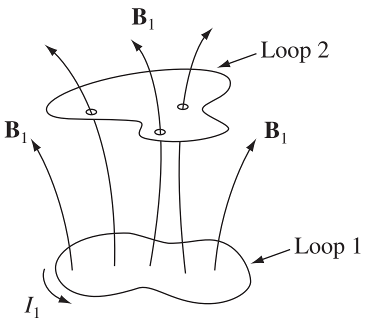


FIGURE 7.30

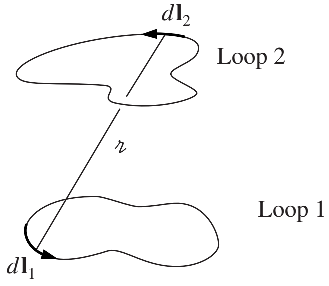


FIGURE 7.31

through loop 2; let Φ_2 be the flux of $\mathbf{B}_1$ through 2. You might have a tough time actually *calculating* $\mathbf{B}_1$, but a glance at the Biot-Savart law,

$$\mathbf{B}_1 = \frac{\mu_0}{4\pi} I_1 \oint \frac{d\mathbf{l}_1 \times \hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2},$$

reveals one significant fact about this field: *It is proportional to the current* I_1. Therefore, so too is the flux through loop 2:

$$\Phi_2 = \int \mathbf{B}_1 \cdot d\mathbf{a}_2.$$

Thus

$$\Phi_2 = M_{21} I_1, \tag{7.22}$$

where M_{21} is the constant of proportionality; it is known as the **mutual inductance** of the two loops.

There is a cute formula for the mutual inductance, which you can derive by expressing the flux in terms of the vector potential, and invoking Stokes' theorem:

$$\Phi_2 = \int \mathbf{B}_1 \cdot d\mathbf{a}_2 = \int (\boldsymbol{\nabla} \times \mathbf{A}_1) \cdot d\mathbf{a}_2 = \oint \mathbf{A}_1 \cdot d\mathbf{l}_2.$$

Now, according to Eq. 5.66,

$$\mathbf{A}_1 = \frac{\mu_0 I_1}{4\pi} \oint \frac{d\mathbf{l}_1}{\mathscr{r}},$$

and hence

$$\Phi_2 = \frac{\mu_0 I_1}{4\pi} \oint \left(\oint \frac{d\mathbf{l}_1}{\mathscr{r}} \right) \cdot d\mathbf{l}_2.$$

Evidently

$$M_{21} = \frac{\mu_0}{4\pi} \oint \oint \frac{d\mathbf{l}_1 \cdot d\mathbf{l}_2}{\mathscr{r}}. \tag{7.23}$$

This is the **Neumann formula**; it involves a double line integral—one integration around loop 1, the other around loop 2 (Fig. 7.31). It's not very useful for practical calculations, but it does reveal two important things about mutual inductance:

1. M_{21} is a purely geometrical quantity, having to do with the sizes, shapes, and relative positions of the two loops.
2. The integral in Eq. 7.23 is unchanged if we switch the roles of loops 1 and 2; it follows that

$$M_{21} = M_{12}. \tag{7.24}$$

This is an astonishing conclusion: *Whatever the shapes and positions of the loops, the flux through 2 when we run a current I around 1 is identical to the flux through 1 when we send the same current I around 2*. We may as well drop the subscripts and call them both M.

Example 7.10. A short solenoid (length l and radius a, with n_1 turns per unit length) lies on the axis of a very long solenoid (radius b, n_2 turns per unit length) as shown in Fig. 7.32. Current I flows in the short solenoid. What is the flux through the long solenoid?

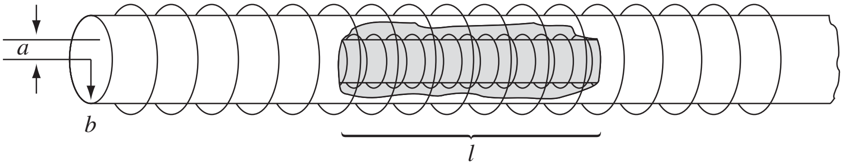


FIGURE 7.32

Solution

Since the inner solenoid is short, it has a very complicated field; moreover, it puts a different flux through each turn of the outer solenoid. It would be a *miserable* task to compute the total flux this way. However, if we exploit the equality of the mutual inductances, the problem becomes very easy. Just look at the reverse situation: run the current I through the *outer* solenoid, and calculate the flux through the *inner* one. The field inside the long solenoid is constant:

$$B = \mu_0 n_2 I$$

(Eq. 5.59), so the flux through a single loop of the short solenoid is

$$B\pi a^2 = \mu_0 n_2 I \pi a^2.$$

There are $n_1 l$ turns in all, so the total flux through the inner solenoid is

$$\Phi = \mu_0 \pi a^2 n_1 n_2 l I.$$

This is also the flux a current I in the *short* solenoid would put through the *long* one, which is what we set out to find. Incidentally, the mutual inductance, in this case, is

$$M = \mu_0 \pi a^2 n_1 n_2 l.$$

Suppose, now, that you *vary* the current in loop 1. The flux through loop 2 will vary accordingly, and Faraday's law says this changing flux will induce an emf in loop 2:

$$\mathcal{E}_2 = -\frac{d\Phi_2}{dt} = -M\frac{dI_1}{dt}. \tag{7.25}$$

(In quoting Eq. 7.22—which was based on the Biot-Savart law—I am tacitly assuming that the currents change slowly enough for the system to be considered quasistatic.) What a remarkable thing: Every time you change the current in loop 1, an induced current flows in loop 2—even though there are no wires connecting them!

Come to think of it, a changing current not only induces an emf in any nearby loops, it also induces an emf in the source loop *itself* (Fig 7.33). Once again, the field (and therefore also the flux) is proportional to the current:

$$\Phi = LI. \tag{7.26}$$

The constant of proportionality L is called the **self inductance** (or simply the **inductance**) of the loop. As with M, it depends on the geometry (size and shape) of the loop. If the current changes, the emf induced in the loop is

$$\mathcal{E} = -L\frac{dI}{dt}. \tag{7.27}$$

Inductance is measured in **henries** (H); a henry is a volt-second per ampere.

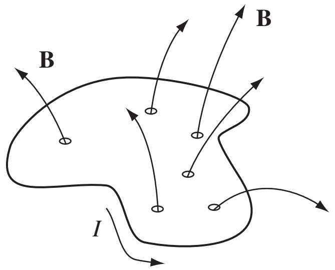


FIGURE 7.33

Example 7.11. Find the self-inductance of a toroidal coil with rectangular cross section (inner radius a, outer radius b, height h), that carries a total of N turns.

Solution

The magnetic field inside the toroid is (Eq. 5.60)

$$B = \frac{\mu_0 N I}{2\pi s}.$$

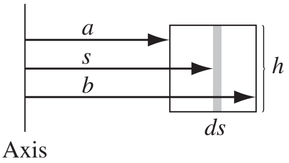


FIGURE 7.34

The flux through a single turn (Fig. 7.34) is

$$\int \mathbf{B} \cdot d\mathbf{a} = \frac{\mu_0 N I}{2\pi} h \int_a^b \frac{1}{s}\, ds = \frac{\mu_0 N I h}{2\pi} \ln\left(\frac{b}{a}\right).$$

The *total* flux is N times this, so the self-inductance (Eq. 7.26) is

$$L = \frac{\mu_0 N^2 h}{2\pi} \ln\left(\frac{b}{a}\right). \tag{7.28}$$

Inductance (like capacitance) is an intrinsically *positive* quantity. Lenz's law, which is enforced by the minus sign in Eq. 7.27, dictates that the emf is in such a direction as to *oppose* any *change in current*. For this reason, it is called a **back emf**. Whenever you try to alter the current in a wire, you must fight against this back emf. Inductance plays somewhat the same role in electric circuits that *mass* plays in mechanical systems: The greater L is, the harder it is to change the current, just as the larger the mass, the harder it is to change an object's velocity.

Example 7.12. Suppose a current I is flowing around a loop, when someone suddenly cuts the wire. The current drops "instantaneously" to zero. This generates a whopping back emf, for although I may be small, dI/dt is enormous. (That's why you sometimes draw a spark when you unplug an iron or toaster—electromagnetic induction is desperately trying to keep the current going, even if it has to jump the gap in the circuit.)

Nothing so dramatic occurs when you plug *in* a toaster or iron. In this case induction opposes the sudden *increase* in current, prescribing instead a smooth and continuous buildup. Suppose, for instance, that a battery (which supplies a constant emf $\mathcal{E}_0$) is connected to a circuit of resistance R and inductance L (Fig. 7.35). What current flows?

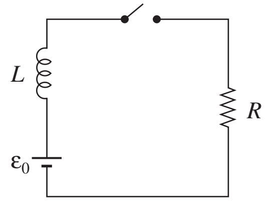


FIGURE 7.35

Solution

The total emf in this circuit is $\mathcal{E}_0$ from the battery plus $-L(dI/dt)$ from the inductance. Ohm's law, then, says[17]

$$\mathcal{E}_0 - L\frac{dI}{dt} = IR.$$

This is a first-order differential equation for I as a function of time. The general solution, as you can show for yourself, is

$$I(t) = \frac{\mathcal{E}_0}{R} + ke^{-(R/L)t},$$

where k is a constant to be determined by the initial conditions. In particular, if you close the switch at time $t = 0$, so $I(0) = 0$, then $k = -\mathcal{E}_0/R$, and

$$I(t) = \frac{\mathcal{E}_0}{R}\left[1 - e^{-(R/L)t}\right]. \tag{7.29}$$

This function is plotted in Fig. 7.36. Had there been no inductance in the circuit, the current would have jumped immediately to $\mathcal{E}_0/R$. In practice, *every* circuit has *some* self-inductance, and the current approaches $\mathcal{E}_0/R$ asymptotically. The quantity $\tau \equiv L/R$ is the **time constant**; it tells you how long the current takes to reach a substantial fraction (roughly two-thirds) of its final value.

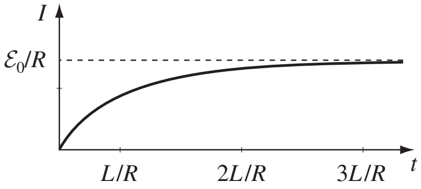


FIGURE 7.36

[17] Notice that $-L(dI/dt)$ goes on the *left* side of the equation—it is part of the emf that establishes the voltage across the resistor.

Problem 7.22 A small loop of wire (radius a) is held a distance z above the center of a large loop (radius b), as shown in Fig. 7.37. The planes of the two loops are parallel, and perpendicular to the common axis.

(a) Suppose current I flows in the big loop. Find the flux through the little loop. (The little loop is so small that you may consider the field of the big loop to be essentially constant.)

(b) Suppose current I flows in the little loop. Find the flux through the big loop. (The little loop is so small that you may treat it as a magnetic dipole.)

(c) Find the mutual inductances, and confirm that $M_{12} = M_{21}$.

Problem 7.23 A square loop of wire, of side a, lies midway between two long wires, $3a$ apart, and in the same plane. (Actually, the long wires are sides of a large rectangular loop, but the short ends are so far away that they can be neglected.) A clockwise current I in the square loop is gradually increasing: $dI/dt = k$ (a constant). Find the emf induced in the big loop. Which way will the induced current flow?

Problem 7.24 Find the self-inductance per unit length of a long solenoid, of radius R, carrying n turns per unit length.

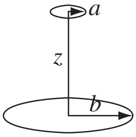


FIGURE 7.37

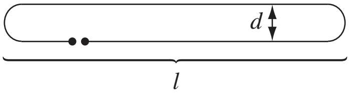


FIGURE 7.38

Problem 7.25 Try to compute the self-inductance of the "hairpin" loop shown in Fig. 7.38. (Neglect the contribution from the ends; most of the flux comes from the long straight section.) You'll run into a snag that is characteristic of many self-inductance calculations. To get a definite answer, assume the wire has a tiny radius ϵ, and ignore any flux through the wire itself.

Problem 7.26 An alternating current $I(t) = I_0 \cos(\omega t)$ (amplitude 0.5 A, frequency 60 Hz) flows down a straight wire, which runs along the axis of a toroidal coil with rectangular cross section (inner radius 1 cm, outer radius 2 cm, height 1 cm, 1000 turns). The coil is connected to a 500 Ω resistor.

(a) In the quasistatic approximation, what emf is induced in the toroid? Find the current, $I_R(t)$, in the resistor.

(b) Calculate the back emf in the coil, due to the current $I_R(t)$. What is the ratio of the amplitudes of this back emf and the "direct" emf in (a)?

Problem 7.27 A capacitor C is charged up to a voltage V and connected to an inductor L, as shown schematically in Fig. 7.39. At time $t = 0$, the switch S is closed. Find the current in the circuit as a function of time. How does your answer change if a resistor R is included in series with C and L?

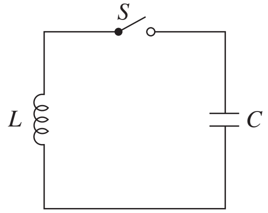


FIGURE 7.39

7.2.4 ■ Energy in Magnetic Fields

It takes a certain amount of energy to start a current flowing in a circuit. I'm not talking about the energy delivered to the resistors and converted into heat—that is irretrievably lost, as far as the circuit is concerned, and can be large or small, depending on how long you let the current run. What I am concerned with, rather, is the work you must do *against the back emf* to get the current going. This is a *fixed* amount, and it is *recoverable*: you get it back when the current is turned off. In the meantime, it represents energy latent in the circuit; as we'll see in a moment, it can be regarded as energy stored in the magnetic field.

The work done on a unit charge, against the back emf, in one trip around the circuit is $-\mathcal{E}$ (the minus sign records the fact that this is the work done *by you against* the emf, not the work done by the emf). The amount of charge per unit time passing down the wire is I. So the total work done per unit time is

$$\frac{dW}{dt} = -\mathcal{E}I = LI\frac{dI}{dt}.$$

If we start with zero current and build it up to a final value I, the work done (integrating the last equation over time) is

$$\boxed{W = \frac{1}{2}LI^2.} \tag{7.30}$$

It does not depend on how *long* we take to crank up the current, only on the geometry of the loop (in the form of L) and the final current I.

There is a nicer way to write W, which has the advantage that it is readily generalized to surface and volume currents. Remember that the flux Φ through the loop is equal to LI (Eq. 7.26). On the other hand,

$$\Phi = \int \mathbf{B} \cdot d\mathbf{a} = \int (\nabla \times \mathbf{A}) \cdot d\mathbf{a} = \oint \mathbf{A} \cdot d\mathbf{l},$$

where the line integral is around the perimeter of the loop. Thus

$$LI = \oint \mathbf{A} \cdot d\mathbf{l},$$

and therefore

$$W = \frac{1}{2} I \oint \mathbf{A} \cdot d\mathbf{l} = \frac{1}{2} \oint (\mathbf{A} \cdot \mathbf{I})\, dl. \tag{7.31}$$

In this form, the generalization to volume currents is obvious:

$$W = \frac{1}{2} \int_{\mathcal{V}} (\mathbf{A} \cdot \mathbf{J})\, d\tau. \tag{7.32}$$

But we can do even better, and express W entirely in terms of the magnetic field: Ampère's law, $\boldsymbol{\nabla} \times \mathbf{B} = \mu_0 \mathbf{J}$, lets us eliminate $\mathbf{J}$:

$$W = \frac{1}{2\mu_0} \int \mathbf{A} \cdot (\boldsymbol{\nabla} \times \mathbf{B})\, d\tau. \tag{7.33}$$

Integration by parts transfers the derivative from $\mathbf{B}$ to $\mathbf{A}$; specifically, product rule 6 states that

$$\boldsymbol{\nabla} \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\boldsymbol{\nabla} \times \mathbf{A}) - \mathbf{A} \cdot (\boldsymbol{\nabla} \times \mathbf{B}),$$

so

$$\mathbf{A} \cdot (\boldsymbol{\nabla} \times \mathbf{B}) = \mathbf{B} \cdot \mathbf{B} - \boldsymbol{\nabla} \cdot (\mathbf{A} \times \mathbf{B}).$$

Consequently,

$$\begin{aligned} W &= \frac{1}{2\mu_0} \left[\int B^2\, d\tau - \int \boldsymbol{\nabla} \cdot (\mathbf{A} \times \mathbf{B})\, d\tau \right] \\ &= \frac{1}{2\mu_0} \left[\int_{\mathcal{V}} B^2\, d\tau - \oint_{\mathcal{S}} (\mathbf{A} \times \mathbf{B}) \cdot d\mathbf{a} \right], \end{aligned} \tag{7.34}$$

where $\mathcal{S}$ is the surface bounding the volume $\mathcal{V}$.

Now, the integration in Eq. 7.32 is to be taken over the *entire volume occupied by the current*. But any region *larger* than this will do just as well, for $\mathbf{J}$ is zero out there anyway. In Eq. 7.34, the larger the region we pick the greater is the contribution from the volume integral, and therefore the smaller is that of the surface integral (this makes sense: as the surface gets farther from the current, both $\mathbf{A}$ and $\mathbf{B}$ decrease). In particular, if we agree to integrate over *all* space, then the surface integral goes to zero, and we are left with

$$\boxed{W = \frac{1}{2\mu_0} \int_{\text{all space}} B^2\, d\tau.} \tag{7.35}$$

In view of this result, we say the energy is "stored in the magnetic field," in the amount $(B^2/2\mu_0)$ per unit volume. This is a nice way to think of it, though someone looking at Eq. 7.32 might prefer to say that the energy is stored in the *current distribution*, in the amount $\frac{1}{2}(\mathbf{A} \cdot \mathbf{J})$ per unit volume. The distinction is one of bookkeeping; the important quantity is the total energy W, and we need not worry about where (if anywhere) the energy is "located."

You might find it strange that it takes energy to set up a magnetic field—after all, magnetic fields *themselves* do no work. The point is that producing a magnetic field, where previously there was none, requires *changing* the field, and a changing **B**-field, according to Faraday, induces an *electric* field. The latter, of course, *can* do work. In the beginning, there is no **E**, and at the end there is no **E**; but in between, while **B** is building up, there *is* an **E**, and it is against *this* that the work is done. (You see why I could not calculate the energy stored in a magnetostatic field back in Chapter 5.) In the light of this, it is extraordinary how similar the magnetic energy formulas are to their electrostatic counterparts:[18]

$$W_{\text{elec}} = \frac{1}{2}\int (V\rho)\, d\tau = \frac{\epsilon_0}{2}\int E^2\, d\tau, \qquad \text{(2.43 and 2.45)}$$

$$W_{\text{mag}} = \frac{1}{2}\int (\mathbf{A}\cdot\mathbf{J})\, d\tau = \frac{1}{2\mu_0}\int B^2\, d\tau. \qquad \text{(7.32 and 7.35)}$$

Example 7.13. A long coaxial cable carries current I (the current flows down the surface of the inner cylinder, radius a, and back along the outer cylinder, radius b) as shown in Fig. 7.40. Find the magnetic energy stored in a section of length l.

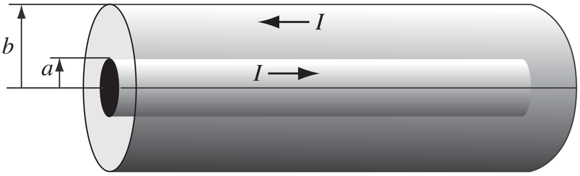


FIGURE 7.40

Solution

According to Ampère's law, the field between the cylinders is

$$\mathbf{B} = \frac{\mu_0 I}{2\pi s}\hat{\boldsymbol{\phi}}.$$

Elsewhere, the field is zero. Thus, the energy per unit volume is

$$\frac{1}{2\mu_0}\left(\frac{\mu_0 I}{2\pi s}\right)^2 = \frac{\mu_0 I^2}{8\pi^2 s^2}.$$

The energy in a cylindrical shell of length l, radius s, and thickness ds, then, is

$$\left(\frac{\mu_0 I^2}{8\pi^2 s^2}\right) 2\pi l s\, ds = \frac{\mu_0 I^2 l}{4\pi}\left(\frac{ds}{s}\right).$$

[18] For an illuminating confirmation of Eq. 7.35, using the method of Prob. 2.44, see T. H. Boyer, *Am. J. Phys.* **69**, 1 (2001).

Integrating from a to b, we have:

$$W = \frac{\mu_0 I^2 l}{4\pi} \ln\left(\frac{b}{a}\right).$$

By the way, this suggests a very simple way to calculate the self-inductance of the cable. According to Eq. 7.30, the energy can also be written as $\frac{1}{2}LI^2$. Comparing the two expressions,[19]

$$L = \frac{\mu_0 l}{2\pi} \ln\left(\frac{b}{a}\right).$$

This method of calculating self-inductance is especially useful when the current is not confined to a single path, but spreads over some surface or volume, so that different parts of the current enclose different amounts of flux. In such cases, it can be very tricky to get the inductance directly from Eq. 7.26, and it is best to let Eq. 7.30 *define* L.

Problem 7.28 Find the energy stored in a section of length l of a long solenoid (radius R, current I, n turns per unit length), (a) using Eq. 7.30 (you found L in Prob. 7.24); (b) using Eq. 7.31 (we worked out $\mathbf{A}$ in Ex. 5.12); (c) using Eq. 7.35; (d) using Eq. 7.34 (take as your volume the cylindrical tube from radius $a < R$ out to radius $b > R$).

Problem 7.29 Calculate the energy stored in the toroidal coil of Ex. 7.11, by applying Eq. 7.35. Use the answer to check Eq. 7.28.

Problem 7.30 A long cable carries current in one direction uniformly distributed over its (circular) cross section. The current returns along the surface (there is a very thin insulating sheath separating the currents). Find the self-inductance per unit length.

Problem 7.31 Suppose the circuit in Fig. 7.41 has been connected for a long time when suddenly, at time $t = 0$, switch S is thrown from A to B, bypassing the battery.

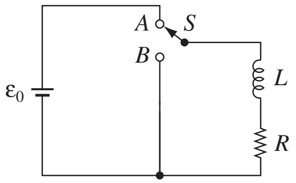


FIGURE 7.41

[19] Notice the similarity to Eq. 7.28—in a sense, the rectangular toroid *is* a short coaxial cable, turned on its side.

(a) What is the current at any subsequent time t?

(b) What is the total energy delivered to the resistor?

(c) Show that this is equal to the energy originally stored in the inductor.

Problem 7.32 Two tiny wire loops, with areas $\mathbf{a}_1$ and $\mathbf{a}_2$, are situated a displacement $\boldsymbol{\mathscr{r}}$ apart (Fig. 7.42).

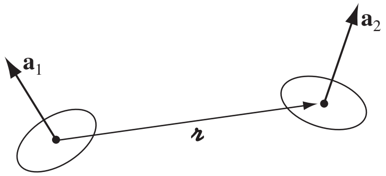


FIGURE 7.42

(a) Find their mutual inductance. [*Hint:* Treat them as magnetic dipoles, and use Eq. 5.88.] Is your formula consistent with Eq. 7.24?

(b) Suppose a current I_1 is flowing in loop 1, and we propose to turn on a current I_2 in loop 2. How much work must be done, against the mutually induced emf, to keep the current I_1 flowing in loop 1? In light of this result, comment on Eq. 6.35.

Problem 7.33 An infinite cylinder of radius R carries a uniform surface charge σ. We propose to set it spinning about its axis, at a final angular velocity ω_f. How much work will this take, per unit length? Do it two ways, and compare your answers:

(a) Find the magnetic field and the induced electric field (in the quasistatic approximation), inside and outside the cylinder, in terms of ω, $\dot{\omega}$, and s (the distance from the axis). Calculate the torque you must exert, and from that obtain the work done per unit length ($W = \int N\,d\phi$).

(b) Use Eq. 7.35 to determine the energy stored in the resulting magnetic field.

7.3 ■ MAXWELL'S EQUATIONS

7.3.1 ■ Electrodynamics Before Maxwell

So far, we have encountered the following laws, specifying the divergence and curl of electric and magnetic fields:

$$
\begin{array}{llll}
\text{(i)} & \nabla \cdot \mathbf{E} = \dfrac{1}{\epsilon_0}\rho & & \text{(Gauss's law),} \\[2ex]
\text{(ii)} & \nabla \cdot \mathbf{B} = 0 & & \text{(no name),} \\[2ex]
\text{(iii)} & \nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t} & & \text{(Faraday's law),} \\[2ex]
\text{(iv)} & \nabla \times \mathbf{B} = \mu_0 \mathbf{J} & & \text{(Ampère's law).}
\end{array}
$$

These equations represent the state of electromagnetic theory in the mid-nineteenth century, when Maxwell began his work. They were not written in so compact a form, in those days, but their physical content was familiar. Now, it happens that there is a fatal inconsistency in these formulas. It has to do with the old rule that divergence of curl is always zero. If you apply the divergence to number (iii), everything works out:

$$\nabla \cdot (\nabla \times \mathbf{E}) = \nabla \cdot \left(-\frac{\partial \mathbf{B}}{\partial t}\right) = -\frac{\partial}{\partial t}(\nabla \cdot \mathbf{B}).$$

The left side is zero because divergence of curl is zero; the right side is zero by virtue of equation (ii). But when you do the same thing to number (iv), you get into trouble:

$$\nabla \cdot (\nabla \times \mathbf{B}) = \mu_0(\nabla \cdot \mathbf{J}); \tag{7.36}$$

the left side must be zero, but the right side, in general, is *not*. For *steady* currents, the divergence of **J** is zero, but when we go beyond magnetostatics Ampère's law cannot be right.

There's another way to see that Ampère's law is bound to fail for nonsteady currents. Suppose we're in the process of charging up a capacitor (Fig. 7.43). In integral form, Ampère's law reads

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 I_{\text{enc}}.$$

I want to apply it to the Amperian loop shown in the diagram. How do I determine I_{enc}? Well, it's the total current passing through the loop, or, more precisely, the current piercing a surface that has the loop for its boundary. In this case, the *simplest* surface lies in the plane of the loop—the wire punctures this surface, so $I_{\text{enc}} = I$. Fine—but what if I draw instead the balloon-shaped surface in Fig. 7.43? *No* current passes through *this* surface, and I conclude that $I_{\text{enc}} = 0$! We never had this problem in magnetostatics because the conflict arises only when charge

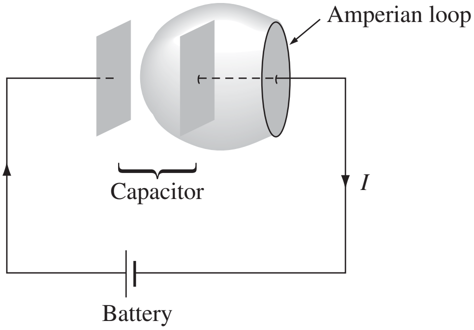


FIGURE 7.43

is piling up somewhere (in this case, on the capacitor plates). But for *nonsteady* currents (such as this one) "the current enclosed by the loop" is an ill-defined notion; it depends entirely on what surface you use. (If this seems pedantic to you—"obviously one should use the plane surface"—remember that the Amperian loop could be some contorted shape that doesn't even lie in a plane.)

Of course, we had no right to *expect* Ampère's law to hold outside of magnetostatics; after all, we derived it from the Biot-Savart law. However, in Maxwell's time there was no *experimental* reason to doubt that Ampère's law was of wider validity. The flaw was a purely theoretical one, and Maxwell fixed it by purely theoretical arguments.

7.3.2 ■ How Maxwell Fixed Ampère's Law

The problem is on the right side of Eq. 7.36, which *should be* zero, but *isn't*. Applying the continuity equation (5.29) and Gauss's law, the offending term can be rewritten:

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial t}(\epsilon_0 \nabla \cdot \mathbf{E}) = -\nabla \cdot \left(\epsilon_0 \frac{\partial \mathbf{E}}{\partial t}\right).$$

If we were to combine $\epsilon_0(\partial \mathbf{E}/\partial t)$ with $\mathbf{J}$, in Ampère's law, it would be just right to kill off the extra divergence:

$$\boxed{\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}.} \tag{7.37}$$

(Maxwell himself had other reasons for wanting to add this quantity to Ampère's law. To him, the rescue of the continuity equation was a happy dividend rather than a primary motive. But today we recognize this argument as a far more compelling one than Maxwell's, which was based on a now-discredited model of the ether.)[20]

Such a modification changes nothing, as far as magneto*statics* is concerned: when $\mathbf{E}$ is constant, we still have $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$. In fact, Maxwell's term is hard to detect in ordinary electromagnetic experiments, where it must compete for attention with $\mathbf{J}$—that's why Faraday and the others never discovered it in the laboratory. However, it plays a crucial role in the propagation of electromagnetic waves, as we'll see in Chapter 9.

Apart from curing the defect in Ampère's law, Maxwell's term has a certain aesthetic appeal: Just as a changing *magnetic* field induces an *electric* field (Faraday's law), so[21]

A changing electric field induces a magnetic field.

[20] For the history of this subject, see A. M. Bork, *Am. J. Phys.* **31**, 854 (1963).

[21] See footnote 8 (page 313) for commentary on the word "induce." The same issue arises here: Should a changing electric field be regarded as an independent source of magnetic field (along with current)? In a proximate sense it does function as a source, but since the electric field itself was produced by charges and currents, they alone are the "ultimate" sources of **E** and **B**. See S. E. Hill, *Phys. Teach.* **49**, 343 (2011); for a contrary view, see C. Savage, *Phys. Teach.* **50**, 226 (2012).

Of course, theoretical convenience and aesthetic consistency are only *suggestive*—there might, after all, be other ways to doctor up Ampère's law. The real confirmation of Maxwell's theory came in 1888 with Hertz's experiments on electromagnetic waves.

Maxwell called his extra term the **displacement current**:

$$\mathbf{J}_d \equiv \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}. \tag{7.38}$$

(It's a misleading name; $\epsilon_0(\partial \mathbf{E}/\partial t)$ has nothing to do with current, except that it adds to $\mathbf{J}$ in Ampère's law.) Let's see now how displacement current resolves the paradox of the charging capacitor (Fig. 7.43). If the capacitor plates are very close together (I didn't *draw* them that way, but the calculation is simpler if you assume this), then the electric field between them is

$$E = \frac{1}{\epsilon_0}\sigma = \frac{1}{\epsilon_0}\frac{Q}{A},$$

where Q is the charge on the plate and A is its area. Thus, between the plates

$$\frac{\partial E}{\partial t} = \frac{1}{\epsilon_0 A}\frac{dQ}{dt} = \frac{1}{\epsilon_0 A} I.$$

Now, Eq. 7.37 reads, in integral form,

$$\oint \mathbf{B}\cdot d\mathbf{l} = \mu_0 I_{\text{enc}} + \mu_0\epsilon_0 \int \left(\frac{\partial \mathbf{E}}{\partial t}\right)\cdot d\mathbf{a}. \tag{7.39}$$

If we choose the *flat* surface, then $E = 0$ and $I_{\text{enc}} = I$. If, on the other hand, we use the balloon-shaped surface, then $I_{\text{enc}} = 0$, but $\int(\partial \mathbf{E}/\partial t)\cdot d\mathbf{a} = I/\epsilon_0$. So we get the same answer for either surface, though in the first case it comes from the conduction current, and in the second from the displacement current.

Example 7.14. Imagine two concentric metal spherical shells (Fig. 7.44).

The inner one (radius a) carries a charge $Q(t)$, and the outer one (radius b) an opposite charge $-Q(t)$. The space between them is filled with Ohmic material of conductivity σ, so a radial current flows:

$$\mathbf{J} = \sigma\mathbf{E} = \sigma\frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}; \quad I = -\dot{Q} = \int \mathbf{J}\cdot d\mathbf{a} = \frac{\sigma Q}{\epsilon_0}.$$

This configuration is spherically symmetrical, so the magnetic field has to be zero (the only direction it could possibly point is radial, and $\nabla\cdot\mathbf{B} = 0 \Rightarrow \oint \mathbf{B}\cdot d\mathbf{a} = B(4\pi r^2) = 0$, so $\mathbf{B} = \mathbf{0}$). *What?* I thought currents produce magnetic fields! Isn't that what Biot-Savart and Ampère taught us? How can there be a $\mathbf{J}$ with no accompanying $\mathbf{B}$?

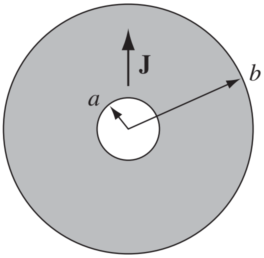


FIGURE 7.44

Solution

This is not a static configuration: Q, $\mathbf{E}$, and $\mathbf{J}$ are all functions of time; Ampère and Biot-Savart do not apply. The displacement current

$$J_d = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \frac{1}{4\pi}\frac{\dot{Q}}{r^2}\,\hat{\mathbf{r}} = -\sigma \frac{Q}{4\pi\epsilon_0 r^2}\,\hat{\mathbf{r}}$$

exactly cancels the conduction current (in Eq. 7.37), and the magnetic field (determined by $\nabla \cdot \mathbf{B} = 0$, $\nabla \times \mathbf{B} = \mathbf{0}$) is indeed zero.

Problem 7.34 A fat wire, radius a, carries a constant current I, uniformly distributed over its cross section. A narrow gap in the wire, of width $w \ll a$, forms a parallel-plate capacitor, as shown in Fig. 7.45. Find the magnetic field in the gap, at a distance $s < a$ from the axis.

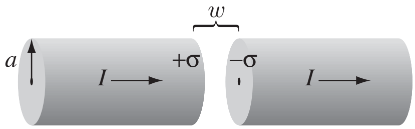


FIGURE 7.45

Problem 7.35 The preceding problem was an artificial model for the charging capacitor, designed to avoid complications associated with the current spreading out over the surface of the plates. For a more realistic model, imagine *thin* wires that connect to the centers of the plates (Fig. 7.46a). Again, the current I is constant, the radius of the capacitor is a, and the separation of the plates is $w \ll a$. Assume that the current flows out over the plates in such a way that the surface charge is uniform, at any given time, and is zero at $t = 0$.

(a) Find the electric field between the plates, as a function of t.

(b) Find the displacement current through a circle of radius s in the plane midway between the plates. Using this circle as your "Amperian loop," and the flat surface that spans it, find the magnetic field at a distance s from the axis.

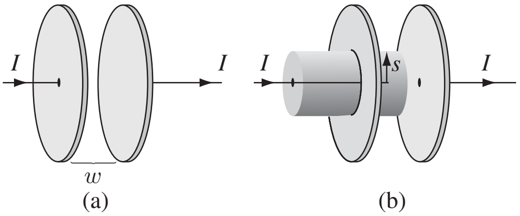


FIGURE 7.46

(c) Repeat part (b), but this time use the cylindrical surface in Fig. 7.46(b), which is open at the right end and extends to the left through the plate and terminates outside the capacitor. Notice that the displacement current through this surface is zero, and there are two contributions to I_{enc}.[22]

Problem 7.36 Refer to Prob. 7.16, to which the correct answer was

$$\mathbf{E}(s,t) = \frac{\mu_0 I_0 \omega}{2\pi} \sin(\omega t) \ln\left(\frac{a}{s}\right) \hat{\mathbf{z}}.$$

(a) Find the displacement current density $\mathbf{J}_d$.

(b) Integrate it to get the total displacement current,

$$I_d = \int \mathbf{J}_d \cdot d\mathbf{a}.$$

(c) Compare I_d and I. (What's their ratio?) If the outer cylinder were, say, 2 mm in diameter, how high would the frequency have to be, for I_d to be 1% of I? [This problem is designed to indicate why Faraday never discovered displacement currents, and why it is ordinarily safe to ignore them unless the frequency is extremely high.]

7.3.3 ■ Maxwell's Equations

In the last section we put the finishing touches on Maxwell's equations:

$$\boxed{\begin{array}{llll} \text{(i)} & \nabla\cdot\mathbf{E} = \dfrac{1}{\epsilon_0}\rho & & \text{(Gauss's law),} \\ \text{(ii)} & \nabla\cdot\mathbf{B} = 0 & & \text{(no name),} \\ \text{(iii)} & \nabla\times\mathbf{E} = -\dfrac{\partial\mathbf{B}}{\partial t} & & \text{(Faraday's law),} \\ \text{(iv)} & \nabla\times\mathbf{B} = \mu_0\mathbf{J} + \mu_0\epsilon_0\dfrac{\partial\mathbf{E}}{\partial t} & & \text{(Ampère's law with} \\ & & & \text{Maxwell's correction).} \end{array}} \tag{7.40}$$

[22] This problem raises an interesting quasi-philosophical question: If you measure **B** in the laboratory, have you detected the effects of displacement current (as (b) would suggest), or merely confirmed the effects of ordinary currents (as (c) implies)? See D. F. Bartlett, *Am. J. Phys.* **58**, 1168 (1990).

Together with the force law,

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{7.41}$$

they summarize the entire theoretical content of classical electrodynamics[23] (save for some special properties of matter, which we encountered in Chapters 4 and 6). Even the continuity equation,

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}, \tag{7.42}$$

which is the mathematical expression of conservation of charge, can be derived from Maxwell's equations by applying the divergence to number (iv).

I have written Maxwell's equations in the traditional way, which emphasizes that they specify the divergence and curl of **E** and **B**. In this form, they reinforce the notion that electric fields can be produced *either* by charges (ρ) *or* by changing magnetic fields ($\partial \mathbf{B}/\partial t$), and magnetic fields can be produced *either* by currents ($\mathbf{J}$) *or* by changing electric fields ($\partial \mathbf{E}/\partial t$). Actually, this is misleading, because $\partial \mathbf{B}/\partial t$ and $\partial \mathbf{E}/\partial t$ are *themselves* due to charges and currents. I think it is logically preferable to write

$$\left.\begin{array}{ll} \text{(i)}\ \nabla \cdot \mathbf{E} = \dfrac{1}{\epsilon_0}\rho, & \text{(iii)}\ \nabla \times \mathbf{E} + \dfrac{\partial \mathbf{B}}{\partial t} = \mathbf{0}, \\[2ex] \text{(ii)}\ \nabla \cdot \mathbf{B} = 0, & \text{(iv)}\ \nabla \times \mathbf{B} - \mu_0\epsilon_0\dfrac{\partial \mathbf{E}}{\partial t} = \mu_0\mathbf{J}, \end{array}\right\} \tag{7.43}$$

with the fields (**E** and **B**) on the left and the sources (ρ and **J**) on the right. This notation emphasizes that all electromagnetic fields are ultimately attributable to charges and currents. Maxwell's equations tell you how *charges* produce *fields*; reciprocally, the force law tells you how *fields* affect *charges*.

Problem 7.37 Suppose

$$\mathbf{E}(\mathbf{r}, t) = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\theta(vt - r)\hat{\mathbf{r}};\quad \mathbf{B}(\mathbf{r}, t) = \mathbf{0}$$

(The theta function is defined in Prob. 1.46b). Show that these fields satisfy all of Maxwell's equations, and determine ρ and **J**. Describe the physical situation that gives rise to these fields.

7.3.4 ■ Magnetic Charge

There is a pleasing symmetry to Maxwell's equations; it is particularly striking in free space, where ρ and **J** vanish:

$$\left.\begin{array}{ll} \nabla \cdot \mathbf{E} = 0, & \nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}, \\[2ex] \nabla \cdot \mathbf{B} = 0, & \nabla \times \mathbf{B} = \mu_0\epsilon_0\dfrac{\partial \mathbf{E}}{\partial t}. \end{array}\right\}$$

[23] Like any differential equations, Maxwell's must be supplemented by suitable *boundary conditions*. Because these are typically "obvious" from the context (e.g. **E** and **B** go to zero at large distances from a localized charge distribution), it is easy to forget that they play an essential role.

If you replace $\mathbf{E}$ by $\mathbf{B}$ and $\mathbf{B}$ by $-\mu_0\epsilon_0\mathbf{E}$, the first pair of equations turns into the second, and vice versa. This symmetry[24] between $\mathbf{E}$ and $\mathbf{B}$ is spoiled, though, by the charge term in Gauss's law and the current term in Ampère's law. You can't help wondering why the corresponding quantities are "missing" from $\nabla\cdot\mathbf{B}=0$ and $\nabla\times\mathbf{E}=-\partial\mathbf{B}/\partial t$. What if we had

$$\left.\begin{array}{ll}\text{(i)}\ \nabla\cdot\mathbf{E}=\dfrac{1}{\epsilon_0}\rho_e, & \text{(iii)}\ \nabla\times\mathbf{E}=-\mu_0\mathbf{J}_m-\dfrac{\partial\mathbf{B}}{\partial t},\\[2ex] \text{(ii)}\ \nabla\cdot\mathbf{B}=\mu_0\rho_m, & \text{(iv)}\ \nabla\times\mathbf{B}=\mu_0\mathbf{J}_e+\mu_0\epsilon_0\dfrac{\partial\mathbf{E}}{\partial t}.\end{array}\right\}\tag{7.44}$$

Then ρ_m would represent the density of magnetic "charge," and ρ_e the density of electric charge; $\mathbf{J}_m$ would be the current of magnetic charge, and $\mathbf{J}_e$ the current of electric charge. Both charges would be conserved:

$$\nabla\cdot\mathbf{J}_m=-\frac{\partial\rho_m}{\partial t},\quad\text{and}\quad\nabla\cdot\mathbf{J}_e=-\frac{\partial\rho_e}{\partial t}.\tag{7.45}$$

The former follows by application of the divergence to (iii), the latter by taking the divergence of (iv).

In a sense, Maxwell's equations *beg* for magnetic charge to exist—it would fit in so nicely. And yet, in spite of a diligent search, no one has ever found any.[25] As far as we know, ρ_m is zero everywhere, and so is $\mathbf{J}_m$; $\mathbf{B}$ is *not* on equal footing with $\mathbf{E}$: there exist stationary sources for $\mathbf{E}$ (electric charges) but none for $\mathbf{B}$. (This is reflected in the fact that magnetic multipole expansions have no monopole term, and magnetic dipoles consist of current loops, not separated north and south "poles.") Apparently God just didn't *make* any magnetic charge. (In *quantum* electrodynamics, by the way, it's a more than merely aesthetic shame that magnetic charge does not seem to exist: Dirac showed that the existence of *magnetic* charge would explain why *electric* charge is *quantized*. See Prob. 8.19.)

Problem 7.38 Assuming that "Coulomb's law" for magnetic charges (q_m) reads

$$\mathbf{F}=\frac{\mu_0}{4\pi}\frac{q_{m_1}q_{m_2}}{\mathscr{r}^2}\hat{\boldsymbol{\mathscr{r}}},\tag{7.46}$$

work out the force law for a monopole q_m moving with velocity $\mathbf{v}$ through electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$.[26]

Problem 7.39 Suppose a magnetic monopole q_m passes through a resistanceless loop of wire with self-inductance L. What current is induced in the loop?[27]

[24] Don't be distracted by the pesky constants μ_0 and ϵ_0; these are present only because the SI system measures $\mathbf{E}$ and $\mathbf{B}$ in different units, and would not occur, for instance, in the Gaussian system.

[25] For an extensive bibliography, see A. S. Goldhaber and W. P. Trower, *Am. J. Phys.* **58**, 429 (1990).

[26] For interesting commentary, see W. Rindler, *Am. J. Phys.* **57**, 993 (1989).

[27] This is one of the methods used to search for monopoles in the laboratory; see B. Cabrera, *Phys. Rev. Lett.* **48**, 1378 (1982).

7.3.5 ■ Maxwell's Equations in Matter

Maxwell's equations in the form 7.40 are complete and correct as they stand. However, when you are working with materials that are subject to electric and magnetic polarization there is a more convenient way to *write* them. For inside polarized matter there will be accumulations of "bound" charge and current, over which you exert no direct control. It would be nice to reformulate Maxwell's equations so as to make explicit reference only to the "free" charges and currents.

We have already learned, from the static case, that an electric polarization **P** produces a bound charge density

$$\rho_b = -\nabla \cdot \mathbf{P} \tag{7.47}$$

(Eq. 4.12). Likewise, a magnetic polarization (or "magnetization") **M** results in a bound current

$$\mathbf{J}_b = \nabla \times \mathbf{M} \tag{7.48}$$

(Eq. 6.13). There's just one new feature to consider in the *non*static case: Any *change* in the electric polarization involves a flow of (bound) charge (call it $\mathbf{J}_p$), which must be included in the total current. For suppose we examine a tiny chunk of polarized material (Fig. 7.47). The polarization introduces a charge density $\sigma_b = P$ at one end and $-\sigma_b$ at the other (Eq. 4.11). If P now *increases* a bit, the charge on each end increases accordingly, giving a net current

$$dI = \frac{\partial \sigma_b}{\partial t} da_\perp = \frac{\partial P}{\partial t} da_\perp .$$

The current density, therefore, is

$$\mathbf{J}_p = \frac{\partial \mathbf{P}}{\partial t}. \tag{7.49}$$

This **polarization current** has nothing to do with the *bound* current $\mathbf{J}_b$. The latter is associated with *magnetization* of the material and involves the spin and orbital motion of electrons; $\mathbf{J}_p$, by contrast, is the result of the linear motion of charge when the electric polarization changes. If **P** points to the right, and is increasing, then each plus charge moves a bit to the right and each minus charge to the left; the cumulative effect is the polarization current $\mathbf{J}_p$. We ought to check that Eq. 7.49 is consistent with the continuity equation:

$$\nabla \cdot \mathbf{J}_p = \nabla \cdot \frac{\partial \mathbf{P}}{\partial t} = \frac{\partial}{\partial t}(\nabla \cdot \mathbf{P}) = -\frac{\partial \rho_b}{\partial t}.$$

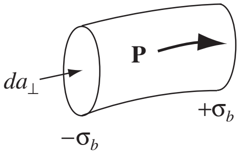


FIGURE 7.47

Yes: The continuity equation *is* satisfied; in fact, $\mathbf{J}_p$ is essential to ensure the conservation of bound charge. (Incidentally, a changing *magnetization* does *not* lead to any analogous accumulation of charge or current. The bound current $\mathbf{J}_b = \nabla \times \mathbf{M}$ varies in response to changes in $\mathbf{M}$, to be sure, but that's about it.)

In view of all this, the total charge density can be separated into two parts:

$$\rho = \rho_f + \rho_b = \rho_f - \nabla \cdot \mathbf{P}, \tag{7.50}$$

and the current density into *three* parts:

$$\mathbf{J} = \mathbf{J}_f + \mathbf{J}_b + \mathbf{J}_p = \mathbf{J}_f + \nabla \times \mathbf{M} + \frac{\partial \mathbf{P}}{\partial t}. \tag{7.51}$$

Gauss's law can now be written as

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}(\rho_f - \nabla \cdot \mathbf{P}),$$

or

$$\nabla \cdot \mathbf{D} = \rho_f, \tag{7.52}$$

where, as in the static case,

$$\mathbf{D} \equiv \epsilon_0 \mathbf{E} + \mathbf{P}. \tag{7.53}$$

Meanwhile, Ampère's law (with Maxwell's term) becomes

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J}_f + \nabla \times \mathbf{M} + \frac{\partial \mathbf{P}}{\partial t} \right) + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t},$$

or

$$\nabla \times \mathbf{H} = \mathbf{J}_f + \frac{\partial \mathbf{D}}{\partial t}, \tag{7.54}$$

where, as before,

$$\mathbf{H} \equiv \frac{1}{\mu_0}\mathbf{B} - \mathbf{M}. \tag{7.55}$$

Faraday's law and $\nabla \cdot \mathbf{B} = 0$ are not affected by our separation of charge and current into free and bound parts, since they do not involve ρ or $\mathbf{J}$.

In terms of *free* charges and currents, then, Maxwell's equations read

$$\boxed{\begin{array}{ll} \text{(i)} \;\; \nabla \cdot \mathbf{D} = \rho_f, & \text{(iii)} \;\; \nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}, \\[2ex] \text{(ii)} \;\; \nabla \cdot \mathbf{B} = 0, & \text{(iv)} \;\; \nabla \times \mathbf{H} = \mathbf{J}_f + \dfrac{\partial \mathbf{D}}{\partial t}. \end{array}} \tag{7.56}$$

Some people regard these as the "true" Maxwell's equations, but please understand that they are in *no way* more "general" than Eq. 7.40; they simply reflect a convenient division of charge and current into free and nonfree parts. And they have the disadvantage of hybrid notation, since they contain both **E** and **D**, both **B** and **H**. They must be supplemented, therefore, by appropriate **constitutive relations**, giving **D** and **H** in terms of **E** and **B**. These depend on the nature of the material; for linear media

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}, \quad \text{and} \quad \mathbf{M} = \chi_m \mathbf{H}, \tag{7.57}$$

so

$$\mathbf{D} = \epsilon \mathbf{E}, \quad \text{and} \quad \mathbf{H} = \frac{1}{\mu}\mathbf{B}, \tag{7.58}$$

where $\epsilon \equiv \epsilon_0(1 + \chi_e)$ and $\mu \equiv \mu_0(1 + \chi_m)$. Incidentally, you'll remember that **D** is called the electric "displacement"; that's why the second term in the Ampère/Maxwell equation (iv) came to be called the **displacement current**. In this context,

$$\mathbf{J}_d \equiv \frac{\partial \mathbf{D}}{\partial t}. \tag{7.59}$$

Problem 7.40 Sea water at frequency $\nu = 4 \times 10^8$ Hz has permittivity $\epsilon = 81\epsilon_0$, permeability $\mu = \mu_0$, and resistivity $\rho = 0.23\ \Omega \cdot \mathrm{m}$. What is the ratio of conduction current to displacement current? [*Hint:* Consider a parallel-plate capacitor immersed in sea water and driven by a voltage $V_0 \cos(2\pi \nu t)$.]

7.3.6 ■ Boundary Conditions

In general, the fields **E, B, D,** and **H** will be discontinuous at a boundary between two different media, or at a surface that carries a charge density σ or a current density **K**. The explicit form of these discontinuities can be deduced from Maxwell's equations (7.56), in their integral form

$$\left.\begin{aligned} &\text{(i)} \quad \oint_{\mathcal{S}} \mathbf{D} \cdot d\mathbf{a} = Q_{f_{\text{enc}}} \\ &\text{(ii)} \quad \oint_{\mathcal{S}} \mathbf{B} \cdot d\mathbf{a} = 0 \end{aligned}\right\} \text{ over any closed surface } \mathcal{S}.$$

$$\left.\begin{aligned} &\text{(iii)} \quad \oint_{\mathcal{P}} \mathbf{E} \cdot d\mathbf{l} = -\frac{d}{dt}\int_{\mathcal{S}} \mathbf{B} \cdot d\mathbf{a} \\ &\text{(iv)} \quad \oint_{\mathcal{P}} \mathbf{H} \cdot d\mathbf{l} = I_{f_{\text{enc}}} + \frac{d}{dt}\int_{\mathcal{S}} \mathbf{D} \cdot d\mathbf{a} \end{aligned}\right\} \begin{array}{l}\text{for any surface } \mathcal{S} \\ \text{bounded by the} \\ \text{closed loop } \mathcal{P}.\end{array}$$

Applying (i) to a tiny, wafer-thin Gaussian pillbox extending just slightly into the material on either side of the boundary (Fig. 7.48), we obtain:

$$\mathbf{D}_1 \cdot \mathbf{a} - \mathbf{D}_2 \cdot \mathbf{a} = \sigma_f\, a.$$

(The positive direction for **a** is *from* 2 *toward* 1. The edge of the wafer contributes nothing in the limit as the thickness goes to zero; nor does any *volume*

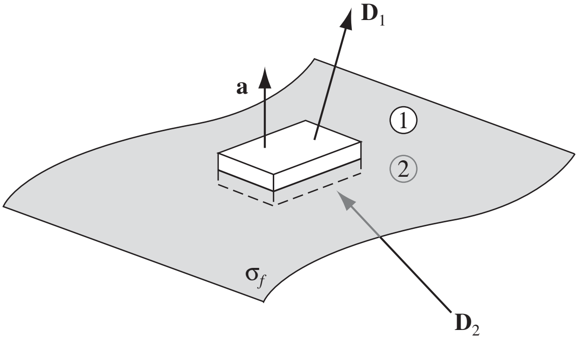


FIGURE 7.48

charge density.) Thus, the component of **D** that is perpendicular to the interface is discontinuous in the amount

$$\boxed{D_1^{\perp} - D_2^{\perp} = \sigma_f.} \tag{7.60}$$

Identical reasoning, applied to equation (ii), yields

$$\boxed{B_1^{\perp} - B_2^{\perp} = 0.} \tag{7.61}$$

Turning to (iii), a very thin Amperian loop straddling the surface gives

$$\mathbf{E}_1 \cdot \mathbf{l} - \mathbf{E}_2 \cdot \mathbf{l} = -\frac{d}{dt}\int_{\mathcal{S}} \mathbf{B} \cdot d\mathbf{a}.$$

But in the limit as the width of the loop goes to zero, the flux vanishes. (I have already dropped the contribution of the two ends to $\oint \mathbf{E} \cdot d\mathbf{l}$, on the same grounds.) Therefore,

$$\boxed{\mathbf{E}_1^{\parallel} - \mathbf{E}_2^{\parallel} = \mathbf{0}.} \tag{7.62}$$

That is, the components of **E** *parallel* to the interface are continuous across the boundary. By the same token, (iv) implies

$$\mathbf{H}_1 \cdot \mathbf{l} - \mathbf{H}_2 \cdot \mathbf{l} = I_{f_{\text{enc}}},$$

where $I_{f_{\text{enc}}}$ is the free current passing through the Amperian loop. No *volume* current density will contribute (in the limit of infinitesimal width), but a *surface* current can. In fact, if $\hat{\mathbf{n}}$ is a unit vector perpendicular to the interface (pointing from 2 toward 1), so that $(\hat{\mathbf{n}} \times \mathbf{l})$ is normal to the Amperian loop (Fig. 7.49), then

$$I_{f_{\text{enc}}} = \mathbf{K}_f \cdot (\hat{\mathbf{n}} \times \mathbf{l}) = (\mathbf{K}_f \times \hat{\mathbf{n}}) \cdot \mathbf{l},$$

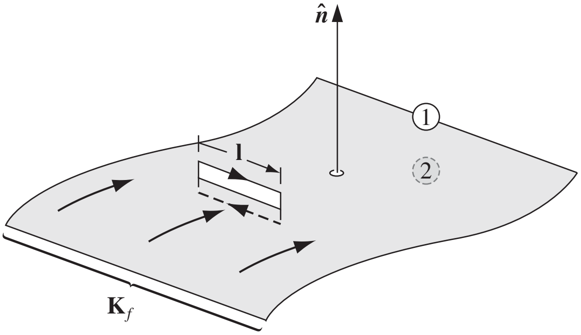


FIGURE 7.49

and hence

$$\boxed{\mathbf{H}_1^{\|} - \mathbf{H}_2^{\|} = \mathbf{K}_f \times \hat{\mathbf{n}}.} \tag{7.63}$$

So the *parallel* components of **H** are discontinuous by an amount proportional to the free surface current density.

Equations 7.60-63 are the general boundary conditions for electrodynamics. In the case of *linear* media, they can be expressed in terms of **E** and **B** alone:

$$\left.\begin{array}{ll} \text{(i)}\ \epsilon_1 E_1^{\perp} - \epsilon_2 E_2^{\perp} = \sigma_f, & \text{(iii)}\ \mathbf{E}_1^{\|} - \mathbf{E}_2^{\|} = \mathbf{0}, \\ \text{(ii)}\ B_1^{\perp} - B_2^{\perp} = 0, & \text{(iv)}\ \dfrac{1}{\mu_1}\mathbf{B}_1^{\|} - \dfrac{1}{\mu_2}\mathbf{B}_2^{\|} = \mathbf{K}_f \times \hat{\mathbf{n}}. \end{array}\right\} \tag{7.64}$$

In particular, if there is no free charge or free current at the interface, then

$$\begin{array}{ll} \text{(i)}\ \epsilon_1 E_1^{\perp} - \epsilon_2 E_2^{\perp} = 0, & \text{(iii)}\ \mathbf{E}_1^{\|} - \mathbf{E}_2^{\|} = \mathbf{0}, \\ \text{(ii)}\ B_1^{\perp} - B_2^{\perp} = 0, & \text{(iv)}\ \dfrac{1}{\mu_1}\mathbf{B}_1^{\|} - \dfrac{1}{\mu_2}\mathbf{B}_2^{\|} = \mathbf{0}. \end{array} \tag{7.65}$$

As we shall see in Chapter 9, these equations are the basis for the theory of reflection and refraction.

More Problems on Chapter 7

! **Problem 7.41** Two long, straight copper pipes, each of radius a, are held a distance $2d$ apart (see Fig. 7.50). One is at potential V_0, the other at $-V_0$. The space surrounding the pipes is filled with weakly conducting material of conductivity σ. Find the current per unit length that flows from one pipe to the other. [*Hint:* Refer to Prob. 3.12.]

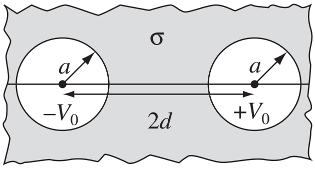


FIGURE 7.50

! **Problem 7.42** A rare case in which the electrostatic field **E** for a circuit can actually be *calculated* is the following:[28] Imagine an infinitely long cylindrical sheet, of uniform resistivity and radius a. A slot (corresponding to the battery) is maintained at $\pm V_0/2$, at $\phi = \pm\pi$, and a steady current flows over the surface, as indicated in Fig. 7.51. According to Ohm's law, then,

$$V(a, \phi) = \frac{V_0\phi}{2\pi}, \quad (-\pi < \phi < +\pi).$$

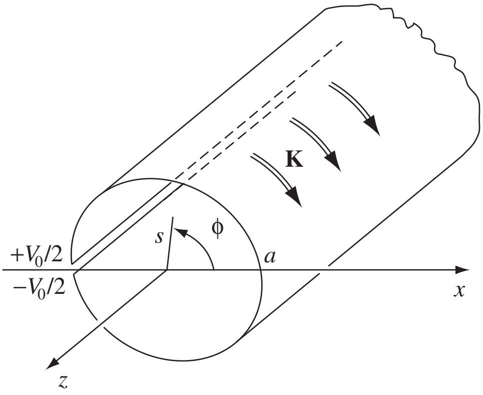


FIGURE 7.51

(a) Use separation of variables in cylindrical coordinates to determine $V(s, \phi)$ inside and outside the cylinder. [*Answer:* $(V_0/\pi)\tan^{-1}[(s\sin\phi)/(a + s\cos\phi)]$, $(s < a)$; $(V_0/\pi)\tan^{-1}[(a\sin\phi)/(s + a\cos\phi)]$, $(s > a)$]

(b) Find the surface charge density on the cylinder. [*Answer:* $(\epsilon_0 V_0/\pi a)\tan(\phi/2)$]

Problem 7.43 The magnetic field outside a long straight wire carrying a steady current I is

$$\mathbf{B} = \frac{\mu_0}{2\pi}\frac{I}{s}\,\hat{\boldsymbol{\phi}}.$$

The *electric field inside* the wire is uniform:

$$\mathbf{E} = \frac{I\rho}{\pi a^2}\,\hat{\mathbf{z}},$$

[28] M. A. Heald, *Am. J. Phys.* **52**, 522 (1984). See also J. A. Hernandes and A. K. T. Assis, *Phys. Rev. E* **68**, 046611 (2003).

where ρ is the resistivity and a is the radius (see Exs. 7.1 and 7.3). *Question:* What is the electric field *outside* the wire?[29] The answer depends on how you complete the circuit. Suppose the current returns along a perfectly conducting grounded coaxial cylinder of radius b (Fig. 7.52). In the region $a < s < b$, the potential $V(s, z)$ satisfies Laplace's equation, with the boundary conditions

$$\text{(i)}\ \ V(a, z) = -\frac{I\rho z}{\pi a^2};\quad \text{(ii)}\ \ V(b, z) = 0.$$

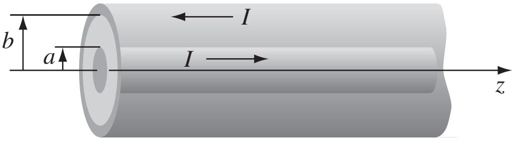


FIGURE 7.52

This does not suffice to determine the answer—we still need to specify boundary conditions at the two ends (though for a *long* wire it shouldn't matter much). In the literature, it is customary to sweep this ambiguity under the rug by simply *stipulating* that $V(s, z)$ is proportional to z: $V(s, z) = zf(s)$. On this assumption:

(a) Determine $f(s)$.

(b) Find $\mathbf{E}(s, z)$.

(c) Calculate the surface charge density $\sigma(z)$ on the wire.

[*Answer:* $V = (-Iz\rho/\pi a^2)[\ln(s/b)/\ln(a/b)]$ This is a peculiar result, since E_s and $\sigma(z)$ are *not* independent of z—as one would certainly expect for a truly *infinite* wire.]

Problem 7.44 In a **perfect conductor**, the conductivity is infinite, so $\mathbf{E} = \mathbf{0}$ (Eq. 7.3), and any net charge resides on the surface (just as it does for an *im*perfect conductor, in electro*statics*).

(a) Show that the magnetic field is constant ($\partial\mathbf{B}/\partial t = \mathbf{0}$), inside a perfect conductor.

(b) Show that the magnetic flux through a perfectly conducting loop is constant.

A **superconductor** is a perfect conductor with the additional property that the (constant) **B** inside is in fact *zero*. (This "flux exclusion" is known as the **Meissner effect**.[30])

[29] This is a famous problem, first analyzed by Sommerfeld, and is known in its most recent incarnation as **Merzbacher's puzzle**. A. Sommerfeld, *Electrodynamics*, p. 125 (New York: Academic Press, 1952); E. Merzbacher, *Am. J. Phys.* **48**, 178 (1980); further references in R. N. Varnay and L. H. Fisher, *Am. J. Phys.* **52**, 1097 (1984).

[30] The Meissner effect is sometimes referred to as "perfect diamagnetism," in the sense that the field inside is not merely *reduced*, but canceled entirely. However, the surface currents responsible for this are *free*, not bound, so the actual *mechanism* is quite different.

(c) Show that the current in a superconductor is confined to the surface.

(d) Superconductivity is lost above a certain critical temperature (T_c), which varies from one material to another. Suppose you had a sphere (radius a) above its critical temperature, and you held it in a uniform magnetic field $B_0\hat{\mathbf{z}}$ while cooling it below T_c. Find the induced surface current density $\mathbf{K}$, as a function of the polar angle θ.

Problem 7.45 A familiar demonstration of superconductivity (Prob. 7.44) is the levitation of a magnet over a piece of superconducting material. This phenomenon can be analyzed using the method of images.[31] Treat the magnet as a perfect dipole $\mathbf{m}$, a height z above the origin (and constrained to point in the z direction), and pretend that the superconductor occupies the entire half-space below the xy plane. Because of the Meissner effect, $\mathbf{B} = \mathbf{0}$ for $z \leq 0$, and since $\mathbf{B}$ is divergenceless, the normal (z) component is continuous, so $B_z = 0$ just *above* the surface. This boundary condition is met by the image configuration in which an identical dipole is placed at $-z$, as a stand-in for the superconductor; the two arrangements therefore produce the same magnetic field in the region $z > 0$.

(a) Which way should the image dipole point ($+z$ or $-z$)?

(b) Find the force on the magnet due to the induced currents in the superconductor (which is to say, the force due to the image dipole). Set it equal to Mg (where M is the mass of the magnet) to determine the height h at which the magnet will "float." [*Hint:* Refer to Prob. 6.3.]

(c) The induced current on the surface of the superconductor (the xy plane) can be determined from the boundary condition on the *tangential* component of $\mathbf{B}$ (Eq. 5.76): $\mathbf{B} = \mu_0(\mathbf{K} \times \hat{\mathbf{z}})$. Using the field you get from the image configuration, show that

$$\mathbf{K} = -\frac{3mrh}{2\pi(r^2 + h^2)^{5/2}}\,\hat{\boldsymbol{\phi}},$$

where r is the distance from the origin.

! **Problem 7.46** If a magnetic dipole levitating above an infinite superconducting plane (Prob. 7.45) is free to rotate, what orientation will it adopt, and how high above the surface will it float?

Problem 7.47 A perfectly conducting spherical shell of radius a rotates about the z axis with angular velocity ω, in a uniform magnetic field $\mathbf{B} = B_0\,\hat{\mathbf{z}}$. Calculate the emf developed between the "north pole" and the equator. [*Answer:* $\frac{1}{2}B_0\omega a^2$]

! **Problem 7.48** Refer to Prob. 7.11 (and use the result of Prob. 5.42): How long does is take a falling *circular* ring (radius a, mass m, resistance R) to cross the bottom of the magnetic field B, at its (changing) terminal velocity?

[31] W. M. Saslow, *Am. J. Phys.* **59**, 16 (1991).

Problem 7.49

(a) Referring to Prob. 5.52(a) and Eq. 7.18, show that

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t}, \tag{7.66}$$

for Faraday-induced electric fields. Check this result by taking the divergence and curl of both sides.

(b) A spherical shell of radius R carries a uniform surface charge σ. It spins about a fixed axis at an angular velocity $\omega(t)$ that changes slowly with time. Find the electric field inside and outside the sphere. [*Hint:* There are *two* contributions here: the Coulomb field due to the charge, and the Faraday field due to the changing $\mathbf{B}$. Refer to Ex. 5.11.]

Problem 7.50 Electrons undergoing cyclotron motion can be sped up by increasing the magnetic field; the accompanying electric field will impart tangential acceleration. This is the principle of the **betatron**. One would like to keep the radius of the orbit constant during the process. Show that this can be achieved by designing a magnet such that the average field over the area of the orbit is twice the field at the circumference (Fig. 7.53). Assume the electrons start from rest in zero field, and that the apparatus is symmetric about the center of the orbit. (Assume also that the electron velocity remains well below the speed of light, so that nonrelativistic mechanics applies.) [*Hint:* Differentiate Eq. 5.3 with respect to time, and use $F = ma = qE$.]

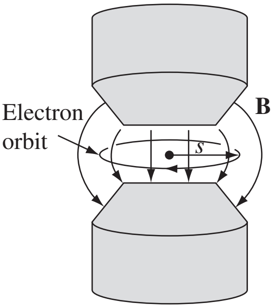


FIGURE 7.53

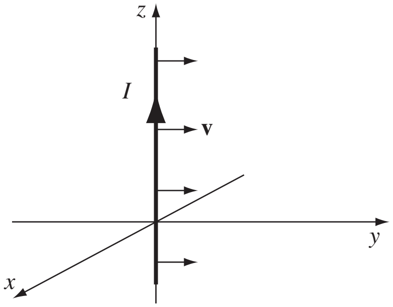


FIGURE 7.54

Problem 7.51 An infinite wire carrying a constant current I in the $\hat{\mathbf{z}}$ direction is moving in the y direction at a constant speed v. Find the electric field, in the quasistatic approximation, at the instant the wire coincides with the z axis (Fig. 7.54). [*Answer:* $-(\mu_0 I v/2\pi s)\sin\phi\,\hat{\mathbf{z}}$]

Problem 7.52 An atomic electron (charge q) circles about the nucleus (charge Q) in an orbit of radius r; the centripetal acceleration is provided, of course, by the Coulomb attraction of opposite charges. Now a small magnetic field dB is slowly turned on, perpendicular to the plane of the orbit. Show that the increase in kinetic energy, dT, imparted by the induced electric field, is just right to sustain circular motion *at the same radius r*. (That's why, in my discussion of diamagnetism, I assumed the radius is fixed. See Sect. 6.1.3 and the references cited there.)

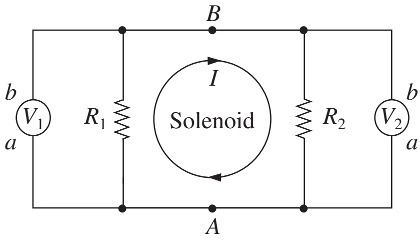


FIGURE 7.55

Problem 7.53 The current in a long solenoid is increasing linearly with time, so the flux is proportional to t: $\Phi = \alpha t$. Two voltmeters are connected to diametrically opposite points (A and B), together with resistors (R_1 and R_2), as shown in Fig. 7.55. What is the reading on each voltmeter? Assume that these are *ideal* voltmeters that draw negligible current (they have huge internal resistance), and that a voltmeter registers $-\int_a^b \mathbf{E} \cdot d\mathbf{l}$ between the terminals and through the meter. [*Answer:* $V_1 = \alpha R_1/(R_1 + R_2)$; $V_2 = -\alpha R_2/(R_1 + R_2)$. Notice that $V_1 \neq V_2$, even though they are connected to the same points![32]]

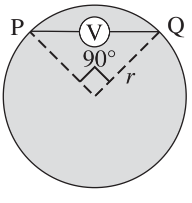


FIGURE 7.56

Problem 7.54 A circular wire loop (radius r, resistance R) encloses a region of uniform magnetic field, B, perpendicular to its plane. The field (occupying the shaded region in Fig. 7.56) increases linearly with time ($B = \alpha t$). An ideal voltmeter (infinite internal resistance) is connected between points P and Q.

(a) What is the current in the loop?

(b) What does the voltmeter read? [*Answer:* $\alpha r^2/2$]

Problem 7.55 In the discussion of motional emf (Sect. 7.1.3) I assumed that the wire loop (Fig. 7.10) has a resistance R; the current generated is then $I = vBh/R$. But what if the wire is made out of perfectly conducting material, so that R is *zero*? In that case, the current is limited only by the back emf associated with the self-inductance L of the loop (which would ordinarily be negligible in comparison with IR). Show that in this régime the loop (mass m) executes simple harmonic motion, and find its frequency.[33] [*Answer:* $\omega = Bh/\sqrt{mL}$]

[32] R. H. Romer, *Am. J. Phys.* **50**, 1089 (1982). See also H. W. Nicholson, *Am. J. Phys.* **73**, 1194 (2005); B. M. McGuyer, *Am. J. Phys.* **80**, 101 (2012).

[33] For a collection of related problems, see W. M. Saslow, *Am. J. Phys.* **55**, 986 (1987), and R. H. Romer, *Eur. J. Phys.* **11**, 103 (1990).

Problem 7.56

(a) Use the Neumann formula (Eq. 7.23) to calculate the mutual inductance of the configuration in Fig. 7.37, assuming a is very small ($a \ll b, a \ll z$). Compare your answer to Prob. 7.22.

(b) For the general case (*not* assuming a is small), show that

$$M = \frac{\mu_0 \pi \beta}{2}\sqrt{ab\beta}\left(1 + \frac{15}{8}\beta^2 + \dots\right),$$

where

$$\beta \equiv \frac{ab}{z^2 + a^2 + b^2}.$$

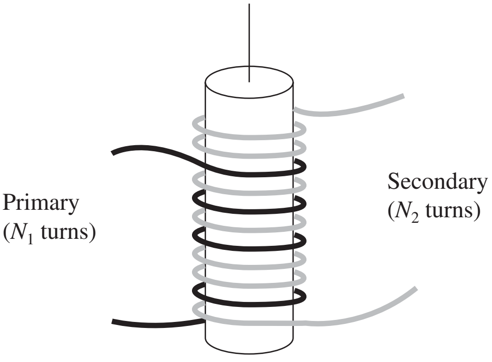


FIGURE 7.57

Problem 7.57 Two coils are wrapped around a cylindrical form in such a way that the *same flux passes through every turn of both coils*. (In practice this is achieved by inserting an iron core through the cylinder; this has the effect of concentrating the flux.) The **primary** coil has N_1 turns and the **secondary** has N_2 (Fig. 7.57). If the current I in the primary is changing, show that the emf in the secondary is given by

$$\frac{\mathcal{E}_2}{\mathcal{E}_1} = \frac{N_2}{N_1}, \tag{7.67}$$

where $\mathcal{E}_1$ is the (back) emf of the primary. [This is a primitive **transformer**—a device for raising or lowering the emf of an alternating current source. By choosing the appropriate number of turns, any desired secondary emf can be obtained. If you think this violates the conservation of energy, study Prob. 7.58.]

Problem 7.58 A transformer (Prob. 7.57) takes an input AC voltage of amplitude V_1, and delivers an output voltage of amplitude V_2, which is determined by the turns ratio ($V_2/V_1 = N_2/N_1$). If $N_2 > N_1$, the output voltage is greater than the input voltage. Why doesn't this violate conservation of energy? *Answer:* Power is the product of voltage and current; if the voltage goes *up*, the current must come *down*. The purpose of this problem is to see exactly how this works out, in a simplified model.

(a) In an ideal transformer, the same flux passes through all turns of the primary and of the secondary. Show that in this case $M^2 = L_1 L_2$, where M is the mutual inductance of the coils, and L_1, L_2 are their individual self-inductances.

(b) Suppose the primary is driven with AC voltage $V_{\text{in}} = V_1 \cos(\omega t)$, and the secondary is connected to a resistor, R. Show that the two currents satisfy the relations

$$L_1 \frac{dI_1}{dt} + M \frac{dI_2}{dt} = V_1 \cos(\omega t); \qquad L_2 \frac{dI_2}{dt} + M \frac{dI_1}{dt} = -I_2 R.$$

(c) Using the result in (a), solve these equations for $I_1(t)$ and $I_2(t)$. (Assume I_1 has no DC component.)

(d) Show that the output voltage ($V_{\text{out}} = I_2 R$) divided by the input voltage (V_{in}) is equal to the turns ratio: $V_{\text{out}}/V_{\text{in}} = N_2/N_1$.

(e) Calculate the input power ($P_{\text{in}} = V_{\text{in}} I_1$) and the output power ($P_{\text{out}} = V_{\text{out}} I_2$), and show that their averages over a full cycle are equal.

Problem 7.59 An infinite wire runs along the z axis; it carries a current $I(z)$ that is a function of z (but not of t), and a charge density $\lambda(t)$ that is a function of t (but not of z).

(a) By examining the charge flowing into a segment dz in a time dt, show that $d\lambda/dt = -dI/dz$. If we stipulate that $\lambda(0) = 0$ and $I(0) = 0$, show that $\lambda(t) = kt,\ I(z) = -kz$, where k is a constant.

(b) Assume for a moment that the process is quasistatic, so the fields are given by Eqs. 2.9 and 5.38. Show that these are in fact the *exact* fields, by confirming that all four of Maxwell's equations are satisfied. (First do it in differential form, for the region $s > 0$, then in integral form for the appropriate Gaussian cylinder/Amperian loop straddling the axis.)

Problem 7.60 Suppose $\mathbf{J}(\mathbf{r})$ is constant in time but $\rho(\mathbf{r}, t)$ is *not*—conditions that might prevail, for instance, during the charging of a capacitor.

(a) Show that the charge density at any particular point is a linear function of time:

$$\rho(\mathbf{r}, t) = \rho(\mathbf{r}, 0) + \dot{\rho}(\mathbf{r}, 0)t,$$

where $\dot{\rho}(\mathbf{r}, 0)$ is the time derivative of ρ at $t = 0$. [*Hint:* Use the continuity equation.]

This is *not* an electrostatic or magnetostatic configuration;[34] nevertheless, rather surprisingly, both Coulomb's law (Eq. 2.8) and the Biot-Savart law (Eq. 5.42) hold, as you can confirm by showing that they satisfy Maxwell's equations. In particular:

[34] Some authors *would* regard this as magnetostatic, since $\mathbf{B}$ is independent of t. For them, the Biot-Savart law is a general rule of magnetostatics, but $\nabla \cdot \mathbf{J} = 0$ and $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$ apply only under the *additional* assumption that ρ is constant. In such a formulation, Maxwell's displacement term can (in this very special case) be *derived* from the Biot-Savart law, by the method of part (b). See D. F. Bartlett, *Am. J. Phys.* **58**, 1168 (1990); D. J. Griffiths and M. A. Heald, *Am. J. Phys.* **59**, 111 (1991).

(b) Show that

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}') \times \hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\, d\tau'$$

obeys Ampère's law *with Maxwell's displacement current term.*

Problem 7.61 The magnetic field of an infinite straight wire carrying a steady current I can be obtained from the *displacement* current term in the Ampère/Maxwell law, as follows: Picture the current as consisting of a uniform line charge λ moving along the z axis at speed v (so that $I = \lambda v$), with a tiny gap of length ϵ, which reaches the origin at time $t = 0$. In the next instant (up to $t = \epsilon/v$) there is no *real* current passing through a circular Amperian loop in the xy plane, but there *is* a *displacement* current, due to the "missing" charge in the gap.

(a) Use Coulomb's law to calculate the z component of the electric field, for points in the xy plane a distance s from the origin, due to a segment of wire with uniform density $-\lambda$ extending from $z_1 = vt - \epsilon$ to $z_2 = vt$.

(b) Determine the flux of this electric field through a circle of radius a in the xy plane.

(c) Find the displacement current through this circle. Show that I_d is equal to I, in the limit as the gap width (ϵ) goes to zero.[35]

Problem 7.62 A certain transmission line is constructed from two thin metal "ribbons," of width w, a very small distance $h \ll w$ apart. The current travels down one strip and back along the other. In each case, it spreads out uniformly over the surface of the ribbon.

(a) Find the capacitance per unit length, $\mathcal{C}$.

(b) Find the inductance per unit length, $\mathcal{L}$.

(c) What is the product $\mathcal{LC}$, numerically? [$\mathcal{L}$ and $\mathcal{C}$ will, of course, vary from one kind of transmission line to another, but their *product* is a universal constant—check, for example, the cable in Ex. 7.13—provided the space between the conductors is a vacuum. In the theory of transmission lines, this product is related to the speed with which a pulse propagates down the line: $v = 1/\sqrt{\mathcal{LC}}$.]

(d) If the strips are insulated from one another by a nonconducting material of permittivity ϵ and permeability μ, what then is the product $\mathcal{LC}$? What is the propagation speed? [*Hint:* see Ex. 4.6; by what factor does L change when an inductor is immersed in linear material of permeability μ?]

Problem 7.63 Prove **Alfven's theorem**: In a perfectly conducting fluid (say, a gas of free electrons), the magnetic flux through any closed loop moving with the fluid is constant in time. (The magnetic field lines are, as it were, "frozen" into the fluid.)

(a) Use Ohm's law, in the form of Eq. 7.2, together with Faraday's law, to prove that if $\sigma = \infty$ and $\mathbf{J}$ is finite, then

$$\frac{\partial \mathbf{B}}{\partial t} = \boldsymbol{\nabla} \times (\mathbf{v} \times \mathbf{B}).$$

[35] For a slightly different approach to the same problem, see W. K. Terry, *Am. J. Phys.* **50**, 742 (1982).

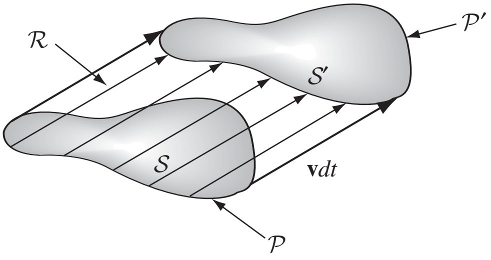


FIGURE 7.58

(b) Let $\mathcal{S}$ be the surface bounded by the loop ($\mathcal{P}$) at time t, and $\mathcal{S}'$ a surface bounded by the loop in its new position ($\mathcal{P}'$) at time $t + dt$ (see Fig. 7.58). The change in flux is

$$d\Phi = \int_{\mathcal{S}'} \mathbf{B}(t + dt) \cdot d\mathbf{a} - \int_{\mathcal{S}} \mathbf{B}(t) \cdot d\mathbf{a}.$$

Use $\nabla \cdot \mathbf{B} = 0$ to show that

$$\int_{\mathcal{S}'} \mathbf{B}(t + dt) \cdot d\mathbf{a} + \int_{\mathcal{R}} \mathbf{B}(t + dt) \cdot d\mathbf{a} = \int_{\mathcal{S}} \mathbf{B}(t + dt) \cdot d\mathbf{a}$$

(where $\mathcal{R}$ is the "ribbon" joining $\mathcal{P}$ and $\mathcal{P}'$), and hence that

$$d\Phi = dt \int_{\mathcal{S}} \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} - \int_{\mathcal{R}} \mathbf{B}(t + dt) \cdot d\mathbf{a}$$

(for infinitesimal dt). Use the method of Sect. 7.1.3 to rewrite the second integral as

$$dt \oint_{\mathcal{P}} (\mathbf{B} \times \mathbf{v}) \cdot d\mathbf{l},$$

and invoke Stokes' theorem to conclude that

$$\frac{d\Phi}{dt} = \int_{\mathcal{S}} \left(\frac{\partial \mathbf{B}}{\partial t} - \nabla \times (\mathbf{v} \times \mathbf{B}) \right) \cdot d\mathbf{a}.$$

Together with the result in (a), this proves the theorem.

Problem 7.64

(a) Show that Maxwell's equations with magnetic charge (Eq. 7.44) are invariant under the **duality transformation**

$$\left.\begin{aligned} \mathbf{E}' &= \mathbf{E}\cos\alpha + c\mathbf{B}\sin\alpha, \\ c\mathbf{B}' &= c\mathbf{B}\cos\alpha - \mathbf{E}\sin\alpha, \\ cq_e' &= cq_e\cos\alpha + q_m\sin\alpha, \\ q_m' &= q_m\cos\alpha - cq_e\sin\alpha, \end{aligned}\right\} \tag{7.68}$$

where $c \equiv 1/\sqrt{\epsilon_0\mu_0}$ and α is an arbitrary rotation angle in "**E/B**-space." Charge and current densities transform in the same way as q_e and q_m. [This means, in particular, that if you know the fields produced by a configuration of *electric* charge, you can immediately (using $\alpha = 90°$) write down the fields produced by the corresponding arrangement of *magnetic* charge.]

(b) Show that the force law (Prob. 7.38)

$$\mathbf{F} = q_e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + q_m\left(\mathbf{B} - \frac{1}{c^2}\mathbf{v} \times \mathbf{E}\right) \tag{7.69}$$

is also invariant under the duality transformation.

Intermission

All of our cards are now on the table, and in a sense my job is done. In the first seven chapters we assembled electrodynamics piece by piece, and now, with Maxwell's equations in their final form, the theory is complete. There are no more laws to be learned, no further generalizations to be considered, and (with perhaps one exception) no lurking inconsistencies to be resolved. If yours is a one-semester course, this would be a reasonable place to stop.

But in another sense we have just arrived at the starting point. We are at last in possession of a full deck—it's time to deal. This is the fun part, in which one comes to appreciate the extraordinary power and richness of electrodynamics. In a full-year course there should be plenty of time to cover the remaining chapters, and perhaps to supplement them with a unit on plasma physics, say, or AC circuit theory, or even a little general relativity. But if you have room for only one topic, I'd recommend Chapter 9, on Electromagnetic Waves (you'll probably want to skim Chapter 8 as preparation). This is the segue to Optics, and is historically the most important application of Maxwell's theory.

CHAPTER

8 Conservation Laws

8.1 ■ CHARGE AND ENERGY

8.1.1 ■ The Continuity Equation

In this chapter we study conservation of energy, momentum, and angular momentum, in electrodynamics. But I want to begin by reviewing the conservation of *charge*, because it is the paradigm for all conservation laws. What precisely does conservation of charge tell us? That the total charge in the universe is constant? Well, sure—that's **global** conservation of charge. But **local** conservation of charge is a much stronger statement: If the charge in some region changes, then exactly that amount of charge must have passed in or out through the surface. The tiger can't simply rematerialize outside the cage; if it got from inside to outside it must have slipped through a hole in the fence.

Formally, the charge in a volume $\mathcal{V}$ is

$$Q(t) = \int_{\mathcal{V}} \rho(\mathbf{r}, t)\, d\tau, \tag{8.1}$$

and the current flowing out through the boundary $\mathcal{S}$ is $\oint_{\mathcal{S}} \mathbf{J} \cdot d\mathbf{a}$, so local conservation of charge says

$$\frac{dQ}{dt} = -\oint_{\mathcal{S}} \mathbf{J} \cdot d\mathbf{a}. \tag{8.2}$$

Using Eq. 8.1 to rewrite the left side, and invoking the divergence theorem on the right, we have

$$\int_{\mathcal{V}} \frac{\partial \rho}{\partial t}\, d\tau = -\int_{\mathcal{V}} \nabla \cdot \mathbf{J}\, d\tau, \tag{8.3}$$

and since this is true for *any* volume, it follows that

$$\boxed{\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{J}.} \tag{8.4}$$

This is the continuity equation—the precise mathematical statement of local conservation of charge. It can be derived from Maxwell's equations—conservation of charge is not an *independent* assumption; it is built into the laws of electrodynamics. It serves as a constraint on the sources (ρ and $\mathbf{J}$). They can't be just *any* old functions—they have to respect conservation of charge.[1]

The purpose of this chapter is to develop the corresponding equations for local conservation of energy and momentum. In the process (and perhaps more important) we will learn how to express the energy density and the momentum density (the analogs to ρ), as well as the energy "current" and the momentum "current" (analogous to $\mathbf{J}$).

8.1.2 ■ Poynting's Theorem

In Chapter 2, we found that the work necessary to assemble a static charge distribution (against the Coulomb repulsion of like charges) is (Eq. 2.45)

$$W_{\mathrm{e}} = \frac{\epsilon_0}{2}\int E^2\,d\tau,$$

where $\mathbf{E}$ is the resulting electric field. Likewise, the work required to get currents going (against the back emf) is (Eq. 7.35)

$$W_{\mathrm{m}} = \frac{1}{2\mu_0}\int B^2\,d\tau,$$

where $\mathbf{B}$ is the resulting magnetic field. This suggests that the total energy stored in electromagnetic fields, per unit volume, is

$$\boxed{u = \frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right).} \qquad (8.5)$$

In this section I will confirm Eq. 8.5, and develop the energy conservation law for electrodynamics.

Suppose we have some charge and current configuration which, at time t, produces fields $\mathbf{E}$ and $\mathbf{B}$. In the next instant, dt, the charges move around a bit. *Question*: How much work, dW, is done by the electromagnetic forces acting on these charges, in the interval dt? According to the Lorentz force law, the work done on a charge q is

$$\mathbf{F}\cdot d\mathbf{l} = q(\mathbf{E}+\mathbf{v}\times\mathbf{B})\cdot\mathbf{v}\,dt = q\mathbf{E}\cdot\mathbf{v}\,dt.$$

In terms of the charge and current densities, $q \to \rho d\tau$ and $\rho\mathbf{v} \to \mathbf{J}$,[2] so the rate at which work is done on all the charges in a volume $\mathcal{V}$ is

$$\frac{dW}{dt} = \int_{\mathcal{V}} (\mathbf{E}\cdot\mathbf{J})\,d\tau. \qquad (8.6)$$

[1] The continuity equation is the *only* such constraint. Any functions $\rho(\mathbf{r}, t)$ and $\mathbf{J}(\mathbf{r}, t)$ consistent with Eq. 8.4 constitute possible charge and current densities, in the sense of admitting solutions to Maxwell's equations.

[2] This is a slippery equation: after all, if charges of both signs are present, the *net* charge density can be zero even when the current is *not*—in fact, this is the case for ordinary current-carrying wires. We should really treat the positive and negative charges separately, and combine the two to get Eq. 8.6, with $\mathbf{J} = \rho_+\mathbf{v}_+ + \rho_-\mathbf{v}_-$.

Evidently $\mathbf{E}\cdot\mathbf{J}$ is the work done per unit time, per unit volume—which is to say, the *power* delivered per unit volume. We can express this quantity in terms of the fields alone, using the Ampère-Maxwell law to eliminate $\mathbf{J}$:

$$\mathbf{E}\cdot\mathbf{J} = \frac{1}{\mu_0}\mathbf{E}\cdot(\nabla\times\mathbf{B}) - \epsilon_0\mathbf{E}\cdot\frac{\partial\mathbf{E}}{\partial t}.$$

From product rule 6,

$$\nabla\cdot(\mathbf{E}\times\mathbf{B}) = \mathbf{B}\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot(\nabla\times\mathbf{B}).$$

Invoking Faraday's law ($\nabla\times\mathbf{E} = -\partial\mathbf{B}/\partial t$), it follows that

$$\mathbf{E}\cdot(\nabla\times\mathbf{B}) = -\mathbf{B}\cdot\frac{\partial\mathbf{B}}{\partial t} - \nabla\cdot(\mathbf{E}\times\mathbf{B}).$$

Meanwhile,

$$\mathbf{B}\cdot\frac{\partial\mathbf{B}}{\partial t} = \frac{1}{2}\frac{\partial}{\partial t}(B^2), \quad \text{and} \quad \mathbf{E}\cdot\frac{\partial\mathbf{E}}{\partial t} = \frac{1}{2}\frac{\partial}{\partial t}(E^2), \tag{8.7}$$

so

$$\mathbf{E}\cdot\mathbf{J} = -\frac{1}{2}\frac{\partial}{\partial t}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right) - \frac{1}{\mu_0}\nabla\cdot(\mathbf{E}\times\mathbf{B}). \tag{8.8}$$

Putting this into Eq. 8.6, and applying the divergence theorem to the second term, we have

$$\frac{dW}{dt} = -\frac{d}{dt}\int_{\mathcal{V}}\frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right)d\tau - \frac{1}{\mu_0}\oint_{\mathcal{S}}(\mathbf{E}\times\mathbf{B})\cdot d\mathbf{a}, \tag{8.9}$$

where $\mathcal{S}$ is the surface bounding $\mathcal{V}$. This is **Poynting's theorem**; it is the "work-energy theorem" of electrodynamics. The first integral on the right is the total energy stored in the fields, $\int u\,d\tau$ (Eq. 8.5). The second term evidently represents the rate at which energy is transported out of $\mathcal{V}$, across its boundary surface, by the electromagnetic fields. Poynting's theorem says, then, that *the work done on the charges by the electromagnetic force is equal to the decrease in energy remaining in the fields, less the energy that flowed out through the surface*.

The *energy per unit time, per unit area*, transported by the fields is called the **Poynting vector**:

$$\boxed{\mathbf{S} \equiv \frac{1}{\mu_0}(\mathbf{E}\times\mathbf{B}).} \tag{8.10}$$

Specifically, $\mathbf{S}\cdot d\mathbf{a}$ is the energy per unit time crossing the infinitesimal surface $d\mathbf{a}$—the energy *flux* (so $\mathbf{S}$ is the **energy flux density**).[3] We will see many

[3] If you're very fastidious, you'll notice a small gap in the logic here: We know from Eq. 8.9 that $\oint\mathbf{S}\cdot d\mathbf{a}$ is the total power passing through a *closed* surface, but this does not prove that $\int\mathbf{S}\cdot d\mathbf{a}$ is the power passing through any *open* surface (there could be an extra term that integrates to zero over all closed surfaces). This is, however, the obvious and natural interpretation; as always, the precise *location* of energy is not really determined in electrodynamics (see Sect. 2.4.4).

applications of the Poynting vector in Chapters 9 and 11, but for the moment I am mainly interested in using it to express Poynting's theorem more compactly:

$$\frac{dW}{dt} = -\frac{d}{dt}\int_{\mathcal{V}} u\,d\tau - \oint_{\mathcal{S}} \mathbf{S}\cdot d\mathbf{a}. \tag{8.11}$$

What if *no* work is done on the charges in $\mathcal{V}$—what if, for example, we are in a region of empty space, where there *is* no charge? In that case $dW/dt = 0$, so

$$\int \frac{\partial u}{\partial t}\,d\tau = -\oint \mathbf{S}\cdot d\mathbf{a} = -\int (\nabla\cdot\mathbf{S})\,d\tau,$$

and hence

$$\frac{\partial u}{\partial t} = -\nabla\cdot\mathbf{S}. \tag{8.12}$$

This is the "continuity equation" for *energy*—u (energy density) plays the role of ρ (charge density), and $\mathbf{S}$ takes the part of $\mathbf{J}$ (current density). It expresses local conservation of electromagnetic energy.

In *general*, though, electromagnetic energy by itself is *not* conserved (nor is the energy of the charges). Of course not! The fields do work on the charges, and the charges create fields—energy is tossed back and forth between them. In the overall energy economy, you must include the contributions of both the matter and the fields.

Example 8.1. When current flows down a wire, work is done, which shows up as Joule heating of the wire (Eq. 7.7). Though there are certainly *easier* ways to do it, the energy per unit time delivered to the wire can be calculated using the Poynting vector. Assuming it's uniform, the electric field parallel to the wire is

$$E = \frac{V}{L},$$

where V is the potential difference between the ends and L is the length of the wire (Fig. 8.1). The magnetic field is "circumferential"; at the surface (radius a) it has the value

$$B = \frac{\mu_0 I}{2\pi a}.$$

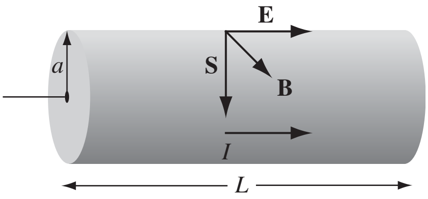


FIGURE 8.1

Accordingly, the magnitude of the Poynting vector is

$$S = \frac{1}{\mu_0} \frac{V}{L} \frac{\mu_0 I}{2\pi a} = \frac{VI}{2\pi a L},$$

and it points radially inward. The energy per unit time passing in through the surface of the wire is therefore

$$\int \mathbf{S} \cdot d\mathbf{a} = S(2\pi a L) = VI,$$

which is exactly what we concluded, on much more direct grounds, in Sect. 7.1.1.[4]

Problem 8.1 Calculate the power (energy per unit time) transported down the cables of Ex. 7.13 and Prob. 7.62, assuming the two conductors are held at potential difference V, and carry current I (down one and back up the other).

Problem 8.2 Consider the charging capacitor in Prob. 7.34.

(a) Find the electric and magnetic fields in the gap, as functions of the distance s from the axis and the time t. (Assume the charge is zero at $t = 0$.)

(b) Find the energy density u_{em} and the Poynting vector $\mathbf{S}$ in the gap. Note especially the *direction* of $\mathbf{S}$. Check that Eq. 8.12 is satisfied.

(c) Determine the total energy in the gap, as a function of time. Calculate the total power flowing into the gap, by integrating the Poynting vector over the appropriate surface. Check that the power input is equal to the rate of increase of energy in the gap (Eq. 8.9—in this case $W = 0$, because there is no charge in the gap). [If you're worried about the fringing fields, do it for a volume of radius $b < a$ well inside the gap.]

8.2 ■ MOMENTUM

8.2.1 ■ Newton's Third Law in Electrodynamics

Imagine a point charge q traveling in along the x axis at a constant speed v. Because it is moving, its electric field is *not* given by Coulomb's law; nevertheless, $\mathbf{E}$ still points radially outward from the instantaneous position of the charge (Fig. 8.2a), as we'll see in Chapter 10. Since, moreover, a moving point charge does not constitute a steady current, its magnetic field is *not* given by the Biot-Savart law. Nevertheless, it's a fact that $\mathbf{B}$ still circles around the axis in a manner suggested by the right-hand rule (Fig. 8.2b); again, the proof will come in Chapter 10.

[4] What about energy flow *down* the wire? For a discussion, see M. K. Harbola, *Am. J. Phys.* **78**, 1203 (2010). For a more sophisticated geometry, see B. S. Davis and L. Kaplan, *Am. J. Phys.* **79**, 1155 (2011).

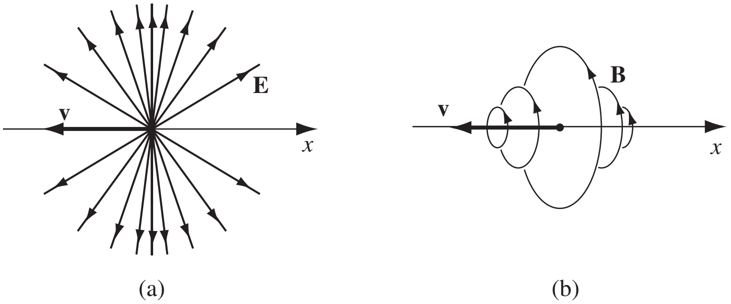


FIGURE 8.2

Now suppose this charge encounters an identical one, proceeding in at the same speed along the y axis. Of course, the electromagnetic force between them would tend to drive them off the axes, but let's assume that they're mounted on tracks, or something, so they're obliged to maintain the same direction and the same speed (Fig. 8.3). The electric force between them is repulsive, but how about the magnetic force? Well, the magnetic field of q_1 points into the page (at the position of q_2), so the magnetic force on q_2 is toward the *right*, whereas the magnetic field of q_2 is *out* of the page (at the position of q_1), and the magnetic force on q_1 is *upward*. *The net electromagnetic force of* q_1 *on* q_2 *is equal but not opposite to the force of* q_2 *on* q_1*, in violation of Newton's third law*. In electro*statics* and magneto*statics* the third law holds, but in electro*dynamics* it does *not*.

Well, that's an interesting curiosity, but then, how often does one actually use the third law, in practice? *Answer:* All the time! For the proof of conservation of momentum rests on the cancellation of internal forces, which follows from the third law. When you tamper with the third law, you are placing conservation of momentum in jeopardy, and there is hardly any principle in physics more sacred than *that*.

Momentum conservation is rescued, in electrodynamics, by the realization that the *fields themselves carry momentum*. This is not so surprising when you

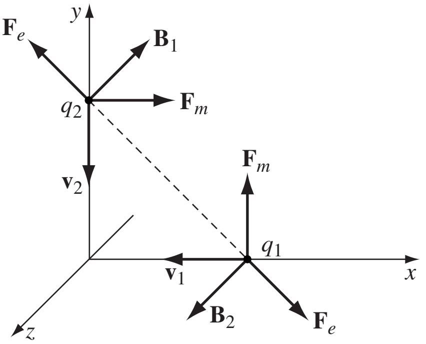


FIGURE 8.3

consider that we have already attributed *energy* to the fields. Whatever momentum is lost to the particles is gained by the fields. Only when the field momentum is added to the mechanical momentum is momentum conservation restored.

8.2.2 ■ Maxwell's Stress Tensor

Let's calculate the total electromagnetic force on the charges in volume $\mathcal{V}$:

$$\mathbf{F} = \int_{\mathcal{V}} (\mathbf{E} + \mathbf{v} \times \mathbf{B})\rho \, d\tau = \int_{\mathcal{V}} (\rho\mathbf{E} + \mathbf{J} \times \mathbf{B}) \, d\tau. \tag{8.13}$$

The *force per unit volume* is

$$\mathbf{f} = \rho\mathbf{E} + \mathbf{J} \times \mathbf{B}. \tag{8.14}$$

As before, I propose to express this in terms of fields alone, eliminating ρ and $\mathbf{J}$ by using Maxwell's equations (i) and (iv):

$$\mathbf{f} = \epsilon_0 (\boldsymbol{\nabla} \cdot \mathbf{E})\mathbf{E} + \left(\frac{1}{\mu_0} \boldsymbol{\nabla} \times \mathbf{B} - \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \times \mathbf{B}.$$

Now

$$\frac{\partial}{\partial t}(\mathbf{E} \times \mathbf{B}) = \left(\frac{\partial \mathbf{E}}{\partial t} \times \mathbf{B} \right) + \left(\mathbf{E} \times \frac{\partial \mathbf{B}}{\partial t} \right),$$

and Faraday's law says

$$\frac{\partial \mathbf{B}}{\partial t} = -\boldsymbol{\nabla} \times \mathbf{E},$$

so

$$\frac{\partial \mathbf{E}}{\partial t} \times \mathbf{B} = \frac{\partial}{\partial t}(\mathbf{E} \times \mathbf{B}) + \mathbf{E} \times (\boldsymbol{\nabla} \times \mathbf{E}).$$

Thus

$$\mathbf{f} = \epsilon_0 \left[(\boldsymbol{\nabla} \cdot \mathbf{E})\mathbf{E} - \mathbf{E} \times (\boldsymbol{\nabla} \times \mathbf{E}) \right] - \frac{1}{\mu_0} \left[\mathbf{B} \times (\boldsymbol{\nabla} \times \mathbf{B}) \right] - \epsilon_0 \frac{\partial}{\partial t}(\mathbf{E} \times \mathbf{B}). \tag{8.15}$$

Just to make things look more symmetrical, let's throw in a term $(\boldsymbol{\nabla} \cdot \mathbf{B})\mathbf{B}$; since $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$, this costs us nothing. Meanwhile, product rule 4 says

$$\boldsymbol{\nabla}(E^2) = 2(\mathbf{E} \cdot \boldsymbol{\nabla})\mathbf{E} + 2\mathbf{E} \times (\boldsymbol{\nabla} \times \mathbf{E}),$$

so

$$\mathbf{E} \times (\boldsymbol{\nabla} \times \mathbf{E}) = \frac{1}{2} \boldsymbol{\nabla}(E^2) - (\mathbf{E} \cdot \boldsymbol{\nabla})\mathbf{E},$$

and the same goes for **B**. Therefore,

$$\mathbf{f} = \epsilon_0\left[(\boldsymbol{\nabla}\cdot\mathbf{E})\mathbf{E} + (\mathbf{E}\cdot\boldsymbol{\nabla})\mathbf{E}\right] + \frac{1}{\mu_0}\left[(\boldsymbol{\nabla}\cdot\mathbf{B})\mathbf{B} + (\mathbf{B}\cdot\boldsymbol{\nabla})\mathbf{B}\right] \\ -\frac{1}{2}\boldsymbol{\nabla}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right) - \epsilon_0\frac{\partial}{\partial t}(\mathbf{E}\times\mathbf{B}). \tag{8.16}$$

Ugly! But it can be simplified by introducing the **Maxwell stress tensor**,

$$T_{ij} \equiv \epsilon_0\left(E_i E_j - \frac{1}{2}\delta_{ij}E^2\right) + \frac{1}{\mu_0}\left(B_i B_j - \frac{1}{2}\delta_{ij}B^2\right). \tag{8.17}$$

The indices i and j refer to the coordinates x, y, and z, so the stress tensor has a total of nine components (T_{xx}, T_{yy}, T_{xz}, T_{yx}, and so on). The **Kronecker delta**, δ_{ij}, is 1 if the indices are the same ($\delta_{xx} = \delta_{yy} = \delta_{zz} = 1$) and zero otherwise ($\delta_{xy} = \delta_{xz} = \delta_{yz} = 0$). Thus

$$T_{xx} = \frac{1}{2}\epsilon_0\left(E_x^2 - E_y^2 - E_z^2\right) + \frac{1}{2\mu_0}\left(B_x^2 - B_y^2 - B_z^2\right),$$

$$T_{xy} = \epsilon_0(E_x E_y) + \frac{1}{\mu_0}(B_x B_y),$$

and so on.

Because it carries *two* indices, where a vector has only one, T_{ij} is sometimes written with a double arrow: $\overleftrightarrow{\mathbf{T}}$. One can form the dot product of $\overleftrightarrow{\mathbf{T}}$ with a vector **a**, in two ways—on the left, and on the right:

$$\left(\mathbf{a}\cdot\overleftrightarrow{\mathbf{T}}\right)_j = \sum_{i=x,y,z} a_i T_{ij}, \quad \left(\overleftrightarrow{\mathbf{T}}\cdot\mathbf{a}\right)_j = \sum_{i=x,y,z} T_{ji} a_i. \tag{8.18}$$

The resulting object, which has one remaining index, is itself a vector. In particular, the divergence of $\overleftrightarrow{\mathbf{T}}$ has as its jth component

$$\left(\boldsymbol{\nabla}\cdot\overleftrightarrow{\mathbf{T}}\right)_j = \epsilon_0\left[(\boldsymbol{\nabla}\cdot\mathbf{E})E_j + (\mathbf{E}\cdot\boldsymbol{\nabla})E_j - \frac{1}{2}\nabla_j E^2\right] \\ + \frac{1}{\mu_0}\left[(\boldsymbol{\nabla}\cdot\mathbf{B})B_j + (\mathbf{B}\cdot\boldsymbol{\nabla})B_j - \frac{1}{2}\nabla_j B^2\right].$$

Thus the force per unit volume (Eq. 8.16) can be written in the much tidier form

$$\mathbf{f} = \boldsymbol{\nabla}\cdot\overleftrightarrow{\mathbf{T}} - \epsilon_0\mu_0\frac{\partial\mathbf{S}}{\partial t}, \tag{8.19}$$

where **S** is the Poynting vector (Eq. 8.10).

The *total* electromagnetic force on the charges in $\mathcal{V}$ (Eq. 8.13) is

$$\mathbf{F} = \oint_{\mathcal{S}} \overleftrightarrow{\mathbf{T}} \cdot d\mathbf{a} - \epsilon_0\mu_0 \frac{d}{dt}\int_{\mathcal{V}} \mathbf{S}\, d\tau. \tag{8.20}$$

(I used the divergence theorem to convert the first term to a surface integral.) In the *static* case the second term drops out, and the electromagnetic force on the charge configuration can be expressed entirely in terms of the stress tensor at the boundary:

$$\mathbf{F} = \oint_{\mathcal{S}} \overleftrightarrow{\mathbf{T}} \cdot d\mathbf{a} \quad \text{(static)}. \tag{8.21}$$

Physically, $\overleftrightarrow{\mathbf{T}}$ is the force per unit area (or **stress**) acting on the surface. More precisely, T_{ij} is the force (per unit area) in the ith direction acting on an element of surface oriented in the jth direction—"diagonal" elements (T_{xx}, T_{yy}, T_{zz}) represent *pressures*, and "off-diagonal" elements (T_{xy}, T_{xz}, etc.) are *shears*.

Example 8.2. Determine the net force on the "northern" hemisphere of a uniformly charged solid sphere of radius R and charge Q (the same as Prob. 2.47, only this time we'll use the Maxwell stress tensor and Eq. 8.21).

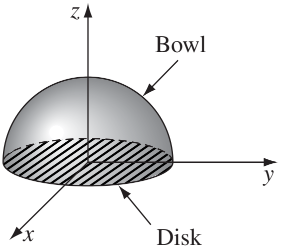


FIGURE 8.4

Solution

The boundary surface consists of two parts—a hemispherical "bowl" at radius R, and a circular disk at $\theta = \pi/2$ (Fig. 8.4). For the bowl,

$$d\mathbf{a} = R^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{r}}$$

and

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0}\frac{Q}{R^2}\,\hat{\mathbf{r}}.$$

In Cartesian components,

$$\hat{\mathbf{r}} = \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}},$$

so

$$T_{zx} = \epsilon_0 E_z E_x = \epsilon_0 \left(\frac{Q}{4\pi\epsilon_0 R^2}\right)^2 \sin\theta\cos\theta\cos\phi,$$

$$T_{zy} = \epsilon_0 E_z E_y = \epsilon_0 \left(\frac{Q}{4\pi\epsilon_0 R^2}\right)^2 \sin\theta\cos\theta\sin\phi,$$

$$T_{zz} = \frac{\epsilon_0}{2}\left(E_z^2 - E_x^2 - E_y^2\right) = \frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0 R^2}\right)^2 (\cos^2\theta - \sin^2\theta). \tag{8.22}$$

The net force is obviously in the z-direction, so it suffices to calculate

$$\left(\overleftrightarrow{\mathbf{T}}\cdot d\mathbf{a}\right)_z = T_{zx}\,da_x + T_{zy}\,da_y + T_{zz}\,da_z = \frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0 R}\right)^2 \sin\theta\cos\theta\,d\theta\,d\phi.$$

The force on the "bowl" is therefore

$$F_{\text{bowl}} = \frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0 R}\right)^2 2\pi\int_0^{\pi/2}\sin\theta\cos\theta\,d\theta = \frac{1}{4\pi\epsilon_0}\frac{Q^2}{8R^2}. \tag{8.23}$$

Meanwhile, for the equatorial disk,

$$d\mathbf{a} = -r\,dr\,d\phi\,\hat{\mathbf{z}}, \tag{8.24}$$

and (since we are now *inside* the sphere)

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0}\frac{Q}{R^3}\mathbf{r} = \frac{1}{4\pi\epsilon_0}\frac{Q}{R^3}r(\cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}).$$

Thus

$$T_{zz} = \frac{\epsilon_0}{2}\left(E_z^2 - E_x^2 - E_y^2\right) = -\frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0 R^3}\right)^2 r^2,$$

and hence

$$\left(\overleftrightarrow{\mathbf{T}}\cdot d\mathbf{a}\right)_z = \frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0 R^3}\right)^2 r^3\,dr\,d\phi.$$

The force on the disk is therefore

$$F_{\text{disk}} = \frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0 R^3}\right)^2 2\pi\int_0^R r^3 dr = \frac{1}{4\pi\epsilon_0}\frac{Q^2}{16R^2}. \tag{8.25}$$

Combining Eqs. 8.23 and 8.25, I conclude that the net force on the northern hemisphere is

$$F = \frac{1}{4\pi\epsilon_0}\frac{3Q^2}{16R^2}. \tag{8.26}$$

Incidentally, in applying Eq. 8.21, *any* volume that encloses all of the charge in question (and no *other* charge) will do the job. For example, in the present case we could use the whole region $z > 0$. In that case the boundary surface consists of the entire xy plane (plus a hemisphere at $r = \infty$, but $E = 0$ out there, so it contributes nothing). In place of the "bowl," we now have the outer portion of the plane ($r > R$). Here

$$T_{zz} = -\frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0}\right)^2 \frac{1}{r^4}$$

(Eq. 8.22 with $\theta = \pi/2$ and $R \to r$), and $d\mathbf{a}$ is given by Eq. 8.24, so

$$\left(\overleftrightarrow{\mathbf{T}} \cdot d\mathbf{a}\right)_z = \frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0}\right)^2 \frac{1}{r^3}\, dr\, d\phi,$$

and the contribution from the plane for $r > R$ is

$$\frac{\epsilon_0}{2}\left(\frac{Q}{4\pi\epsilon_0}\right)^2 2\pi \int_R^\infty \frac{1}{r^3}\, dr = \frac{1}{4\pi\epsilon_0}\frac{Q^2}{8R^2},$$

the same as for the bowl (Eq. 8.23).

I hope you didn't get too bogged down in the details of Ex. 8.2. If so, take a moment to appreciate what happened. We were calculating the force on a solid object, but instead of doing a *volume* integral, as you might expect, Eq. 8.21 allowed us to set it up as a *surface* integral; somehow the stress tensor sniffs out what is going on inside.

! **Problem 8.3** Calculate the force of magnetic attraction between the northern and southern hemispheres of a uniformly charged spinning spherical shell, with radius R, angular velocity ω, and surface charge density σ. [This is the same as Prob. 5.44, but this time use the Maxwell stress tensor and Eq. 8.21.]

Problem 8.4

(a) Consider two equal point charges q, separated by a distance $2a$. Construct the plane equidistant from the two charges. By integrating Maxwell's stress tensor over this plane, determine the force of one charge on the other.

(b) Do the same for charges that are opposite in sign.

8.2.3 ■ Conservation of Momentum

According to Newton's second law, the force on an object is equal to the rate of change of its momentum:

$$\mathbf{F} = \frac{d\mathbf{p}_{\text{mech}}}{dt}.$$

Equation 8.20 can therefore be written in the form[5]

$$\frac{d\mathbf{p}_{\text{mech}}}{dt} = -\epsilon_0\mu_0\frac{d}{dt}\int_{\mathcal{V}}\mathbf{S}\,d\tau + \oint_{\mathcal{S}}\overleftrightarrow{\mathbf{T}}\cdot d\mathbf{a}, \tag{8.27}$$

where $\mathbf{p}_{\text{mech}}$ is the (mechanical) momentum of the particles in volume $\mathcal{V}$. This expression is similar in structure to Poynting's theorem (Eq. 8.11), and it invites an analogous interpretation: The first integral represents *momentum stored in the fields*:

$$\mathbf{p} = \mu_0\epsilon_0\int_{\mathcal{V}}\mathbf{S}\,d\tau, \tag{8.28}$$

while the second integral is the *momentum per unit time flowing in through the surface*.

Equation 8.27 is the statement of *conservation of momentum* in electrodynamics: If the mechanical momentum increases, either the field momentum *de*creases, or else the fields are carrying momentum into the volume through the surface. The momentum *density* in the fields is evidently

$$\boxed{\mathbf{g} = \mu_0\epsilon_0\mathbf{S} = \epsilon_0(\mathbf{E}\times\mathbf{B}),} \tag{8.29}$$

and the momentum flux transported by the fields is $-\overleftrightarrow{\mathbf{T}}$ (specifically, $-\overleftrightarrow{\mathbf{T}}\cdot d\mathbf{a}$ is the electromagnetic momentum per unit time passing through the area $d\mathbf{a}$).

If the mechanical momentum in $\mathcal{V}$ is not changing (for example, if we are talking about a region of empty space), then

$$\int\frac{\partial\mathbf{g}}{\partial t}\,d\tau = \oint\overleftrightarrow{\mathbf{T}}\cdot d\mathbf{a} = \int\nabla\cdot\overleftrightarrow{\mathbf{T}}\,d\tau,$$

and hence

$$\frac{\partial\mathbf{g}}{\partial t} = \nabla\cdot\overleftrightarrow{\mathbf{T}}. \tag{8.30}$$

This is the "continuity equation" for electromagnetic momentum, with $\mathbf{g}$ (momentum density) in the role of ρ (charge density) and $-\overleftrightarrow{\mathbf{T}}$ playing the part of $\mathbf{J}$; it expresses the local conservation of field momentum. But in general (when there *are* charges around) the field momentum by itself, and the mechanical momentum by itself, are *not* conserved—charges and fields exchange momentum, and only the *total* is conserved.

Notice that the Poynting vector has appeared in two quite different roles: $\mathbf{S}$ itself is the energy per unit area, per unit time, transported by the electromagnetic fields, while $\mu_0\epsilon_0\mathbf{S}$ is the momentum per unit volume stored in those fields.[6]

[5] Let's assume the only forces acting are electromagnetic. You can include other forces if you like—both here and in the discussion of energy conservation—but they are just a distraction from the essential story.

[6] This is no coincidence—see R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics* (Reading, Mass.: Addison-Wesley, 1964), Vol. II, Section 27-6.

Similarly, $\overleftrightarrow{\mathbf{T}}$ plays a dual role: $\overleftrightarrow{\mathbf{T}}$ itself is the electromagnetic stress (force per unit area) acting on a surface, and $-\overleftrightarrow{\mathbf{T}}$ describes the flow of momentum (it is the momentum current density) carried by the fields.

Example 8.3. A long coaxial cable, of length l, consists of an inner conductor (radius a) and an outer conductor (radius b). It is connected to a battery at one end and a resistor at the other (Fig. 8.5). The inner conductor carries a uniform charge per unit length λ, and a steady current I to the right; the outer conductor has the opposite charge and current. What is the electromagnetic momentum stored in the fields?

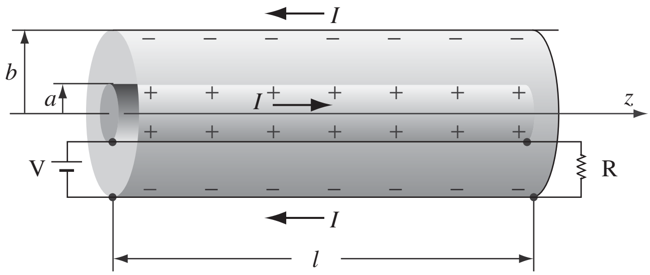


FIGURE 8.5

Solution
The fields are

$$\mathbf{E} = \frac{1}{2\pi\epsilon_0}\frac{\lambda}{s}\,\hat{\mathbf{s}}, \qquad \mathbf{B} = \frac{\mu_0}{2\pi}\frac{I}{s}\,\hat{\boldsymbol{\phi}}.$$

The Poynting vector is therefore

$$\mathbf{S} = \frac{\lambda I}{4\pi^2\epsilon_0 s^2}\,\hat{\mathbf{z}}.$$

So energy is flowing down the line, from the battery to the resistor. In fact, the power transported is

$$P = \int \mathbf{S}\cdot d\mathbf{a} = \frac{\lambda I}{4\pi^2\epsilon_0}\int_a^b \frac{1}{s^2}2\pi s\,ds = \frac{\lambda I}{2\pi\epsilon_0}\ln(b/a) = IV,$$

as it should be.

The *momentum* in the fields is

$$\mathbf{p} = \mu_0\epsilon_0\int \mathbf{S}\,d\tau = \frac{\mu_0\lambda I}{4\pi^2}\,\hat{\mathbf{z}}\int_a^b \frac{1}{s^2}l2\pi s\,ds = \frac{\mu_0\lambda I l}{2\pi}\ln(b/a)\,\hat{\mathbf{z}} = \frac{IVl}{c^2}\,\hat{\mathbf{z}}.$$

This is an astonishing result. The cable is not moving, **E** and **B** are static, and yet we are asked to believe that there is momentum in the fields. If something tells you this cannot be the whole story, you have sound intuitions. But the resolution of this paradox will have to await Chapter 12 (Ex. 12.12).

Suppose now that we turn up the resistance, so the current decreases. The changing magnetic field will induce an electric field (Eq. 7.20):

$$\mathbf{E} = \left[\frac{\mu_0}{2\pi}\frac{dI}{dt}\ln s + K\right]\hat{\mathbf{z}}.$$

This field exerts a force on ±λ:

$$\mathbf{F} = \lambda l\left[\frac{\mu_0}{2\pi}\frac{dI}{dt}\ln a + K\right]\hat{\mathbf{z}} - \lambda l\left[\frac{\mu_0}{2\pi}\frac{dI}{dt}\ln b + K\right]\hat{\mathbf{z}} = -\frac{\mu_0\lambda l}{2\pi}\frac{dI}{dt}\ln(b/a)\,\hat{\mathbf{z}}.$$

The total momentum imparted to the cable, as the current drops from I to 0, is therefore

$$\mathbf{p}_{\text{mech}} = \int \mathbf{F}\,dt = \frac{\mu_0\lambda I l}{2\pi}\ln(b/a)\,\hat{\mathbf{z}},$$

which is precisely the momentum originally stored in the fields.

Problem 8.5 Imagine two parallel infinite sheets, carrying uniform surface charge $+\sigma$ (on the sheet at $z = d$) and $-\sigma$ (at $z = 0$). They are moving in the y direction at constant speed v (as in Problem 5.17).

(a) What is the electromagnetic momentum in a region of area A?

(b) Now suppose the top sheet moves slowly down (speed u) until it reaches the bottom sheet, so the fields disappear. By calculating the total force on the charge ($q = \sigma A$), show that the impulse delivered to the sheet is equal to the momentum originally stored in the fields. [*Hint:* As the upper plate passes by, the magnetic field drops to zero, inducing an electric field that delivers an impulse to the lower plate.]

Problem 8.6 A charged parallel-plate capacitor (with uniform electric field $\mathbf{E} = E\,\hat{\mathbf{z}}$) is placed in a uniform magnetic field $\mathbf{B} = B\,\hat{\mathbf{x}}$, as shown in Fig. 8.6.

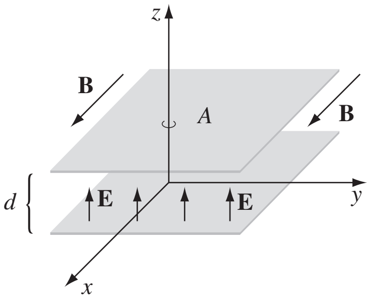


FIGURE 8.6

(a) Find the electromagnetic momentum in the space between the plates.

(b) Now a resistive wire is connected between the plates, along the z axis, so that the capacitor slowly discharges. The current through the wire will experience a magnetic force; what is the total impulse delivered to the system, during the discharge?[7]

Problem 8.7 Consider an infinite parallel-plate capacitor, with the lower plate (at $z = -d/2$) carrying surface charge density $-\sigma$, and the upper plate (at $z = +d/2$) carrying charge density $+\sigma$.

(a) Determine all nine elements of the stress tensor, in the region between the plates. Display your answer as a 3 × 3 matrix:

$$\begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix}$$

(b) Use Eq. 8.21 to determine the electromagnetic force per unit area on the top plate. Compare Eq. 2.51.

(c) What is the electromagnetic momentum per unit area, per unit time, crossing the xy plane (or any other plane parallel to that one, between the plates)?

(d) Of course, there must be *mechanical* forces holding the plates apart—perhaps the capacitor is filled with insulating material under pressure. Suppose we suddenly *remove* the insulator; the momentum flux (c) is now absorbed by the plates, and they begin to move. Find the momentum per unit time delivered to the top plate (which is to say, the force acting on it) and compare your answer to (b). [*Note:* This is not an *additional* force, but rather an alternative way of calculating the *same* force—in (b) we got it from the force law, and in (d) we do it by conservation of momentum.]

8.2.4 ■ Angular Momentum

By now, the electromagnetic fields (which started out as mediators of forces between charges) have taken on a life of their own. They carry *energy* (Eq. 8.5)

$$u = \frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0} B^2\right), \tag{8.31}$$

and *momentum* (Eq. 8.29)

$$\mathbf{g} = \epsilon_0(\mathbf{E} \times \mathbf{B}), \tag{8.32}$$

[7] There is *much* more to be said about this problem, so don't get too excited if your answers to (a) and (b) appear to be consistent. See D. Babson, et al., *Am. J. Phys.* **77**, 826 (2009).

and, for that matter, *angular* momentum:

$$\boldsymbol{\ell} = \mathbf{r} \times \mathbf{g} = \epsilon_0 \left[\mathbf{r} \times (\mathbf{E} \times \mathbf{B})\right]. \tag{8.33}$$

Even perfectly *static* fields can harbor momentum and angular momentum, as long as $\mathbf{E} \times \mathbf{B}$ is nonzero, and it is only when these field contributions are included that the conservation laws are sustained.

Example 8.4. Imagine a very long solenoid with radius R, n turns per unit length, and current I. Coaxial with the solenoid are two long cylindrical (nonconducting) shells of length l—one, *inside* the solenoid at radius a, carries a charge $+Q$, uniformly distributed over its surface; the other, *outside* the solenoid at radius b, carries charge $-Q$ (see Fig. 8.7; l is supposed to be much greater than b). When the current in the solenoid is gradually reduced, the cylinders begin to rotate, as we found in Ex. 7.8. *Question*: Where does the angular momentum come from?[8]

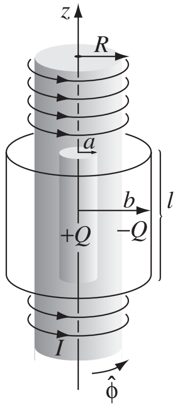


FIGURE 8.7

Solution

It was initially stored in the fields. Before the current was switched off, there was an electric field,

$$\mathbf{E} = \frac{Q}{2\pi\epsilon_0 l}\frac{1}{s}\,\hat{\mathbf{s}} \quad (a < s < b),$$

in the region between the cylinders, and a magnetic field,

$$\mathbf{B} = \mu_0 n I\,\hat{\mathbf{z}} \quad (s < R),$$

[8] This is a variation on the "Feynman disk paradox" (R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures*, vol 2, pp. 17-5 (Reading, Mass.: Addison-Wesley, 1964) suggested by F. L. Boos, Jr. (*Am. J. Phys.* **52**, 756 (1984)). A similar model was proposed earlier by R. H. Romer (*Am. J. Phys.* **34**, 772 (1966)). For further references, see T.-C. E. Ma, *Am. J. Phys.* **54**, 949 (1986).

inside the solenoid. The momentum density (Eq. 8.29) was therefore

$$\mathbf{g} = -\frac{\mu_0 n I Q}{2\pi l s}\,\hat{\boldsymbol{\phi}},$$

in the region $a < s < R$. The z component of the *angular* momentum density was

$$(\mathbf{r} \times \mathbf{g})_z = -\frac{\mu_0 n I Q}{2\pi l},$$

which is *constant* (independent of s). To get the *total* angular momentum in the fields, we simply multiply by the volume, $\pi(R^2 - a^2)l$:[9]

$$\mathbf{L} = -\frac{1}{2}\mu_0 n I Q(R^2 - a^2)\,\hat{\mathbf{z}}. \tag{8.34}$$

When the current is turned off, the changing magnetic field induces a circumferential electric field, given by Faraday's law:

$$\mathbf{E} = \begin{cases} -\dfrac{1}{2}\mu_0 n \dfrac{dI}{dt}\dfrac{R^2}{s}\,\hat{\boldsymbol{\phi}}, & (s > R), \\[2ex] -\dfrac{1}{2}\mu_0 n \dfrac{dI}{dt} s\,\hat{\boldsymbol{\phi}}, & (s < R). \end{cases}$$

Thus the torque on the outer cylinder is

$$\mathbf{N}_b = \mathbf{r} \times (-Q\mathbf{E}) = \frac{1}{2}\mu_0 n Q R^2 \frac{dI}{dt}\hat{\mathbf{z}},$$

and it picks up an angular momentum

$$\mathbf{L}_b = \frac{1}{2}\mu_0 n Q R^2\,\hat{\mathbf{z}} \int_I^0 \frac{dI}{dt}dt = -\frac{1}{2}\mu_0 n I Q R^2\,\hat{\mathbf{z}}.$$

Similarly, the torque on the inner cylinder is

$$\mathbf{N}_a = -\frac{1}{2}\mu_0 n Q a^2 \frac{dI}{dt}\,\hat{\mathbf{z}},$$

and its angular momentum increase is

$$\mathbf{L}_a = \frac{1}{2}\mu_0 n I Q a^2\,\hat{\mathbf{z}}.$$

So it all works out: $\mathbf{L}_{\text{em}} = \mathbf{L}_a + \mathbf{L}_b$. The angular momentum *lost* by the fields is precisely equal to the angular momentum *gained* by the cylinders, and the *total* angular momentum (fields plus matter) is conserved.

[9] The radial component integrates to zero, by symmetry.

Problem 8.8 In Ex. 8.4, suppose that instead of turning off the *magnetic* field (by reducing I) we turn off the *electric* field, by connecting a weakly[10] conducting radial spoke between the cylinders. (We'll have to cut a slot in the solenoid, so the cylinders can still rotate freely.) From the magnetic force on the current in the spoke, determine the total angular momentum delivered to the cylinders, as they discharge (they are now rigidly connected, so they rotate together). Compare the initial angular momentum stored in the fields (Eq. 8.34). (Notice that the *mechanism* by which angular momentum is transferred from the fields to the cylinders is entirely different in the two cases: in Ex. 8.4 it was Faraday's law, but here it is the Lorentz force law.)

Problem 8.9 Two concentric spherical shells carry uniformly distributed charges $+Q$ (at radius a) and $-Q$ (at radius $b > a$). They are immersed in a uniform magnetic field $\mathbf{B} = B_0\,\hat{\mathbf{z}}$.

(a) Find the angular momentum of the fields (with respect to the center).

(b) Now the magnetic field is gradually turned off. Find the torque on each sphere, and the resulting angular momentum of the system.

! **Problem 8.10**[11] Imagine an iron sphere of radius R that carries a charge Q and a uniform magnetization $\mathbf{M} = M\hat{\mathbf{z}}$. The sphere is initially at rest.

(a) Compute the angular momentum stored in the electromagnetic fields.

(b) Suppose the sphere is gradually (and uniformly) demagnetized (perhaps by heating it up past the Curie point). Use Faraday's law to determine the induced electric field, find the torque this field exerts on the sphere, and calculate the total angular momentum imparted to the sphere in the course of the demagnetization.

(c) Suppose instead of *demagnetizing* the sphere we *discharge* it, by connecting a grounding wire to the north pole. Assume the current flows over the surface in such a way that the charge density remains uniform. Use the Lorentz force law to determine the torque on the sphere, and calculate the total angular momentum imparted to the sphere in the course of the discharge. (The magnetic field is discontinuous at the surface ... does this matter?) [*Answer:* $\frac{2}{9}\mu_0 M Q R^2$]

8.3 ■ MAGNETIC FORCES DO NO WORK[12]

This is perhaps a good place to revisit the old paradox that magnetic forces do no work (Eq. 5.11). What about that magnetic crane lifting the carcass of a junked car? *Somebody* is doing work on the car, and if it's not the magnetic field, who

[10] In Ex. 8.4 we turned the current off slowly, to keep things quasistatic; here we reduce the electric field slowly to keep the displacement current negligible.

[11] This version of the Feynman disk paradox was proposed by N. L. Sharma (*Am. J. Phys.* **56**, 420 (1988)); similar models were analyzed by E. M. Pugh and G. E. Pugh, *Am. J. Phys.* **35**, 153 (1967) and by R. H. Romer, *Am. J. Phys.* **35**, 445 (1967).

[12] This section can be skipped without loss of continuity. I include it for those readers who are disturbed by the notion that magnetic forces do no work.

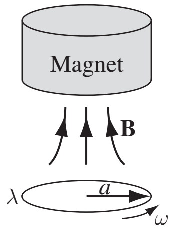


FIGURE 8.8

is it? The car is ferromagnetic; in the presence of the magnetic field, it contains a lot of microscopic magnetic dipoles (spinning electrons, actually), all lined up. The resulting magnetization is equivalent to a bound current running around the surface, so let's model the car as a circular current loop—in fact, let's make it an insulating ring of line charge λ rotating at angular velocity ω (Fig. 8.8).

The upward magnetic force on the loop is (Eq. 6.2)

$$F = 2\pi I a B_s, \tag{8.35}$$

where B_s is the radial component of the magnet's field,[13] and $I = \lambda\omega a$. If the ring rises a distance dz (while the magnet itself stays put), the work done on it is

$$dW = 2\pi a^2 \lambda\omega B_s \, dz. \tag{8.36}$$

This increases the potential energy of the ring. Who did the work? Naively, it appears that the magnetic field is responsible, but we have already learned (Ex. 5.3) that this is not the case—as the ring rises, the magnetic force is perpendicular to the *net* velocity of the charges in the ring, so it does *no* work on them.

At the same time, however, a motional emf is induced in the ring, which opposes the flow of charge, and hence reduces its angular velocity:

$$\mathcal{E} = -\frac{d\Phi}{dt}.$$

Here $d\Phi$ is the flux through the "ribbon" joining the ring at time t to the ring at time $t + dt$ (Fig. 8.9):

$$d\Phi = B_s \, 2\pi a \, dz.$$

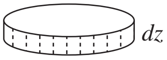


FIGURE 8.9

[13] Note that the field has to be *nonuniform*, or it won't lift the car at all.

Now

$$\mathcal{E} = \oint \mathbf{f} \cdot d\mathbf{l} = f(2\pi a),$$

where **f** is the force per unit charge. So

$$f = -B_s \frac{dz}{dt}, \tag{8.37}$$

the force on a segment of length dl is $f\lambda\, dl$, the torque on the ring is

$$N = a\left(-B_s \frac{dz}{dt}\right)\lambda(2\pi a),$$

and the work done (slowing the rotation) is $N\, d\phi = N\omega\, dt$, or

$$dW = -2\pi a^2 \lambda\omega B_s\, dz. \tag{8.38}$$

The ring slows down, and the rotational energy it loses (Eq. 8.38) is precisely equal to the potential energy it gains (Eq. 8.36). All the magnetic field did was convert energy from one form to another. If you'll permit some sloppy language, the work done by the vertical component of the magnetic force (Eq. 8.35) is equal and opposite to the work done by its horizontal component (Eq. 8.37).[14]

What about the magnet? Is it completely passive in this process? Suppose we model it as a big circular loop (radius b), resting on a table and carrying a current I_b; the "junk car" is a relatively small current loop (radius a), on the floor directly below, carrying a current I_a (Fig. 8.10). This time, just for a change, let's assume both currents are constant (we'll include a regulated power supply in each loop[15]). Parallel currents attract, and we propose to lift the small loop off the floor, keeping careful track of the work done and the agency responsible.

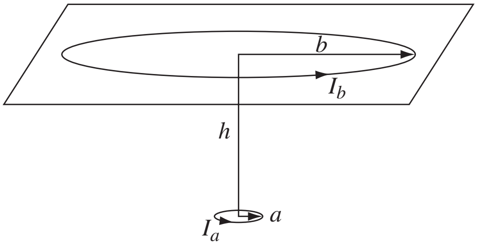


FIGURE 8.10

[14] This argument is essentially the same as the one in Ex. 5.3, except that in this case I told the story in terms of motional emf, instead of the Lorentz force law. But after all, the flux rule is a *consequence* of the Lorentz force law.

[15] The lower loop could be a single spinning electron, in which case quantum mechanics fixes its angular momentum at $\hbar/2$. It might appear that this sustains the current, with no need for a power supply. I will return to this point, but for now let's just keep quantum mechanics out of it.

Let's start by adjusting the currents so the small ring just "floats," a distance h below the table, with the magnetic force exactly balancing the weight ($m_a g$) of the little ring. I'll let you calculate the magnetic force (Prob. 8.11):

$$F_{\text{mag}} = \frac{3\pi}{2}\mu_0 I_a I_b \frac{a^2 b^2 h}{(b^2+h^2)^{5/2}} = m_a g. \tag{8.39}$$

Now the loop rises an infinitesimal distance dz; the work done is equal to the gain in its potential energy

$$dW_g = m_a g\,dz = \frac{3\pi}{2}\mu_0 I_a I_b \frac{a^2 b^2 h}{(b^2+h^2)^{5/2}}\,dz. \tag{8.40}$$

Who did it? The magnetic field? *No!* The work was done by the power supply that sustains the current in loop a (Ex. 5.3). As the loop rises, a motional emf is induced in it. The flux through the loop is

$$\Phi_a = M I_b,$$

where M is the mutual inductance of the two loops:

$$M = \frac{\pi\mu_0}{2}\frac{a^2 b^2}{(b^2+h^2)^{3/2}}$$

(Prob. 7.22). The emf is

$$\begin{aligned}\mathcal{E}_a &= -\frac{d\Phi_a}{dt} = -I_b\frac{dM}{dt} = -I_b\frac{dM}{dh}\frac{dh}{dt}\\ &= -I_b\left(-\frac{3}{2}\right)\frac{\pi\mu_0}{2}\frac{a^2 b^2}{(b^2+h^2)^{5/2}}2h\frac{(-dz)}{dt}.\end{aligned}$$

The work done by the power supply (fighting against this motional emf) is

$$dW_a = -\mathcal{E}_a I_a\,dt = \frac{3\pi}{2}\mu_0 I_a I_b \frac{a^2 b^2 h}{(b^2+h^2)^{5/2}}\,dz \tag{8.41}$$

—same as the work done in lifting the loop (Eq. 8.40).

Meanwhile, however, a *Faraday* emf is induced in the *upper* loop, due to the changing flux from the lower loop:

$$\Phi_b = M I_a \;\Rightarrow\; \mathcal{E}_b = -I_a\frac{dM}{dt},$$

and the work done by the power supply in ring b (to sustain the current I_b) is

$$dW_b = -\mathcal{E}_b I_b\,dt = \frac{3\pi}{2}\mu_0 I_a I_b \frac{a^2 b^2 h}{(b^2+h^2)^{5/2}}\,dz, \tag{8.42}$$

exactly the same as dW_a. That's embarrassing—the power supplies have done *twice* as much work as was necessary to lift the junk car! Where did the "wasted" energy go? *Answer:* It increased the energy stored in the fields. The energy in a system of two current-carrying loops is (see Prob. 8.12)

$$U = \frac{1}{2}L_a I_a^2 + \frac{1}{2}L_b I_b^2 + M I_a I_b, \tag{8.43}$$

so

$$dU = I_a I_b \frac{dM}{dt} dt = dW_b.$$

Remarkably, all four energy increments are the same. If we care to apportion things this way, the power supply in loop a contributes the energy necessary to lift the lower ring, while the power supply in loop b provides the extra energy for the fields. If all we're interested in is the work done to raise the ring, we can ignore the upper loop (and the energy in the fields) altogether.

In both these models, the magnet itself was stationary. That's like lifting a paper clip by holding a magnet over it. But in the case of the magnetic crane, the car stays in contact with the magnet, which is attached to a cable that lifts the whole works. As a model, we might stick the upper loop in a big box, the lower loop in a little box, and crank up the currents so the force of attraction is much greater than $m_a g$; the two boxes snap together, and we attach a string to the upper box and pull up on it (Fig. 8.11).

The same old mechanism (Ex. 5.3) prevails: as the lower loop rises, the magnetic force tilts backwards; its vertical component lifts the loop, but its horizontal component opposes the current, and no net work is done. This time, however, the motional emf is perfectly balanced by the Faraday emf fighting to keep the current going—the flux through the lower loop is not changing. (If you like, the flux is *increasing* because the loop is moving upward, into a region of higher magnetic field, but it is *decreasing* because the magnetic field of the upper loop—at any give point in space—is decreasing as that loop moves up.) No power supply is needed to sustain the current (and for that matter, no power supply is required in the upper loop either, since the energy in the fields is not changing. Who did the work to lift the car? The person pulling up on the rope, obviously. The role of the magnetic field was merely to transmit this energy to the car, via the vertical component of the magnetic force. But the magnetic field itself (as always) did no work.

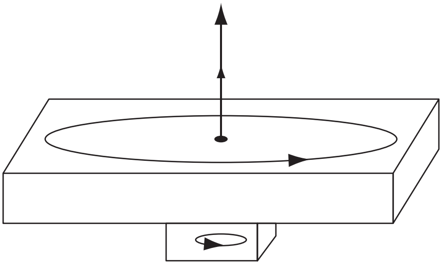


FIGURE 8.11

The fact that magnetic fields do no work follows directly from the Lorentz force law, so if you think you have discovered an exception, you're going to have to explain why that law is incorrect. For example, if magnetic monopoles exist, the force on a particle with electric charge q_e and magnetic charge q_m becomes (Prob. 7.38):

$$\mathbf{F} = q_e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + q_m\left(\mathbf{B} - \epsilon_0\mu_0\mathbf{v} \times \mathbf{E}\right). \tag{8.44}$$

In that case, magnetic fields *can* do work . . . but *only on magnetic charges*. So unless your car is made of monopoles (I don't think so), this doesn't solve the problem.

A somewhat less radical possibility is that in addition to electric charges there exist permanent point magnetic dipoles (electrons?), whose dipole moment **m** is not associated with any electric current, but simply *is*. The Lorentz force law acquires an extra term

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + \nabla(\mathbf{m} \cdot \mathbf{B}).$$

The magnetic field *can* do work on these "intrinsic" dipoles (which experience no motional or Faraday emf, since they enclose no flux). I don't know whether a consistent theory can be constructed in this way, but in any event it is *not* classical electrodynamics, which is predicated on Ampère's assumption that all magnetic phenomena are due to electric charges in motion, and point magnetic dipoles must be interpreted as the limits of tiny current loops.

Problem 8.11 Derive Eq. 8.39. [*Hint:* Treat the lower loop as a magnetic dipole.]

Problem 8.12 Derive Eq. 8.43. [*Hint:* Use the method of Section 7.2.4, building the two currents up from zero to their final values.]

More Problems on Chapter 8

Problem 8.13[16] A very long solenoid of radius a, with n turns per unit length, carries a current I_s. Coaxial with the solenoid, at radius $b \gg a$, is a circular ring of wire, with resistance R. When the current in the solenoid is (gradually) decreased, a current I_r is induced in the ring.

(a) Calculate I_r, in terms of dI_s/dt.

(b) The power ($I_r^2 R$) delivered to the ring must have come from the solenoid. Confirm this by calculating the Poynting vector just outside the solenoid (the *electric* field is due to the changing flux in the solenoid; the *magnetic* field is due to the current in the ring). Integrate over the entire surface of the solenoid, and check that you recover the correct total power.

[16] For extensive discussion, see M. A. Heald, *Am. J. Phys.* **56**, 540 (1988).

Problem 8.14 An infinitely long cylindrical tube, of radius a, moves at constant speed v along its axis. It carries a net charge per unit length λ, uniformly distributed over its surface. Surrounding it, at radius b, is another cylinder, moving with the same velocity but carrying the opposite charge ($-\lambda$). Find:

(a) The energy per unit length stored in the fields.

(b) The momentum per unit length in the fields.

(c) The energy per unit time transported by the fields across a plane perpendicular to the cylinders.

Problem 8.15 A point charge q is located at the center of a toroidal coil of rectangular cross section, inner radius a, outer radius $a + w$, and height h, which carries a total of N tightly-wound turns and current I.

(a) Find the electromagnetic momentum $\mathbf{p}$ of this configuration, assuming that w and h are both much less than a (so you can ignore the variation of the fields over the cross section).

(b) Now the current in the toroid is turned off, quickly enough that the point charge does not move appreciably as the magnetic field drops to zero. Show that the impulse imparted to q is equal to the momentum originally stored in the electromagnetic fields. [*Hint:* You might want to refer to Prob. 7.19.]

Problem 8.16[17] A sphere of radius R carries a uniform polarization $\mathbf{P}$ and a uniform magnetization $\mathbf{M}$ (not necessarily in the same direction). Find the electromagnetic momentum of this configuration. [*Answer:* $(4/9)\pi\mu_0 R^3(\mathbf{M} \times \mathbf{P})$]

Problem 8.17[18] Picture the electron as a uniformly charged spherical shell, with charge e and radius R, spinning at angular velocity ω.

(a) Calculate the total energy contained in the electromagnetic fields.

(b) Calculate the total angular momentum contained in the fields.

(c) According to the Einstein formula ($E = mc^2$), the energy in the fields should contribute to the mass of the electron. Lorentz and others speculated that the *entire* mass of the electron might be accounted for in this way: $U_{\text{em}} = m_e c^2$. Suppose, moreover, that the electron's spin angular momentum is entirely attributable to the electromagnetic fields: $L_{\text{em}} = \hbar/2$. On these two assumptions, determine the radius and angular velocity of the electron. What is their product, ωR? Does this classical model make sense?

Problem 8.18 Work out the formulas for u, $\mathbf{S}$, $\mathbf{g}$, and $\overleftrightarrow{\mathbf{T}}$ in the presence of magnetic charge. [*Hint:* Start with the generalized Maxwell equations (7.44) and Lorentz force law (Eq. 8.44), and follow the derivations in Sections 8.1.2, 8.2.2, and 8.2.3.]

[17] For an interesting discussion and references, see R. H. Romer, *Am. J. Phys.* **63**, 777 (1995).

[18] See J. Higbie, *Am. J. Phys.* **56**, 378 (1988).

! **Problem 8.19**[19] Suppose you had an electric charge q_e and a magnetic monopole q_m. The field of the electric charge is

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0}\frac{q_e}{\scriptstyle{r}^2}\hat{\boldsymbol{\scriptstyle{r}}}$$

(of course), and the field of the magnetic monopole is

$$\mathbf{B} = \frac{\mu_0}{4\pi}\frac{q_m}{\scriptstyle{r}^2}\hat{\boldsymbol{\scriptstyle{r}}}.$$

Find the total angular momentum stored in the fields, if the two charges are separated by a distance d. [*Answer:* $(\mu_0/4\pi)q_e q_m$.][20]

Problem 8.20 Consider an ideal stationary magnetic dipole $\mathbf{m}$ in a static electric field $\mathbf{E}$. Show that the fields carry momentum

$$\mathbf{p} = -\epsilon_0\mu_0(\mathbf{m}\times\mathbf{E}). \tag{8.45}$$

[*Hint:* There are several ways to do this. The simplest method is to start with $\mathbf{p} = \epsilon_0\int(\mathbf{E}\times\mathbf{B})\,d\tau$, write $\mathbf{E} = -\boldsymbol{\nabla}V$, and use integration by parts to show that

$$\mathbf{p} = \epsilon_0\mu_0\int V\mathbf{J}\,d\tau.$$

So far, this is valid for *any* localized static configuration. For a current confined to an infinitesimal neighborhood of the origin we can approximate $V(\mathbf{r}) \approx V(\mathbf{0}) - \mathbf{E}(\mathbf{0})\cdot\mathbf{r}$. Treat the dipole as a current loop, and use Eqs. 5.82 and 1.108.][21]

Problem 8.21 Because the cylinders in Ex. 8.4 are left rotating (at angular velocities ω_a and ω_b, say), there is actually a residual magnetic field, and hence angular momentum in the fields, even after the current in the solenoid has been extinguished. If the cylinders are heavy, this correction will be negligible, but it is interesting to do the problem *without* making that assumption.[22]

(a) Calculate (in terms of ω_a and ω_b) the final angular momentum in the fields. [Define $\boldsymbol{\omega} = \omega\,\hat{\mathbf{z}}$, so ω_a and ω_b could be positive or negative.]

(b) As the cylinders begin to rotate, their changing magnetic field induces an extra azimuthal electric field, which, in turn, will make an additional contribution to

[19] This system is known as **Thomson's dipole**. See I. Adawi, *Am. J. Phys.* **44**, 762 (1976) and *Phys. Rev.* **D31**, 3301 (1985), and K. R. Brownstein, *Am. J. Phys.* **57**, 420 (1989), for discussion and references.

[20] Note that this result is *independent of the separation distance* d! It points from q_e toward q_m. In quantum mechanics, angular momentum comes in half-integer multiples of $\hbar$, so this result suggests that if magnetic monopoles exist, electric and magnetic charge must be quantized: $\mu_0 q_e q_m/4\pi = n\hbar/2$, for $n = 1, 2, 3, \ldots$, an idea first proposed by Dirac in 1931. If even *one* monopole is lurking somewhere in the universe, this would "explain" why electric charge comes in discrete units. (However, see D. Singleton, *Am. J. Phys.* **66**, 697 (1998) for a cautionary note.)

[21] As it stands, Eq. 8.45 is valid only for *ideal* dipoles. But $\mathbf{g}$ is linear in $\mathbf{B}$, and therefore, if $\mathbf{E}$ is held fixed, obeys the superposition principle: For a *collection* of magnetic dipoles, the total momentum is the (vector) sum of the momenta for each one separately. In particular, if $\mathbf{E}$ is *uniform* over a localized steady current distribution, then Eq. 8.45 is valid for the whole thing, only now $\mathbf{m}$ is the *total* magnetic dipole moment.

[22] This problem was suggested by Paul DeYoung.

the torques. Find the resulting extra angular momentum, and compare it with your result in (a). [*Answer:* $-\mu_0 Q^2\omega_b(b^2 - a^2)/4\pi l\,\hat{\mathbf{z}}$]

Problem 8.22[23] A point charge q is a distance $a > R$ from the axis of an infinite solenoid (radius R, n turns per unit length, current I). Find the linear momentum and the angular momentum (with respect to the origin) in the fields. (Put q on the x axis, with the solenoid along z; treat the solenoid as a nonconductor, so you don't need to worry about induced charges on its surface.) [*Answer:* $\mathbf{p} = (\mu_0 qnIR^2/2a)\,\hat{\mathbf{y}}$; $\mathbf{L} = \mathbf{0}$]

Problem 8.23

(a) Carry through the argument in Sect. 8.1.2, starting with Eq. 8.6, but using $\mathbf{J}_f$ in place of $\mathbf{J}$. Show that the Poynting vector becomes

$$\mathbf{S} = \mathbf{E}\times\mathbf{H}, \tag{8.46}$$

and the rate of change of the energy density in the fields is

$$\frac{\partial u}{\partial t} = \mathbf{E}\cdot\frac{\partial \mathbf{D}}{\partial t} + \mathbf{H}\cdot\frac{\partial \mathbf{B}}{\partial t}.$$

For *linear* media, show that[24]

$$u = \frac{1}{2}(\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H}). \tag{8.47}$$

(b) In the same spirit, reproduce the argument in Sect. 8.2.2, starting with Eq. 8.15, with ρ_f and $\mathbf{J}_f$ in place of ρ and $\mathbf{J}$. Don't bother to construct the Maxwell stress tensor, but do show that the momentum density is[25]

$$\mathbf{g} = \mathbf{D}\times\mathbf{B}. \tag{8.48}$$

Problem 8.24

A circular disk of radius R and mass M carries n point charges (q), attached at regular intervals around its rim. At time $t = 0$ the disk lies in the xy plane, with its center at the origin, and is rotating about the z axis with angular velocity ω_0, when it is released. The disk is immersed in a (time-independent) external magnetic field

$$\mathbf{B}(s, z) = k(-s\,\hat{\mathbf{s}} + 2z\,\hat{\mathbf{z}}),$$

where k is a constant.

(a) Find the position of the center if the ring, $z(t)$, and its angular velocity, $\omega(t)$, as functions of time. (Ignore gravity.)

(b) Describe the motion, and check that the total (kinetic) energy—translational plus rotational—is constant, confirming that the magnetic force does no work.[26]

[23] See F. S. Johnson, B. L. Cragin, and R. R. Hodges, *Am. J. Phys.* **62**, 33 (1994), and B. Y.-K. Hu, *Eur. J. Phys.* **33**, 873 (2012), for discussion of this and related problems.

[24] Refer to Sect. 4.4.3 for the meaning of "energy" in this context.

[25] For over 100 years there has been a raging debate (still not completely resolved) as to whether the field momentum in polarizable/magnetizable media is Eq. 8.48 (Minkowski's candidate) or $\epsilon_0\mu_0$ $(\mathbf{E}\times\mathbf{H})$ (Abraham's). See D. J. Griffiths, *Am. J. Phys.* **80**, 7 (2012).

[26] This cute problem is due to K. T. McDonald, http://puhep1.princeton.edu/mcdonald/examles/disk.pdf (who draws a somewhat different conclusion).

Class: B. Tech (Unit II)

I have taken all course materials for Unit II from Book Concept of Modern Physics by Arthur Besier, Shobhit Mahajan & S. Rai Choudhury (McGraw Hill Education).

Students can download this book form given web address;

Web Address : **https://b-ok.cc/book/2700591/864ac0**

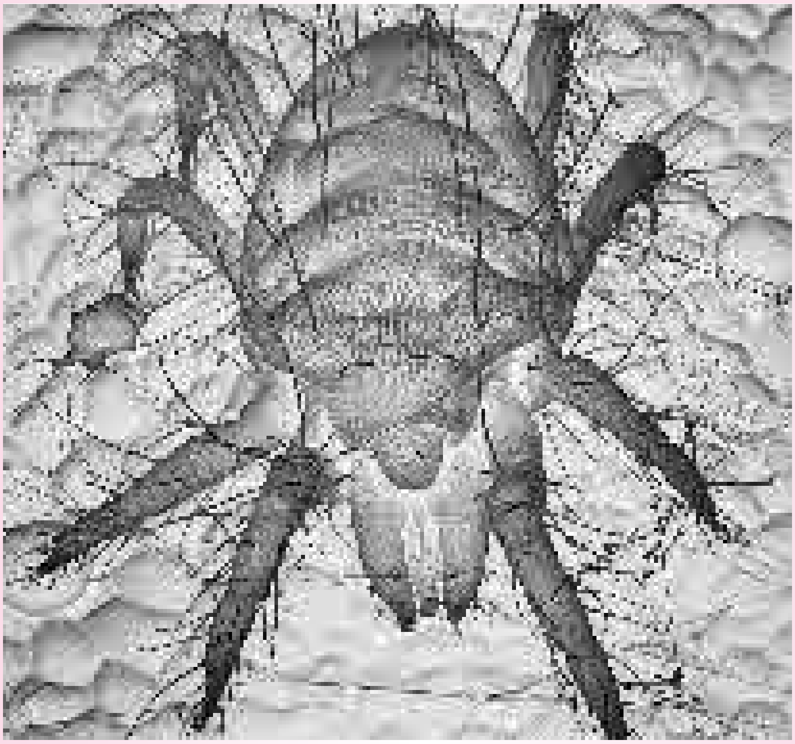
All topics of unit II (Quantum Mechanics) have been taken from **Chapter 3 & Chapter 5** from above said book (**https://b-ok.cc/book/2700591/864ac0**). I am sending pdf file of Chapter 3 & Chapter 5.

UNIT-1I: Quantum Mechanics **(8 Hours)**

Origin of the quantum Mechanics, Interpretation of Wave function, Normalization, Schrodinger time-independent & time-dependent equations, basic postulates of the quantum Mechanics, Probability Current Density, Expectation values, Operators, Hermitian operators, Communication relation between Position & Momentum operators; Applications of Schrödinger equation in Particle in a box, Single step barrier**,** Harmonic Oscillator, Problems.

CHAPTER 3

Wave Properties of Particles



In a scanning electron microscope, an electron beam that scans a specimen causes secondary electrons to be ejected in numbers that vary with the angle of the surface. A suitable data display suggests the three-dimensional form of the specimen. The high resolution of this image of a red spider mite on a leaf is a consequence of the wave nature of moving electrons.

Looking back, it may seem odd that two decades passed between the 1905 discovery of the particle properties of waves and the 1924 speculation that particles might show wave behavior. It is one thing, however, to suggest a revolutionary concept to explain otherwise mysterious data and quite another to suggest an equally revolutionary concept without a strong experimental mandate. The latter is just what Louis de Broglie did in 1924 when he proposed that moving objects have wave as well as particle characteristics. So different was the scientific climate at the time from that around the turn of the century that de Broglie's ideas soon received respectful attention, whereas the earlier quantum theory of light of Planck and Einstein had been largely ignored despite its striking empirical support. The existence of de Broglie waves was experimentally demonstrated by 1927, and the duality principle they represent provided the starting point for Schrödinger's successful development of quantum mechanics in the previous year.

3.1 DE BROGLIE WAVES

A moving body behaves in certain ways as though it has a wave nature

A photon of light of frequency ν has the momentum

$$p = \frac{h\nu}{c} = \frac{h}{\lambda}$$

since $\lambda\nu = c$. The wavelength of a photon is therefore specified by its momentum according to the relation

Photon wavelength

$$\lambda = \frac{h}{p} \tag{3.1}$$

De Broglie suggested that Eq. (3.1) is a completely general one that applies to material particles as well as to photons. The momentum of a particle of mass m and velocity v is $p = \gamma m v$, and its **de Broglie wavelength** is accordingly

De Broglie wavelength

$$\lambda = \frac{h}{\gamma m v} \tag{3.2}$$

Louis de Broglie (1892–1987), although coming from a French family long identified with diplomacy and the military and initially a student of history, eventually followed his older brother Maurice in a career in physics. His doctoral thesis in 1924 contained the proposal that moving bodies have wave properties that complement their particle properties: these "seemingly incompatible conceptions can each represent an aspect of the truth. . . . They may serve in turn to represent the facts without ever entering into direct conflict." Part of de Broglie's inspiration came from Bohr's theory of the hydrogen atom, in which the electron is supposed to follow only certain orbits around the nucleus. "This fact suggested to me the idea that electrons . . . could not be considered simply as particles but that periodicity must be assigned to them also." Two years later Erwin Schrödinger used the concept of de Broglie waves to develop a general theory that he and others applied to explain a wide variety of atomic phenomena. The existence of de Broglie waves was confirmed in diffraction experiments with electron beams in 1927, and in 1929 de Broglie received the Nobel Prize.

The greater the particle's momentum, the shorter its wavelength. In Eq. (3.2) γ is the relativistic factor

$$\gamma = \frac{1}{\sqrt{1 - \upsilon^2/c^2}}$$

As in the case of em waves, the wave and particle aspects of moving bodies can never be observed at the same time. We therefore cannot ask which is the "correct" description. All that can be said is that in certain situations a moving body resembles a wave and in others it resembles a particle. Which set of properties is most conspicuous depends on how its de Broglie wavelength compares with its dimensions and the dimensions of whatever it interacts with.

Example 3.1

Find the de Broglie wavelengths of (*a*) a 46-g golf ball with a velocity of 30 m/s, and (*b*) an electron with a velocity of 10^7 m/s.

Solution

(*a*) Since $\upsilon \ll c$, we can let $\gamma = 1$. Hence

$$\lambda = \frac{h}{m\upsilon} = \frac{6.63 \times 10^{-34}\ \text{J} \cdot \text{s}}{(0.046\ \text{kg})(30\ \text{m/s})} = 4.8 \times 10^{-34}\ \text{m}$$

The wavelength of the golf ball is so small compared with its dimensions that we would not expect to find any wave aspects in its behavior.

(*b*) Again $\upsilon \ll c$, so with $m = 9.1 \times 10^{-31}$ kg, we have

$$\lambda = \frac{h}{m\upsilon} = \frac{6.63 \times 10^{-34}\ \text{J} \cdot \text{s}}{(9.1 \times 10^{-31}\ \text{kg})(10^7\ \text{m/s})} = 7.3 \times 10^{-11}\ \text{m}$$

The dimensions of atoms are comparable with this figure—the radius of the hydrogen atom, for instance, is 5.3×10^{-11} m. It is therefore not surprising that the wave character of moving electrons is the key to understanding atomic structure and behavior.

Example 3.2

Find the kinetic energy of a proton whose de Broglie wavelength is 1.000 fm $= 1.000 \times 10^{-15}$ m, which is roughly the proton diameter.

Solution

A relativistic calculation is needed unless pc for the proton is much smaller than the proton rest energy of $E_0 = 0.938$ GeV. To find out, we use Eq. (3.2) to determine pc:

$$pc = (\gamma m\upsilon)c = \frac{hc}{\lambda} = \frac{(4.136 \times 10^{-15}\ \text{eV} \cdot \text{s})(2.998 \times 10^8\ \text{m/s})}{1.000 \times 10^{-15}\ \text{m}} = 1.240 \times 10^9\ \text{eV}$$
$$= 1.2410\ \text{GeV}$$

Since $pc > E_0$ a relativistic calculation is required. From Eq. (1.24) the total energy of the proton is

$$E = \sqrt{E_0^2 + p^2c^2} = \sqrt{(0.938\ \text{GeV})^2 + (1.2340\ \text{GeV})^2} = 1.555\ \text{GeV}$$

The corresponding kinetic energy is

$$\text{KE} = E - E_0 = (1.555 - 0.938) \text{ GeV} = 0.617 \text{ GeV} = 617 \text{ MeV}$$

De Broglie had no direct experimental evidence to support his conjecture. However, he was able to show that it accounted in a natural way for the energy quantization—the restriction to certain specific energy values—that Bohr had had to postulate in his 1913 model of the hydrogen atom. (This model is discussed in Chap. 4.) Within a few years Eq. (3.2) was verified by experiments involving the diffraction of electrons by crystals. Before we consider one of these experiments, let us look into the question of what kind of wave phenomenon is involved in the matter waves of de Broglie.

3.2 WAVES OF WHAT?

Waves of probability

In water waves, the quantity that varies periodically is the height of the water surface. In sound waves, it is pressure. In light waves, electric and magnetic fields vary. What is it that varies in the case of matter waves?

The quantity whose variations make up matter waves is called the **wave function,** symbol Ψ (the Greek letter psi). The value of the wave function associated with a moving body at the particular point x, y, z in space at the time t is related to the likelihood of finding the body there at the time.

Max Born (1882–1970) grew up in Breslau, then a German city but today part of Poland, and received a doctorate in applied mathematics at Göttingen in 1907. Soon afterward he decided to concentrate on physics, and was back in Göttingen in 1909 as a lecturer. There he worked on various aspects of the theory of crystal lattices, his "central interest" to which he often returned in later years. In 1915, at Planck's recommendation, Born became professor of physics in Berlin where, among his other activities, he played piano to Einstein's violin. After army service in World War I and a period at Frankfurt University, Born was again in Göttingen, now as professor of physics. There a remarkable center of theoretical physics developed under his leadership: Heisenberg and Pauli were among his assistants and Fermi, Dirac, Wigner, and Goeppert were among those who worked with him, just to name future Nobel Prize winners. In those days, Born wrote, "There was complete freedom of teaching and learning in German universities, with no class examinations, and no control of students. The University just offered lectures and the student had to decide for himself which he wished to attend."

Born was a pioneer in going from "the bright realm of classical physics into the still dark and unexplored underworld of the new quantum mechanics;" he was the first to use the latter term. From Born came the basic concept that the wave function Ψ of a particle is related to the probability of finding it. He began with an idea of Einstein, who "sought to make the duality of particles (light quanta or photons) and waves comprehensible by interpreting the square of the optical wave amplitude as probability density for the occurrence of photons. This idea could at once be extended to the Ψ-function: $|\Psi|^2$ must represent the probability density for electrons (or other particles). To assert this was easy; but how was it to be proved? For this purpose atomic scattering processes suggested themselves." Born's development of the quantum theory of atomic scattering (collisions of atoms with various particles) not only verified his "new way of thinking about the phenomena of nature" but also founded an important branch of theoretical physics.

Born left Germany in 1933 at the start of the Nazi period, like so many other scientists. He became a British subject and was associated with Cambridge and then Edinburg universities until he retired in 1953. Finding the Scottish climate harsh and wishing to contribute to the democratization of postwar Germany, Born spent the rest of his life in Bad Pyrmont, a town near Göttingen. His textbooks on modern physics and on optics were standard works on these subjects for many years.

The wave function Ψ itself, however, has no direct physical significance. There is a simple reason why Ψ cannot by interpreted in terms of an experiment. The probability that something be in a certain place at a given time must lie between 0 (the object is definitely not there) and 1 (the object is definitely there). An intermediate probability, say 0.2, means that there is a 20% chance of finding the object. But the amplitude of a wave can be negative as well as positive, and a negative probability, say -0.2, is meaningless. Hence Ψ by itself cannot be an observable quantity.

This objection does not apply to $|\Psi|^2$, the square of the absolute value of the wave function, which is known as **probability density:**

> The probability of experimentally finding the body described by the wave function Ψ at the point x, y, z, at the time t is proportional to the value of $|\Psi|^2$ there at t.

A large value of $|\Psi|^2$ means the strong possibility of the body's presence, while a small value of $|\Psi|^2$ means the slight possibility of its presence. As long as $|\Psi|^2$ is not actually 0 somewhere, however, there is a definite chance, however small, of detecting it there. This interpretation was first made by Max Born in 1926.

There is a big difference between the probability of an event and the event itself. Although we can speak of the wave function Ψ that describes a particle as being spread out in space, this does not mean that the particle itself is thus spread out. When an experiment is performed to detect electrons, for instance, a whole electron is either found at a certain time and place or it is not; there is no such thing as a 20 percent of an electron. However, it is entirely possible for there to be a 20 percent chance that the electron be found at that time and place, and it is this likelihood that is specified by $|\Psi|^2$.

W. L. Bragg, the pioneer in x-ray diffraction, gave this loose but vivid interpretation: "The dividing line between the wave and particle nature of matter and radiation is the moment 'now.' As this moment steadily advances through time it coagulates a wavy future into a particle past. . . . Everything in the future is a wave, everything in the past is a particle." If "the moment 'now' " is understood to be the time a measurement is performed, this is a reasonable way to think about the situation. (The philosopher Søren Kierkegaard may have been anticipating this aspect of modern physics when he wrote, "Life can only be understood backwards, but it must be lived forwards.")

Alternatively, if an experiment involves a great many identical objects all described by the same wave function Ψ, the *actual density* (number per unit volume) of objects at x, y, z at the time t is proportional to the corresponding value of $|\Psi|^2$. It is instructive to compare the connection between Ψ and the density of particles it describes with the connection discussed in Sec. 2.4 between the electric field E of an electromagnetic wave and the density N of photons associated with the wave.

While the wavelength of the de Broglie waves associated with a moving body is given by the simple formula $\lambda = h/\gamma m v$, to find their amplitude Ψ as a function of position and time is often difficult. How to calculate Ψ is discussed in Chap. 5 and the ideas developed there are applied to the structure of the atom in Chap. 6. Until then we can assume that we know as much about Ψ as each situation requires.

3.3 DESCRIBING A WAVE

A general formula for waves

How fast do de Broglie waves travel? Since we associate a de Broglie wave with a moving body, we expect that this wave has the same velocity as that of the body. Let us see if this is true.

If we call the de Broglie wave velocity v_p, we can apply the usual formula

$$v_p = \nu\lambda$$

to find v_p. The wavelength λ is simply the de Broglie wavelength $\lambda = h/\gamma m v$. To find the frequency, we equate the quantum expression $E = h\nu$ with the relativistic formula for total energy $E = \gamma mc^2$ to obtain

$$h\nu = \gamma mc^2$$

$$\nu = \frac{\gamma mc^2}{h}$$

The de Broglie wave velocity is therefore

De Broglie phase velocity

$$v_p = \nu\lambda = \left(\frac{\gamma mc^2}{h}\right)\left(\frac{h}{\gamma m v}\right) = \frac{c^2}{v} \tag{3.3}$$

Because the particle velocity v must be less than the velocity of light c, the de Broglie waves always travel faster than light! In order to understand this unexpected result, we must look into the distinction between **phase velocity** and **group velocity.** (Phase velocity is what we have been calling wave velocity.)

Let us begin by reviewing how waves are described mathematically. For simplicity we consider a string stretched along the x axis whose vibrations are in the y direction, as in Fig. 3.1, and are simple harmonic in character. If we choose $t = 0$ when the displacement y of the string at $x = 0$ is a maximum, its displacement at any future time t at the same place is given by the formula

$$y = A \cos 2\pi\nu t \tag{3.4}$$

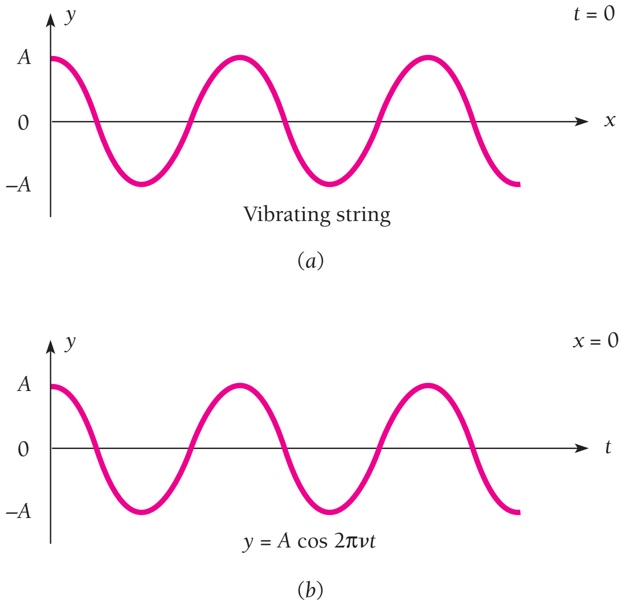


Figure 3.1 (*a*) The appearance of a wave in a stretched string at a certain time. (*b*) How the displacement of a point on the string varies with time.

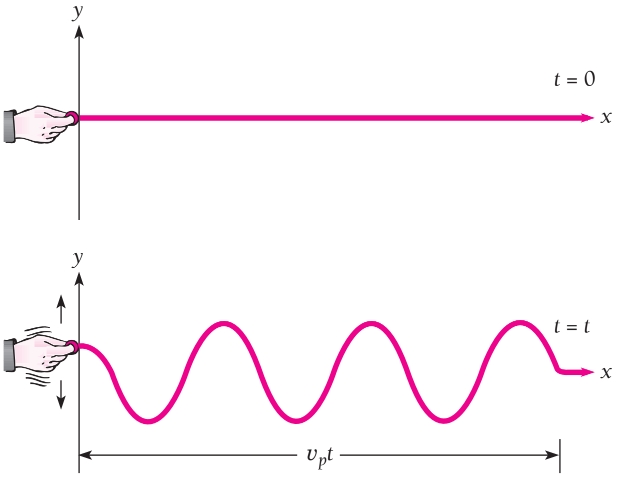


Figure 3.2 Wave propagation.

where A is the amplitude of the vibrations (that is, their maximum displacement on either side of the x axis) and ν their frequency.

Equation (3.4) tells us what the displacement of a single point on the string is as a function of time t. A complete description of wave motion in a stretched string, however, should tell us what y is at *any* point on the string at *any* time. What we want is a formula giving y as a function of both x and t.

To obtain such a formula, let us imagine that we shake the string at $x = 0$ when $t = 0$, so that a wave starts to travel down the string in the $+x$ direction (Fig. 3.2). This wave has some speed v_p that depends on the properties of the string. The wave travels the distance $x = v_p t$ in the time t, so the time interval between the formation of the wave at $x = 0$ and its arrival at the point x is x/v_p. Hence the displacement y of the string at x at any time t is exactly the same as the value of y at $x = 0$ *at the earlier time* $t - x/v_p$. By simply replacing t in Eq. (3.4) with $t - x/v_p$, then, we have the desired formula giving y in terms of both x and t:

Wave formula

$$y = A \cos 2\pi\nu\left(t - \frac{x}{v_p}\right) \tag{3.5}$$

As a check, we note that Eq. (3.5) reduces to Eq. (3.4) at $x = 0$.

Equation (3.5) may be rewritten

$$y = A \cos 2\pi\left(\nu t - \frac{\nu x}{v_p}\right)$$

Since the wave speed v_p is given by $v_p = \nu\lambda$ we have

Wave formula

$$y = A \cos 2\pi\left(\nu t - \frac{x}{\lambda}\right) \tag{3.6}$$

Equation (3.6) is often more convenient to use than Eq. (3.5).

Perhaps the most widely used description of a wave, however, is still another form of Eq. (3.5). The quantities **angular frequency** ω and **wave number** k are defined by the formulas

Angular frequency
$$\omega = 2\pi\nu \tag{3.7}$$

Wave number
$$k = \frac{2\pi}{\lambda} = \frac{\omega}{v_p} \tag{3.8}$$

The unit of ω is the radian per second and that of k is the radian per meter. Angular frequency gets its name from uniform circular motion, where a particle that moves around a circle ν times per second sweeps out $2\pi\nu$ rad/s. The wave number is equal to the number of radians corresponding to a wave train 1 m long, since there are 2π rad in one complete wave.

In terms of ω and k, Eq. (3.5) becomes

Wave formula
$$y = A \cos (\omega t - kx) \tag{3.9}$$

In three dimensions k becomes a vector $\mathbf{k}$ normal to the wave fronts and x is replaced by the radius vector $\mathbf{r}$. The scalar product $\mathbf{k} \cdot \mathbf{r}$ is then used instead of kx in Eq. (3.9).

3.4 PHASE AND GROUP VELOCITIES

A group of waves need not have the same velocity as the waves themselves

The amplitude of the de Broglie waves that correspond to a moving body reflects the probability that it will be found at a particular place at a particular time. It is clear that de Broglie waves cannot be represented simply by a formula resembling Eq. (3.9), which describes an indefinite series of waves all with the same amplitude A. Instead, we expect the wave representation of a moving body to correspond to a **wave packet,** or **wave group,** like that shown in Fig. 3.3, whose waves have amplitudes upon which the likelihood of detecting the body depends.

A familiar example of how wave groups come into being is the case of **beats.** When two sound waves of the same amplitude but of slightly different frequencies are produced simultaneously, the sound we hear has a frequency equal to the average of the two original frequencies and its amplitude rises and falls periodically. The amplitude fluctuations occur as many times per second as the difference between the two original frequencies. If the original sounds have frequencies of, say, 440 and 442 Hz, we will hear a fluctuating sound of frequency 441 Hz with two loudness peaks, called beats, per second. The production of beats is illustrated in Fig. 3.4.

A way to mathematically describe a wave group, then, is in terms of a superposition of individual waves of different wavelengths whose interference with one another results in the variation in amplitude that defines the group shape. If the velocities of the waves are the same, the velocity with which the wave group travels is the common phase velocity. However, if the phase velocity varies with wavelength, the different individual waves do not proceed together. This situation is called **dispersion.** As a result the wave group has a velocity different from the phase velocities of the waves that make it up. This is the case with de Broglie waves.

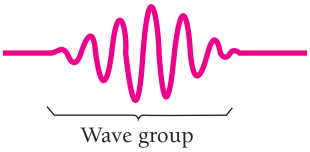


Figure 3.3 A wave group.

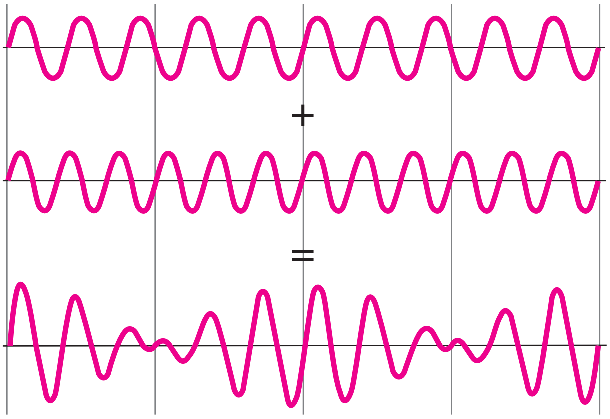


Figure 3.4 Beats are produced by the superposition of two waves with different frequencies.

It is not hard to find the velocity $\boldsymbol{v}_g$ with which a wave group travels. Let us suppose that the wave group arises from the combination of two waves that have the same amplitude A but differ by an amount $\Delta\omega$ in angular frequency and an amount Δk in wave number. We may represent the original waves by the formulas

$$y_1 = A \cos(\omega t - kx)$$
$$y_2 = A \cos[(\omega + \Delta\omega)t - (k + \Delta k)x]$$

The resultant displacement y at any time t and any position x is the sum of y_1 and y_2. With the help of the identity

$$\cos\alpha + \cos\beta = 2\cos\tfrac{1}{2}(\alpha + \beta)\cos\tfrac{1}{2}(\alpha - \beta)$$

and the relation

$$\cos(-\theta) = \cos\theta$$

we find that

$$\begin{aligned} y &= y_1 + y_2 \\ &= 2A\cos\tfrac{1}{2}[(2\omega + \Delta\omega)t - (2k + \Delta k)x]\cos\tfrac{1}{2}(\Delta\omega\, t - \Delta k\, x) \end{aligned}$$

Since $\Delta\omega$ and Δk are small compared with ω and k respectively,

$$2\omega + \Delta\omega \approx 2\omega$$
$$2k + \Delta k \approx 2k$$

and so

Beats

$$y = 2A\cos(\omega t - kx)\cos\left(\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x\right) \tag{3.10}$$

Equation (3.10) represents a wave of angular frequency ω and wave number k that has superimposed upon it a modulation of angular frequency $\frac{1}{2}\Delta\omega$ and of wave number $\frac{1}{2}\Delta k$.

The effect of the modulation is to produce successive wave groups, as in Fig. 3.4. The phase velocity v_p is

Phase velocity

$$v_p = \frac{\omega}{k} \tag{3.11}$$

and the velocity v_g of the wave groups is

Group velocity

$$v_g = \frac{\Delta\omega}{\Delta k} \tag{3.12}$$

When ω and k have continuous spreads instead of the two values in the preceding discussion, the group velocity is instead given by

Group velocity

$$v_g = \frac{d\omega}{dk} \tag{3.13}$$

Depending on how phase velocity varies with wave number in a particular situation, the group velocity may be less or greater than the phase velocities of its member waves. If the phase velocity is the same for all wavelengths, as is true for light waves in empty space, the group and phase velocities are the same.

The angular frequency and wave number of the de Broglie waves associated with a body of mass m moving with the velocity v are

$$\omega = 2\pi\nu = \frac{2\pi\gamma mc^2}{h}$$

Angular frequency of de Broglie waves

$$= \frac{2\pi mc^2}{h\sqrt{1 - v^2/c^2}} \tag{3.14}$$

$$k = \frac{2\pi}{\lambda} = \frac{2\pi\gamma mv}{h}$$

Wave number of de Broglie waves

$$= \frac{2\pi mv}{h\sqrt{1 - v^2/c^2}} \tag{3.15}$$

Both ω and k are functions of the body's velocity v.

The group velocity v_g of the de Broglie waves associated with the body is

$$v_g = \frac{d\omega}{dk} = \frac{d\omega/dv}{dk/dv}$$

Now

$$\frac{d\omega}{dv} = \frac{2\pi mv}{h(1 - v^2/c^2)^{3/2}}$$

$$\frac{dk}{dv} = \frac{2\pi m}{h(1 - v^2/c^2)^{3/2}}$$

Electron Microscopes

The wave nature of moving electrons is the basis of the electron microscope, the first of which was built in 1932. The resolving power of any optical instrument, which is limited by diffraction, is proportional to the wavelength of whatever is used to illuminate the specimen. In the case of a good microscope that uses visible light, the maximum useful magnification is about 500×; higher magnifications give larger images but do not reveal any more detail. Fast electrons, however, have wavelengths very much shorter than those of visible light and are easily controlled by electric and magnetic fields because of their charge. X-rays also have short wavelengths, but it is not (yet?) possible to focus them adequately.

In an electron microscope, current-carrying coils produce magnetic fields that act as lenses to focus an electron beam on a specimen and then produce an enlarged image on a fluorescent screen or photographic plate (Fig. 3.5). To prevent the beam from being scattered and thereby blurring the image, a thin specimen is used and the entire system is evacuated.

The technology of magnetic "lenses" does not permit the full theoretical resolution of electron waves to be realized in practice. For instance, 100-keV electrons have wavelengths of 0.0037 nm, but the actual resolution they can provide in an electron microscope may be only about 0.1 nm. However, this is still a great improvement on the ~200-nm resolution of an optical microscope, and magnifications of over 1,000,000× have been achieved with electron microscopes.

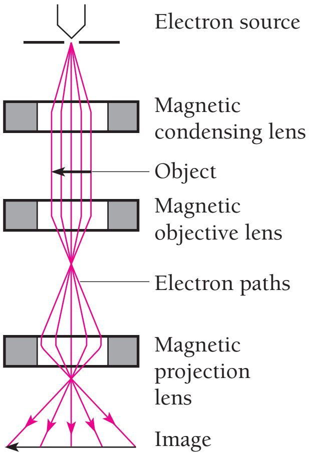
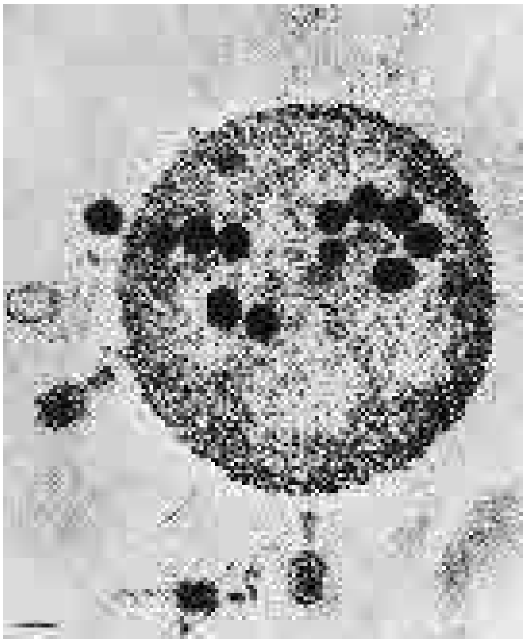
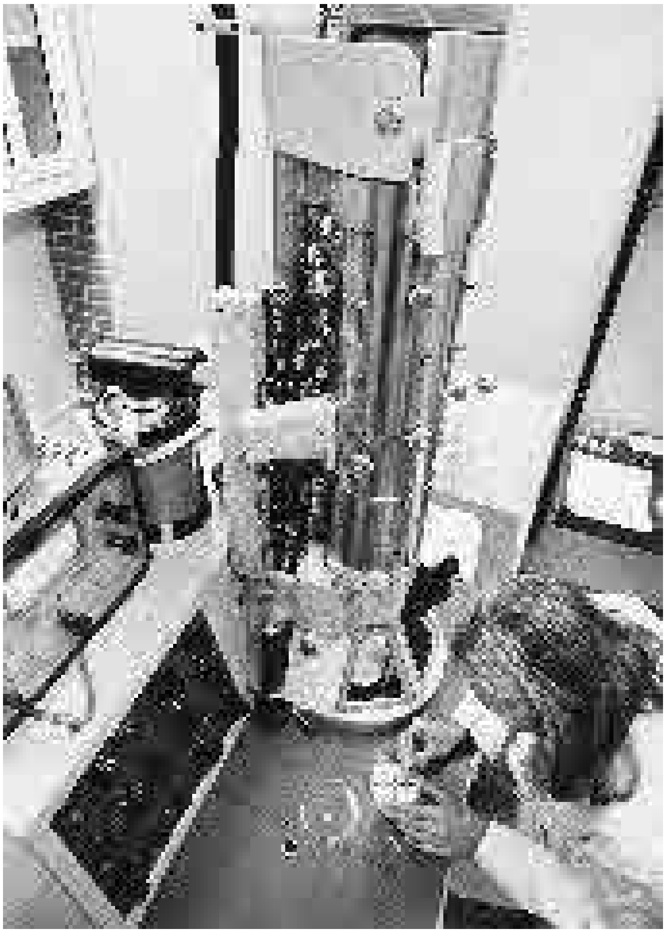


Figure 3.5 Because the wavelengths of the fast electrons in an electron microscope are shorter than those of the light waves in an optical microscope, the electron microscope can produce sharp images at higher magnifications. The electron beam in an electron microscope is focused by magnetic fields.



Electron micrograph showing bacteriophage viruses in an *Escherichia coli* bacterium. The bacterium is approximately 1 μm across.



An electron microscope.

and so the group velocity turns out to be

De Broglie group velocity

$$\upsilon_g = \upsilon \tag{3.16}$$

The de Broglie wave group associated with a moving body travels with the same velocity as the body.

The phase velocity υ_p of de Broglie waves is, as we found earlier,

De Broglie phase velocity

$$\upsilon_p = \frac{\omega}{k} = \frac{c^2}{\upsilon} \tag{3.3}$$

This exceeds both the velocity of the body υ and the velocity of light c, since $\upsilon < c$. However, υ_p has no physical significance because the motion of the wave group, not the motion of the individual waves that make up the group, corresponds to the motion of the body, and $\upsilon_g < c$ as it should be. The fact that $\upsilon_p > c$ for de Broglie waves therefore does not violate special relativity.

Example 3.3

An electron has a de Broglie wavelength of 2.00 pm = 2.00×10^{-12} m. Find its kinetic energy and the phase and group velocities of its de Broglie waves.

Solution

(*a*) The first step is to calculate pc for the electron, which is

$$pc = \frac{hc}{\lambda} = \frac{(4.136 \times 10^{-15}\ \text{eV}\cdot\text{s})(3.00 \times 10^{8}\ \text{m/s})}{2.00 \times 10^{-12}\ \text{m}} = 6.20 \times 10^{5}\ \text{eV}$$
$$= 620\ \text{keV}$$

The rest energy of the electron is $E_0 = 511$ keV, so

$$\text{KE} = E - E_0 = \sqrt{E_0^2 + (pc)^2} - E_0 = \sqrt{(511\ \text{keV})^2 + (620\ \text{keV})^2} - 511\ \text{keV}$$
$$= 803\ \text{keV} - 511\ \text{keV} = 292\ \text{keV}$$

(*b*) The electron velocity can be found from

$$E = \frac{E_0}{\sqrt{1 - \upsilon^2/c^2}}$$

to be

$$\upsilon = c\sqrt{1 - \frac{E_0^2}{E^2}} = c\sqrt{1 - \left(\frac{511\ \text{keV}}{803\ \text{keV}}\right)^2} = 0.771c$$

Hence the phase and group velocities are respectively

$$\upsilon_p = \frac{c^2}{\upsilon} = \frac{c^2}{0.771c} = 1.30c$$
$$\upsilon_g = \upsilon = 0.771c$$

3.5 PARTICLE DIFFRACTION

An experiment that confirms the existence of de Broglie waves

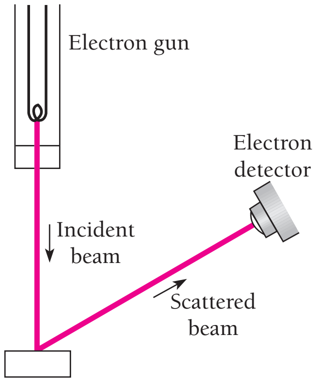


Figure 3.6 The Davisson-Germer experiment.

A wave effect with no analog in the behavior of Newtonian particles is diffraction. In 1927 Clinton Davisson and Lester Germer in the United States and G. P. Thomson in England independently confirmed de Broglie's hypothesis by demonstrating that electron beams are diffracted when they are scattered by the regular atomic arrays of crystals. (All three received Nobel Prizes for their work. J. J. Thomson, G. P.'s father, had earlier won a Nobel Prize for verifying the particle nature of the electron: the wave-particle duality seems to have been the family business.) We shall look at the experiment of Davisson and Germer because its interpretation is more direct.

Davisson and Germer were studying the scattering of electrons from a solid using an apparatus like that sketched in Fig. 3.6. The energy of the electrons in the primary beam, the angle at which they reach the target, and the position of the detector could all be varied. Classical physics predicts that the scattered electrons will emerge in all directions with only a moderate dependence of their intensity on scattering angle and even less on the energy of the primary electrons. Using a block of nickel as the target, Davisson and Germer verified these predictions.

In the midst of their work an accident occurred that allowed air to enter their apparatus and oxidize the metal surface. To reduce the oxide to pure nickel, the target was baked in a hot oven. After this treatment, the target was returned to the apparatus and the measurements resumed.

Now the results were very different. Instead of a continuous variation of scattered electron intensity with angle, distinct maxima and minima were observed whose positions depended upon the electron energy! Typical polar graphs of electron intensity after the accident are shown in Fig. 3.7. The method of plotting is such that the intensity at any angle is proportional to the distance of the curve at that angle from the point of scattering. If the intensity were the same at all scattering angles, the curves would be circles centered on the point of scattering.

Two questions come to mind immediately: What is the reason for this new effect? Why did it not appear until after the nickel target was baked?

De Broglie's hypothesis suggested that electron waves were being diffracted by the target, much as x-rays are diffracted by planes of atoms in a crystal. This idea received

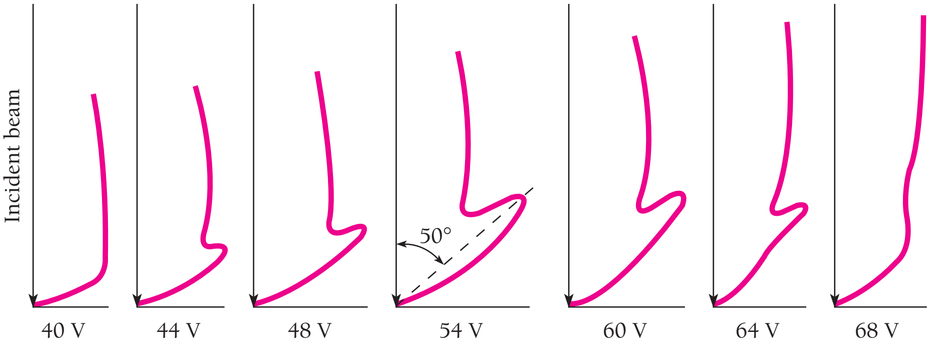


Figure 3.7 Results of the Davisson-Germer experiment, showing how the number of scattered electrons varied with the angle between the incoming beam and the crystal surface. The Bragg planes of atoms in the crystal were not parallel to the crystal surface, so the angles of incidence and scattering relative to one family of these planes were both 65° (see Fig. 3.8).

support when it was realized that heating a block of nickel at high temperature causes the many small individual crystals of which it is normally composed to form into a single large crystal, all of whose atoms are arranged in a regular lattice.

Let us see whether we can verify that de Broglie waves are responsible for the findings of Davisson and Germer. In a particular case, a beam of 54-eV electrons was directed perpendicularly at the nickel target and a sharp maximum in the electron distribution occurred at an angle of 50° with the original beam. The angles of incidence and scattering relative to the family of Bragg planes shown in Fig. 3.8 are both 65°. The spacing of the planes in this family, which can be measured by x-ray diffraction, is 0.091 nm. The Bragg equation for maxima in the diffraction pattern is

$$n\lambda = 2d \sin \theta \tag{2.13}$$

Here $d = 0.091$ nm and $\theta = 65°$. For $n = 1$ the de Broglie wavelength λ of the diffracted electrons is

$$\lambda = 2d \sin \theta = (2)(0.091 \text{ nm})(\sin 65°) = 0.165 \text{ nm}$$

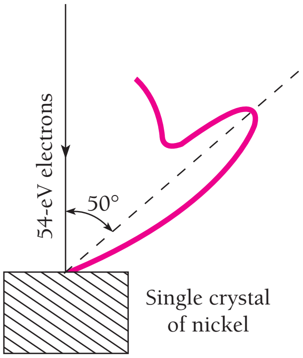


Figure 3.8 The diffraction of the de Broglie waves by the target is responsible for the results of Davisson and Germer.

Now we use de Broglie's formula $\lambda = h/\gamma m v$ to find the expected wavelength of the electrons. The electron kinetic energy of 54 eV is small compared with its rest energy mc^2 of 0.51 MeV, so we can let $\gamma = 1$. Since

$$\text{KE} = \tfrac{1}{2}mv^2$$

the electron momentum mv is

$$\begin{aligned} mv &= \sqrt{2m\text{KE}} \\ &= \sqrt{(2)(9.1 \times 10^{-31} \text{ kg})(54 \text{ eV})(1.6 \times 10^{-19} \text{ J/eV})} \\ &= 4.0 \times 10^{-24} \text{ kg} \cdot \text{m/s} \end{aligned}$$

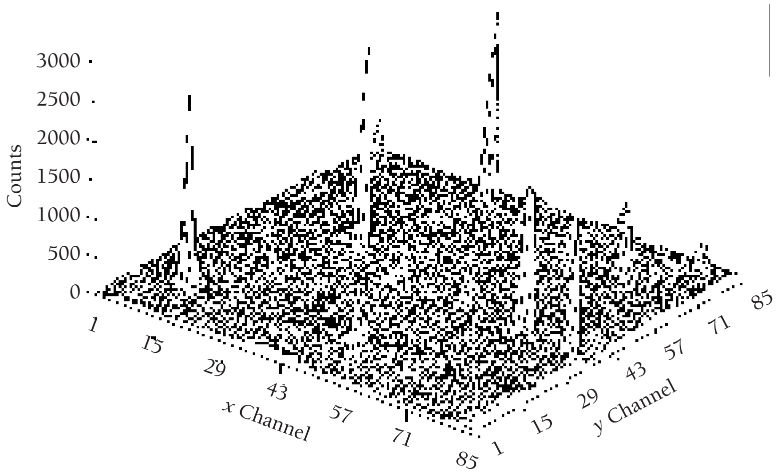
The electron wavelength is therefore

$$\lambda = \frac{h}{mv} = \frac{6.63 \times 10^{-34} \text{ J} \cdot \text{s}}{4.0 \times 10^{-24} \text{ kg} \cdot \text{m/s}} = 1.66 \times 10^{-10} \text{ m} = 0.166 \text{ nm}$$

which agrees well with the observed wavelength of 0.165 nm. The Davisson-Germer experiment thus directly verifies de Broglie's hypothesis of the wave nature of moving bodies.

Analyzing the Davisson-Germer experiment is actually less straightforward than indicated above because the energy of an electron increases when it enters a crystal by an amount equal to the work function of the surface. Hence the electron speeds in the experiment were greater inside the crystal and the de Broglie wavelengths there shorter than the values outside. Another complication arises from interference between waves diffracted by different families of Bragg planes, which restricts the occurrence of maxima to certain combinations of electron energy and angle of incidence rather than merely to any combination that obeys the Bragg equation.

Electrons are not the only bodies whose wave behavior can be demonstrated. The diffraction of neutrons and of whole atoms when scattered by suitable crystals has been observed, and in fact neutron diffraction, like x-ray and electron diffraction, has been used for investigating crystal structures.



Neutron diffraction by a quartz crystal. The peaks represent directions in which constructive interference occurred. (*Courtesy Frank J. Rotella and Arthur J. Schultz, Argonne National Laboratory*)

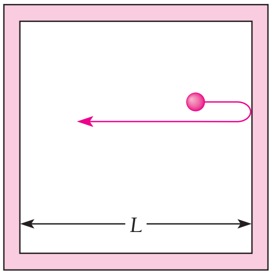


Figure 3.9 A particle confined to a box of width L. The particle is assumed to move back and forth along a straight line between the walls of the box.

3.6 PARTICLE IN A BOX

Why the energy of a trapped particle is quantized

The wave nature of a moving particle leads to some remarkable consequences when the particle is restricted to a certain region of space instead of being able to move freely.

The simplest case is that of a particle that bounces back and forth between the walls of a box, as in Fig. 3.9. We shall assume that the walls of the box are infinitely hard, so the particle does not lose energy each time it strikes a wall, and that its velocity is sufficiently small so that we can ignore relativistic considerations. Simple as it is, this model situation requires fairly elaborate mathematics in order to be properly analyzed, as we shall learn in Chap. 5. However, even a relatively crude treatment can reveal the essential results.

From a wave point of view, a particle trapped in a box is like a standing wave in a string stretched between the box's walls. In both cases the wave variable (transverse displacement for the string, wave function Ψ for the moving particle) must be 0 at the walls, since the waves stop there. The possible de Broglie wavelengths of the particle in the box therefore are determined by the width L of the box, as in Fig. 3.10. The longest wavelength is specified by $\lambda = 2L$, the next by $\lambda = L$, then $\lambda = 2L/3$, and so forth. The general formula for the permitted wavelengths is

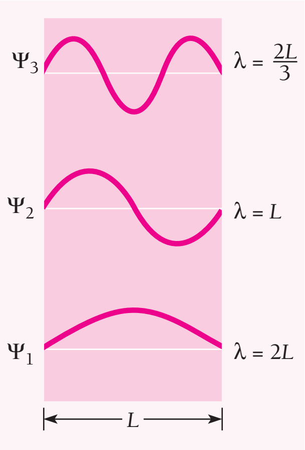


Figure 3.10 Wave functions of a particle trapped in a box L wide.

De Broglie wavelengths of trapped particle

$$\lambda_n = \frac{2L}{n} \qquad n = 1, 2, 3, \ldots \tag{3.17}$$

Because $mv = h/\lambda$, the restrictions on de Broglie wavelength λ imposed by the width of the box are equivalent to limits on the momentum of the particle and, in turn, to limits on its kinetic energy. The kinetic energy of a particle of momentum mv is

$$\mathrm{KE} = \tfrac{1}{2}mv^2 = \frac{(mv)^2}{2m} = \frac{h^2}{2m\lambda^2}$$

The permitted wavelengths are $\lambda_n = 2L/n$, and so, because the particle has no potential energy in this model, the only energies it can have are

Particle in a box
$$E_n = \frac{n^2h^2}{8mL^2} \qquad n = 1, 2, 3, \ldots \tag{3.18}$$

Each permitted energy is called an **energy level**, and the integer n that specifies an energy level E_n is called its **quantum number.**

We can draw three general conclusions from Eq. (3.18). These conclusions apply to *any* particle confined to a certain region of space (even if the region does not have a well-defined boundary), for instance an atomic electron held captive by the attraction of the positively charged nucleus.

1 A trapped particle cannot have an arbitrary energy, as a free particle can. The fact of its confinement leads to restrictions on its wave function that allow the particle to have only certain specific energies and no others. Exactly what these energies are depends on the mass of the particle and on the details of how it is trapped.

2 A trapped particle cannot have zero energy. Since the de Broglie wavelength of the particle is $\boldsymbol{\lambda} = h/m\boldsymbol{v}$, a speed of $\boldsymbol{v} = 0$ means an infinite wavelength. But there is no way to reconcile an infinite wavelength with a trapped particle, so such a particle must have at least some kinetic energy. The exclusion of $E = 0$ for a trapped particle, like the limitation of E to a set of discrete values, is a result with no counterpart in classical physics, where all non-negative energies, including zero, are allowed.

3 Because Planck's constant is so small—only 6.63×10^{-34} J · s—quantization of energy is conspicuous only when m and L are also small. This is why we are not aware of energy quantization in our own experience. Two examples will make this clear.

Example 3.4

An electron is in a box 0.10 nm across, which is the order of magnitude of atomic dimensions. Find its permitted energies.

Solution

Here $m = 9.1 \times 10^{-31}$ kg and $L = 0.10$ nm $= 1.0 \times 10^{-10}$ m, so that the permitted electron energies are

$$E_n = \frac{(n^2)(6.63 \times 10^{-34}\ \text{J} \cdot \text{s})^2}{(8)(9.1 \times 10^{-31}\ \text{kg})(1.0 \times 10^{-10}\ \text{m})^2} = 6.0 \times 10^{-18} n^2\ \text{J}$$
$$= 38n^2\ \text{eV}$$

The minimum energy the electron can have is 38 eV, corresponding to $n = 1$. The sequence of energy levels continues with $E_2 = 152$ eV, $E_3 = 342$ eV, $E_4 = 608$ eV, and so on (Fig. 3.11). If such a box existed, the quantization of a trapped electron's energy would be a prominent feature of the system. (And indeed energy quantization is prominent in the case of an atomic electron.)

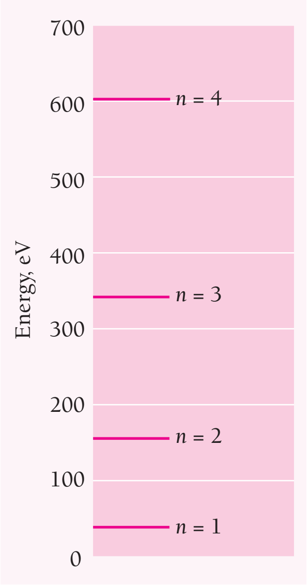


Figure 3.11 Energy levels of an electron confined to a box 0.1 nm wide.

Example 3.5

A 10-g marble is in a box 10 cm across. Find its permitted energies.

Solution

With $m = 10$ g $= 1.0 \times 10^{-2}$ kg and $L = 10$ cm $= 1.0 \times 10^{-1}$ m,

$$E_n = \frac{(n^2)(6.63 \times 10^{-34}\ \text{J} \cdot \text{s})^2}{(8)(1.0 \times 10^{-2}\ \text{kg})(1.0 \times 10^{-1}\ \text{m})^2}$$
$$= 5.5 \times 10^{-64} n^2\ \text{J}$$

The minimum energy the marble can have is 5.5×10^{-64} J, corresponding to $n = 1$. A marble with this kinetic energy has a speed of only 3.3×10^{-31} m/s and therefore cannot be experimentally distinguished from a stationary marble. A reasonable speed a marble might have is, say, $\frac{1}{3}$ m/s—which corresponds to the energy level of quantum number $n = 10^{30}$! The permissible energy levels are so very close together, then, that there is no way to determine whether the marble can take on only those energies predicted by Eq. (3.18) or any energy whatever. Hence in the domain of everyday experience, quantum effects are imperceptible, which accounts for the success of Newtonian mechanics in this domain.

3.7 UNCERTAINTY PRINCIPLE 1

We cannot know the future because we cannot know the present

To regard a moving particle as a wave group implies that there are fundamental limits to the accuracy with which we can measure such "particle" properties as position and momentum.

To make clear what is involved, let us look at the wave group of Fig. 3.3. The particle that corresponds to this wave group may be located anywhere within the group at a given time. Of course, the probability density $|\Psi|^2$ is a maximum in the middle of the group, so it is most likely to be found there. Nevertheless, we may still find the particle anywhere that $|\Psi|^2$ is not actually 0.

The narrower its wave group, the more precisely a particle's position can be specified (Fig. 3.12*a*). However, the wavelength of the waves in a narrow packet is not well defined; there are not enough waves to measure λ accurately. This means that since $\lambda = h/\gamma mv$, the particle's momentum γmv is not a precise quantity. If we make a series of momentum measurements, we will find a broad range of values.

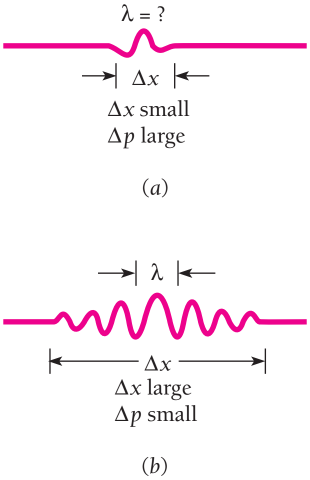


Figure 3.12 (*a*) A narrow de Broglie wave group. The position of the particle can be precisely determined, but the wavelength (and hence the particle's momentum) cannot be established because there are not enough waves to measure accurately. (*b*) A wide wave group. Now the wavelength can be precisely determined but not the position of the particle.

On the other hand, a wide wave group, such as that in Fig. 3.12*b*, has a clearly defined wavelength. The momentum that corresponds to this wavelength is therefore a precise quantity, and a series of measurements will give a narrow range of values. But where is the particle located? The width of the group is now too great for us to be able to say exactly where the particle is at a given time.

Thus we have the **uncertainty principle:**

It is impossible to know both the exact position and exact momentum of an object at the same time.

This principle, which was discovered by Werner Heisenberg in 1927, is one of the most significant of physical laws.

A formal analysis supports the above conclusion and enables us to put it on a quantitative basis. The simplest example of the formation of wave groups is that given in Sec. 3.4, where two wave trains slightly different in angular frequency ω and wave number k were superposed to yield the series of groups shown in Fig. 3.4. A moving body corresponds to a single wave group, not a series of them, but a single wave group can also be thought of in terms of the superposition of trains of harmonic waves. However, an infinite number of wave trains with different frequencies, wave numbers, and amplitudes is required for an isolated group of arbitrary shape, as in Fig. 3.13.

At a certain time t, the wave group $\Psi(x)$ can be represented by the **Fourier integral**

$$\Psi(x) = \int_0^\infty g(k) \cos kx \, dk \qquad (3.19)$$

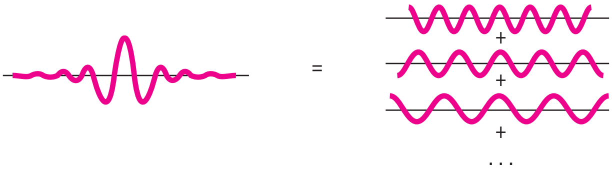


Figure 3.13 An isolated wave group is the result of superposing an infinite number of waves with different wavelengths. The narrower the wave group, the greater the range of wavelengths involved. A narrow de Broglie wave group thus means a well-defined position (Δx smaller) but a poorly defined wavelength and a large uncertainty Δp in the momentum of the particle the group represents. A wide wave group means a more precise momentum but a less precise position.

where the function $g(k)$ describes how the amplitudes of the waves that contribute to $\Psi(x)$ vary with wave number k. This function is called the **Fourier transform** of $\Psi(x)$, and it specifies the wave group just as completely as $\Psi(x)$ does. Figure 3.14 contains graphs of the Fourier transforms of a pulse and of a wave group. For comparison, the Fourier transform of an infinite train of harmonic waves is also included. There is only a single wave number in this case, of course.

Strictly speaking, the wave numbers needed to represent a wave group extend from $k = 0$ to $k = \infty$, but for a group whose length Δx is finite, the waves whose amplitudes $g(k)$ are appreciable have wave numbers that lie within a finite interval Δk. As Fig. 3.14 indicates, the narrower the group, the broader the range of wave numbers needed to describe it, and vice versa.

The relationship between the distance Δx and the wave-number spread Δk depends upon the shape of the wave group and upon how Δx and Δk are defined. The minimum value of the product $\Delta x\,\Delta k$ occurs when the envelope of the group has the familiar bell shape of a Gaussian function. In this case the Fourier transform happens to be a Gaussian function also. If Δx and Δk are taken as the standard deviations of the respective functions $\Psi(x)$ and $g(k)$, then this minimum value is $\Delta x\,\Delta k = \frac{1}{2}$. Because wave groups in general do not have Gaussian forms, it is more realistic to express the relationship between Δx and Δk as

$$\Delta x\,\Delta k \geq \tfrac{1}{2} \qquad (3.20)$$

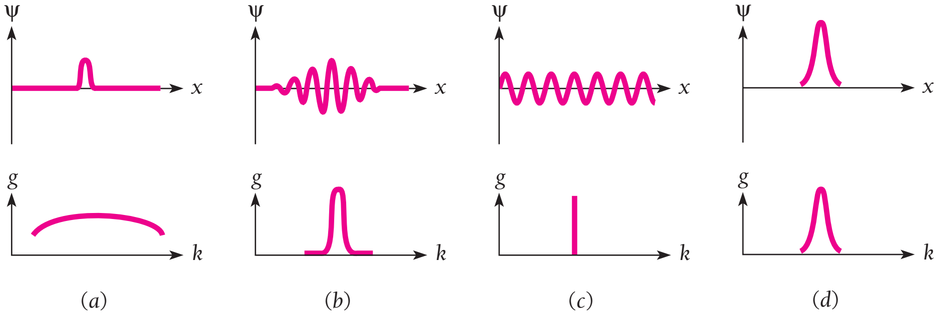


Figure 3.14 The wave functions and Fourier transforms for (*a*) a pulse, (*b*) a wave group, (*c*) a wave train, and (*d*) a Gaussian distribution. A brief disturbance needs a broader range of frequencies to describe it than a disturbance of greater duration. The Fourier transform of a Gaussian function is also a Gaussian function.

Gaussian Function

When a set of measurements is made of some quantity x in which the experimental errors are random, the result is often a **Gaussian distribution** whose form is the bell-shaped curve shown in Fig. 3.15. The **standard deviation σ** of the measurements is a measure of the spread of x values about the mean of x_0, where σ equals the square root of the average of the squared deviations from x_0. If N measurements were made,

Standard deviation
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_1 - x_0)^2}$$

The width of a Gaussian curve at half its maximum value is 2.35σ.

The ***Gaussian function*** $f(x)$ that describes the above curve is given by

Gaussian function
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-x_0)^2/2\sigma^2}$$

where $f(x)$ is the probability that the value x be found in a particular measurement. Gaussian functions occur elsewhere in physics and mathematics as well. (Gabriel Lippmann had this to say about the Gaussian function: "Experimentalists think that it is a mathematical theorem while mathematicians believe it to be an experimental fact.")

The probability that a measurement lie inside a certain range of x values, say between x_1 and x_2, is given by the area of the $f(x)$ curve between these limits. This area is the integral

$$P_{x_1x_2} = \int_{x_1}^{x_2} f(x)\,dx$$

An interesting questions is what fraction of a series of measurements has values within a standard deviation of the mean value x_0. In this case $x_1 = x_0 - \sigma$ and $x_2 = x_0 + \sigma$, and

$$P_{x_0\pm\sigma} = \int_{x_0-\sigma}^{x_0+\sigma} f(x)\,dx = 0.683$$

Hence 68.3 percent of the measurements fall in this interval, which is shaded in Fig. 3.15. A similar calculation shows that 95.4 percent of the measurements fall within two standard deviations of the mean value.

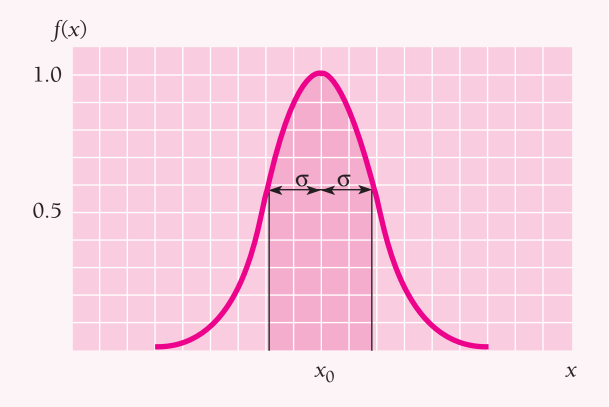


Figure 3.15 A Gaussian distribution. The probability of finding a value of x is given by the Gaussian function $f(x)$. The mean value of x is x_0, and the total width of the curve at half its maximum value is 2.35σ, where σ is the standard deviation of the distribution. The total probability of finding a value of x within a standard deviation of x_0 is equal to the shaded area and is 68.3 percent.

The de Broglie wavelength of a particle of momentum p is $\lambda = h/p$ and the corresponding wave number is

$$k = \frac{2\pi}{\lambda} = \frac{2\pi p}{h}$$

In terms of wave number the particle's momentum is therefore

$$p = \frac{hk}{2\pi}$$

Hence an uncertainty Δk in the wave number of the de Broglie waves associated with the particle results in an uncertainty Δp in the particle's momentum according to the formula

$$\Delta p = \frac{h\,\Delta k}{2\pi}$$

Since $\Delta x\,\Delta k \geq \frac{1}{2}$, $\Delta k \geq 1/(2\Delta x)$ and

Uncertainty principle

$$\Delta x\,\Delta p \geq \frac{h}{4\pi} \tag{3.21}$$

This equation states that the product of the uncertainty Δx in the position of an object at some instant and the uncertainty Δp in its momentum component in the x direction at the same instant is equal to or greater than $h/4\pi$.

If we arrange matters so that Δx is small, corresponding to a narrow wave group, then Δp will be large. If we reduce Δp in some way, a broad wave group is inevitable and Δx will be large.

Werner Heisenberg (1901–1976) was born in Duisberg, Germany, and studied theoretical physics at Munich, where he also became an enthusiastic skier and mountaineer. At Göttingen in 1924 as an assistant to Max Born, Heisenberg became uneasy about mechanical models of the atom: "Any picture of the atom that our imagination is able to invent is for that very reason defective," he later remarked. Instead he conceived an abstract approach using matrix algebra. In 1925, together with Born and Pascual Jordan, Heisenberg developed this approach into a consistent theory of quantum mechanics, but it was so difficult to understand and apply that it had very little impact on physics at the time. Schrödinger's wave formulation of quantum mechanics the following year was much more successful; Schrödinger and others soon showed that the wave and matrix versions of quantum mechanics were mathematically equivalent.

In 1927, working at Bohr's institute in Copenhagen, Heisenberg developed a suggestion by Wolfgang Pauli into the uncertainty principle. Heisenberg initially felt that this principle was a consequence of the disturbances inevitably produced by any measuring process. Bohr, on the other hand, thought that the basic cause of the uncertainties was the wave-particle duality, so that they were built into the natural world rather than solely the result of measurement. After much argument Heisenberg came around to Bohr's view. (Einstein, always skeptical about quantum mechanics, said after a lecture by Heisenberg on the uncertainty principle: "Marvelous, what ideas the young people have these days. But I don't believe a word of it.") Heisenberg received the Nobel Prize in 1932.

Heisenberg was one of the very few distinguished scientists to remain in Germany during the Nazi period. In World War II he led research there on atomic weapons, but little progress had been made by the war's end. Exactly why remains unclear, although there is no evidence that Heisenberg, as he later claimed, had moral qualms about creating such weapons and more or less deliberately dragged his feet. Heisenberg recognized early that "an explosive of unimaginable consequences" could be developed, and he and his group should have been able to have gotten farther than they did. In fact, alarmed by the news that Heisenberg was working on an atomic bomb, the U.S. government sent the former Boston Red Sox catcher Moe Berg to shoot Heisenberg during a lecture in neutral Switzerland in 1944. Berg, sitting in the second row, found himself uncertain from Heisenberg's remarks about how advanced the German program was, and kept his gun in his pocket.

These uncertainties are due not to inadequate apparatus but to the imprecise character in nature of the quantities involved. Any instrumental or statistical uncertainties that arise during a measurement only increase the product $\Delta x\, \Delta p$. Since we cannot know exactly both where a particle is right now and what its momentum is, we cannot say anything definite about where it will be in the future or how fast it will be moving then. *We cannot know the future for sure because we cannot know the present for sure*. But our ignorance is not total: we can still say that the particle is more likely to be in one place than another and that its momentum is more likely to have a certain value than another.

H-Bar

The quantity $h/2\pi$ appears often in modern physics because it turns out to be the basic unit of angular momentum. It is therefore customary to abbreviate $h/2\pi$ by the symbol $\hbar$ ("h-bar"):

$$\hbar = \frac{h}{2\pi} = 1.054 \times 10^{-34}\ \text{J} \cdot \text{s}$$

In the remainder of this book $\hbar$ is used in place of $h/2\pi$. In terms of $\hbar$, the uncertainty principle becomes

Uncertainty principle

$$\Delta x\, \Delta p \geq \frac{\hbar}{2} \tag{3.22}$$

Example 3.6

A measurement establishes the position of a proton with an accuracy of $\pm 1.00 \times 10^{-11}$ m. Find the uncertainty in the proton's position 1.00 s later. Assume $\boldsymbol{v} \ll c$.

Solution

Let us call the uncertainty in the proton's position Δx_0 at the time $t = 0$. The uncertainty in its momentum at this time is therefore, from Eq. (3.22),

$$\Delta p \geq \frac{\hbar}{2\Delta x_0}$$

Since $\boldsymbol{v} \ll c$, the momentum uncertainty is $\Delta p = \Delta(m\boldsymbol{v}) = m\, \Delta\boldsymbol{v}$ and the uncertainty in the proton's velocity is

$$\Delta \boldsymbol{v} = \frac{\Delta p}{m} \geq \frac{\hbar}{2m\, \Delta x_0}$$

The distance x the proton covers in the time t cannot be known more accurately than

$$\Delta x = t\, \Delta \boldsymbol{v} \geq \frac{\hbar t}{2m\, \Delta x_0}$$

Hence Δx is inversely proportional to Δx_0: the *more* we know about the proton's position at $t = 0$, the *less* we know about its later position at $t > 0$. The value of Δx at $t = 1.00$ s is

$$\Delta x \geq \frac{(1.054 \times 10^{-34}\ \text{J} \cdot \text{s})(1.00\ \text{s})}{(2)(1.672 \times 10^{-27}\ \text{kg})(1.00 \times 10^{-11}\ \text{m})}$$

$$\geq 3.15 \times 10^{3}\ \text{m}$$

This is 3.15 km—nearly 2 mi! What has happened is that the original wave group has spread out to a much wider one (Fig. 3.16). This occurred because the phase velocities of the component waves vary with wave number and a large range of wave numbers must have been present to produce the narrow original wave group. See Fig. 3.14.

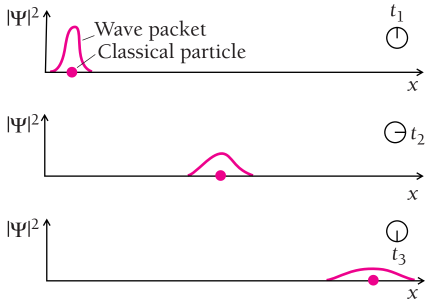


Figure 3.16 The wave packet that corresponds to a moving packet is a composite of many individual waves, as in Fig. 3.13. The phase velocities of the individual waves vary with their wave lengths. As a result, as the particle moves, the wave packet spreads out in space. The narrower the original wavepacket—that is, the more precisely we know its position at that time—the more it spreads out because it is made up of a greater span of waves with different phase velocities.

3.8 UNCERTAINTY PRINCIPLE II

A particle approach gives the same result

The uncertainty principle can be arrived at from the point of view of the particle properties of waves as well as from the point of view of the wave properties of particles.

We might want to measure the position and momentum of an object at a certain moment. To do so, we must touch it with something that will carry the required information back to us. That is, we must poke it with a stick, shine light on it, or perform some similar act. The measurement process itself thus requires that the object be interfered with in some way. If we consider such interferences in detail, we are led to the same uncertainty principle as before even without taking into account the wave nature of moving bodies.

Suppose we look at an electron using light of wavelength λ, as in Fig. 3.17. Each photon of this light has the momentum h/λ. When one of these photons bounces off the electron (which must happen if we are to "see" the electron), the electron's

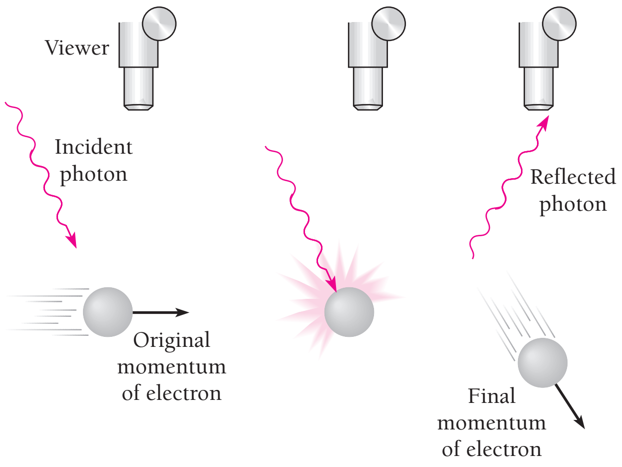


Figure 3.17 An electron cannot be observed without changing its momentum.

original momentum will be changed. The exact amount of the change Δp cannot be predicted, but it will be of the same order of magnitude as the photon momentum h/λ. Hence

$$\Delta p \approx \frac{h}{\lambda} \tag{3.23}$$

The longer the wavelength of the observing photon, the smaller the uncertainty in the electron's momentum.

Because light is a wave phenomenon as well as a particle phenomenon, we cannot expect to determine the electron's location with perfect accuracy regardless of the instrument used. A reasonable estimate of the minimum uncertainty in the measurement might be one photon wavelength, so that

$$\Delta x \geq \lambda \tag{3.24}$$

The shorter the wavelength, the smaller the uncertainty in location. However, if we use light of short wavelength to increase the accuracy of the position measurement, there will be a corresponding decrease in the accuracy of the momentum measurement because the higher photon momentum will disturb the electron's motion to a greater extent. Light of long wavelength will give a more accurate momentum but a less accurate position.

Combining Eqs. (3.23) and (3.24) gives

$$\Delta x \, \Delta p \geq h \tag{3.25}$$

This result is consistent with Eq. (3.22), $\Delta x \, \Delta p \geq \hbar/2$.

Arguments like the preceding one, although superficially attractive, must be approached with caution. The argument above implies that the electron can possess a definite position and momentum at any instant and that it is the measurement process that introduces the indeterminacy in $\Delta x \, \Delta p$. On the contrary, *this indeterminacy is inherent in the nature of a moving body*. The justification for the many "derivations" of this kind is first, they show it is impossible to imagine a way around the uncertainty principle; and second, they present a view of the principle that can be appreciated in a more familiar context than that of wave groups.

3.9 APPLYING THE UNCERTAINTY PRINCIPLE

A useful tool, not just a negative statement

Planck's constant h is so small that the limitations imposed by the uncertainty principle are significant only in the realm of the atom. On such a scale, however, this principle is of great help in understanding many phenomena. It is worth keeping in mind that the lower limit of $\hbar/2$ for $\Delta x \, \Delta p$ is rarely attained. More usually $\Delta x \, \Delta p \geq \hbar$, or even (as we just saw) $\Delta x \, \Delta p \geq h$.

Example 3.7

A typical atomic nucleus is about 5.0×10^{-15} m in radius. Use the uncertainty principle to place a lower limit on the energy an electron must have if it is to be part of a nucleus.

Solution

Letting $\Delta x = 5.0 \times 10^{-5}$ m we have

$$\Delta p \geq \frac{\hbar}{2\Delta x} \geq \frac{1.054 \times 10^{-34}\ \text{J}\cdot\text{s}}{(2)(5.0 \times 10^{-15}\ \text{m})} \geq 1.1 \times 10^{-20}\ \text{kg}\cdot\text{m/s}$$

If this is the uncertainty in a nuclear electron's momentum, the momentum p itself must be at least comparable in magnitude. An electron with such a momentum has a kinetic energy KE many times greater than its rest energy mc^2. From Eq. (1.24) we see that we can let KE $= pc$ here to a sufficient degree of accuracy. Therefore

$$\text{KE} = pc \geq (1.1 \times 10^{-20}\ \text{kg}\cdot\text{m/s})(3.0 \times 10^{8}\ \text{m/s}) \geq 3.3 \times 10^{-12}\ \text{J}$$

Since 1 eV $= 1.6 \times 10^{-19}$ J, the kinetic energy of an electron must exceed 20 MeV if it is to be inside a nucleus. Experiments show that the electrons emitted by certain unstable nuclei never have more than a small fraction of this energy, from which we conclude that nuclei cannot contain electrons. The electron an unstable nucleus may emit comes into being at the moment the nucleus decays (see Secs. 11.3 and 12.5).

Example 3.8

A hydrogen atom is 5.3×10^{-11} m in radius. Use the uncertainty principle to estimate the minimum energy an electron can have in this atom.

Solution

Here we find that with $\Delta x = 5.3 \times 10^{-11}$ m.

$$\Delta p \geq \frac{\hbar}{2\Delta x} \geq 9.9 \times 10^{-25}\ \text{kg}\cdot\text{m/s}$$

An electron whose momentum is of this order of magnitude behaves like a classical particle, and its kinetic energy is

$$\text{KE} = \frac{p^2}{2m} \geq \frac{(9.9 \times 10^{-25}\ \text{kg}\cdot\text{m/s})^2}{(2)(9.1 \times 10^{-31}\ \text{kg})} \geq 5.4 \times 10^{-19}\ \text{J}$$

which is 3.4 eV. The kinetic energy of an electron in the lowest energy level of a hydrogen atom is actually 13.6 eV.

Energy and Time

Another form of the uncertainty principle concerns energy and time. We might wish to measure the energy E emitted during the time interval Δt in an atomic process. If the energy is in the form of em waves, the limited time available restricts the accuracy with which we can determine the frequency ν of the waves. Let us assume that the minimum uncertainty in the number of waves we count in a wave group is one wave. Since the frequency of the waves under study is equal to the number of them we count divided by the time interval, the uncertainty $\Delta\nu$ in our frequency measurement is

$$\Delta\nu \geq \frac{1}{\Delta t}$$

The corresponding energy uncertainty is

$$\Delta E = h\,\Delta\nu$$

and so

$$\Delta E \geq \frac{h}{\Delta t} \qquad \text{or} \qquad \Delta E\,\Delta t \geq h$$

A more precise calculation based on the nature of wave groups changes this result to

Uncertainties in energy and time

$$\Delta E\,\Delta t \geq \frac{\hbar}{2} \tag{3.26}$$

Equation (3.26) states that the product of the uncertainty ΔE in an energy measurement and the uncertainty Δt in the time at which the measurement is made is equal to or greater than $\hbar/2$. This result can be derived in other ways as well and is a general one not limited to em waves.

Example 3.9

An "excited" atom gives up its excess energy by emitting a photon of characteristic frequency, as described in Chap. 4. The average period that elapses between the excitation of an atom and the time it radiates is 1.0×10^{-8} s. Find the inherent uncertainty in the frequency of the photon.

Solution

The photon energy is uncertain by the amount

$$\Delta E \geq \frac{\hbar}{2\Delta t} \geq \frac{1.054 \times 10^{-34}\ \text{J}\cdot\text{s}}{2(1.0 \times 10^{-8}\ \text{s})} \geq 5.3 \times 10^{-27}\ \text{J}$$

The corresponding uncertainty in the frequency of light is

$$\Delta\nu = \frac{\Delta E}{h} \geq 8 \times 10^{6}\ \text{Hz}$$

This is the irreducible limit to the accuracy with which we can determine the frequency of the radiation emitted by an atom. As a result, the radiation from a group of excited atoms does not appear with the precise frequency ν. For a photon whose frequency is, say, 5.0×10^{14} Hz, $\Delta\nu/\nu = 1.6 \times 10^{-8}$. In practice, other phenomena such as the doppler effect contribute more than this to the broadening of spectral lines.

EXERCISES

It is only the first step that takes the effort. —Marquise du Deffand

3.1 De Broglie Waves

1. A photon and a particle have the same wavelength. Can anything be said about how their linear momenta compare? About how the photon's energy compares with the particle's total energy? About how the photon's energy compares with the particle's kinetic energy?

2. Find the de Broglie wavelength of (*a*) an electron whose speed is 1.0×10^8 m/s, and (*b*) an electron whose speed is 2.0×10^8 m/s.

3. Find the de Broglie wavelength of a 1.0-mg grain of sand blown by the wind at a speed of 20 m/s.

4. Find the de Broglie wavelength of the 40-keV electrons used in a certain electron microscope.

5. By what percentage will a nonrelativistic calculation of the de Broglie wavelength of a 100-keV electron be in error?

6. Find the de Broglie wavelength of a 1.00-MeV proton. Is a relativistic calculation needed?

7. The atomic spacing in rock salt, NaCl, is 0.282 nm. Find the kinetic energy (in eV) of a neutron with a de Broglie wavelength of 0.282 nm. Is a relativistic calculation needed? Such neutrons can be used to study crystal structure.

8. Find the kinetic energy of an electron whose de Broglie wavelength is the same as that of a 100-keV x-ray.

9. Green light has a wavelength of about 550 nm. Through what potential difference must an electron be accelerated to have this wavelength?

10. Show that the de Broglie wavelength of a particle of mass m and kinetic energy KE is given by

$$\lambda = \frac{hc}{\sqrt{\text{KE}(\text{KE} + 2mc^2)}}$$

11. Show that if the total energy of a moving particle greatly exceeds its rest energy, its de Broglie wavelength is nearly the same as the wavelength of a photon with the same total energy.

12. (*a*) Derive a relativistically correct formula that gives the de Broglie wavelength of a charged particle in terms of the potential difference V through which it has been accelerated. (*b*) What is the nonrelativistic approximation of this formula, valid for $eV \ll mc^2$?

3.4 Phase and Group Velocities

13. An electron and a proton have the same velocity. Compare the wavelengths and the phase and group velocities of their de Broglie waves.

14. An electron and a proton have the same kinetic energy. Compare the wavelengths and the phase and group velocities of their de Broglie waves.

15. Verify the statement in the text that, if the phase velocity is the same for all wavelengths of a certain wave phenomenon (that is, there is no dispersion), the group and phase velocities are the same.

16. The phase velocity of ripples on a liquid surface is $\sqrt{2\pi S/\lambda\rho}$, where S is the surface tension and ρ the density of the liquid. Find the group velocity of the ripples.

17. The phase velocity of ocean waves is $\sqrt{g\lambda/2\pi}$, where g is the acceleration of gravity. Find the group velocity of ocean waves.

18. Find the phase and group velocities of the de Broglie waves of an electron whose speed is $0.900c$.

19. Find the phase and group velocities of the de Broglie waves of an electron whose kinetic energy is 500 keV.

20. Show that the group velocity of a wave is given by $v_g = d\nu/d(1/\lambda)$.

21. (*a*) Show that the phase velocity of the de Broglie waves of a particle of mass m and de Broglie wavelength λ is given by

$$v_p = c\sqrt{1 + \left(\frac{mc\lambda}{h}\right)^2}$$

(*b*) Compare the phase and group velocities of an electron whose de Broglie wavelength is exactly 1×10^{-13} m.

22. In his original paper, de Broglie suggested that $E = h\nu$ and $p = h/\lambda$, which hold for electromagnetic waves, are also valid for moving particles. Use these relationships to show that the group velocity v_g of a de Broglie wave group is given by dE/dp, and with the help of Eq. (1.24), verify that $v_g = v$ for a particle of velocity v.

3.5 Particle Diffraction

23. What effect on the scattering angle in the Davisson-Germer experiment does increasing the electron energy have?

24. A beam of neutrons that emerges from a nuclear reactor contains neutrons with a variety of energies. To obtain neutrons with an energy of 0.050 eV, the beam is passed through a crystal whose atomic planes are 0.20 nm apart. At what angles relative to the original beam will the desired neutrons be diffracted?

25. In Sec. 3.5 it was mentioned that the energy of an electron entering a crystal increases, which reduces its de Broglie wavelength. Consider a beam of 54-eV electrons directed at a nickel target. The potential energy of an electron that enters the target changes by 26 eV. (*a*) Compare the electron speeds outside and inside the target. (*b*) Compare the respective de Broglie wavelengths.

26. A beam of 50-keV electrons is directed at a crystal and diffracted electrons are found at an angle of 50° relative to the original beam. What is the spacing of the atomic planes of the crystal? A relativistic calculation is needed for λ.

3.6 Particle in a Box

27. Obtain an expression for the energy levels (in MeV) of a neutron confined to a one-dimensional box 1.00×10^{-14} m wide. What is the neutron's minimum energy? (The diameter of an atomic nucleus is of this order of magnitude.)

28. The lowest energy possible for a certain particle trapped in a certain box is 1.00 eV. (*a*) What are the next two higher energies the particle can have? (*b*) If the particle is an electron, how wide is the box?

29. A proton in a one-dimensional box has an energy of 400 keV in its first excited state. How wide is the box?

3.7 Uncertainty Principle I
3.8 Uncertainty Principle II
3.9 Applying the Uncertainty Principle

30. Discuss the prohibition of $E = 0$ for a particle trapped in a box L wide in terms of the uncertainty principle. How does the minimum momentum of such a particle compare with the momentum uncertainty required by the uncertainty principle if we take $\Delta x = L$?

31. The atoms in a solid possess a certain minimum **zero-point energy** even at 0 K, while no such restriction holds for the molecules in an ideal gas. Use the uncertainty principle to explain these statements.

32. Compare the uncertainties in the velocities of an electron and a proton confined in a 1.00-nm box.

33. The position and momentum of a 1.00-keV electron are simultaneously determined. If its position is located to within 0.100 nm, what is the percentage of uncertainty in its momentum?

34. (*a*) How much time is needed to measure the kinetic energy of an electron whose speed is 10.0 m/s with an uncertainty of no more than 0.100 percent? How far will the electron have traveled in this period of time? (*b*) Make the same calculations for a 1.00-g insect whose speed is the same. What do these sets of figures indicate?

35. How accurately can the position of a proton with $v \ll c$ be determined without giving it more than 1.00 keV of kinetic energy?

36. (*a*) Find the magnitude of the momentum of a particle in a box in its nth state. (*b*) The minimum change in the particle's momentum that a measurement can cause corresponds to a change of ± 1 in the quantum number n. If $\Delta x = L$, show that $\Delta p \, \Delta x \geq \hbar/2$.

37. A marine radar operating at a frequency of 9400 MHz emits groups of electromagnetic waves 0.0800 μs in duration. The time needed for the reflections of these groups to return indicates the distance to a target. (*a*) Find the length of each group and the number of waves it contains. (*b*) What is the approximate minimum bandwidth (that is, spread of frequencies) the radar receiver must be able to process?

38. An unstable elementary particle called the eta meson has a rest mass of 549 MeV/c^2 and a mean lifetime of 7.00×10^{-19} s. What is the uncertainty in its rest mass?

39. The frequency of oscillation of a harmonic oscillator of mass m and spring constant C is $\nu = \sqrt{C/m}/2\pi$. The energy of the oscillator is $E = p^2/2m + Cx^2/2$, where p is its momentum when its displacement from the equilibrium position is x. In classical physics the minimum energy of the oscillator is $E_{\min} = 0$. Use the uncertainty principle to find an expression for E in terms of x only and show that the minimum energy is actually $E_{\min} = h\nu/2$ by setting $dE/dx = 0$ and solving for $E_{\min}$.

40. (*a*) Verify that the uncertainty principle can be expressed in the form $\Delta L \, \Delta\theta \geq \hbar/2$, where ΔL is the uncertainty in the angular momentum of a particle and $\Delta\theta$ is the uncertainty in its angular position. (*Hint*: Consider a particle of mass m moving in a circle of radius r at the speed v, for which $L = mvr$.) (*b*) At what uncertainty in L will the angular position of a particle become completely indeterminate?

CHAPTER 5

Quantum Mechanics

Scanning tunneling micrograph of gold atoms on a carbon (graphite) substrate. The cluster of gold atoms is about 1.5 nm across and three atoms high.

Although the Bohr theory of the atom, which can be extended further than was done in Chap. 4, is able to account for many aspects of atomic phenomena, it has a number of severe limitations as well. First of all, it applies only to hydrogen and one-electron ions such as He^+ and Li^{2+}—it does not even work for ordinary helium. The Bohr theory cannot explain why certain spectral lines are more intense than others (that is, why certain transitions between energy levels have greater probabilities of occurrence than others). It cannot account for the observation that many spectral lines actually consist of several separate lines whose wavelengths differ slightly. And perhaps most important, it does not permit us to obtain what a really successful theory of the atom should make possible: an understanding of how individual atoms interact with one another to endow macroscopic aggregates of matter with the physical and chemical properties we observe.

The preceding objections to the Bohr theory are not put forward in an unfriendly way, for the theory was one of those seminal achievements that transform scientific thought, but rather to emphasize that a more general approach to atomic phenomena is required. Such an approach was developed in 1925 and 1926 by Erwin Schrödinger, Werner Heisenberg, Max Born, Paul Dirac, and others under the apt name of **quantum mechanics.** "The discovery of quantum mechanics was nearly a total surprise. It described the physical world in a way that was fundamentally new. It seemed to many of us a miracle," noted Eugene Wigner, one of the early workers in the field. By the early 1930s the application of quantum mechanics to problems involving nuclei, atoms, molecules, and matter in the solid state made it possible to understand a vast body of data ("a large part of physics and the whole of chemistry," according to Dirac) and—vital for any theory—led to predictions of remarkable accuracy. Quantum mechanics has survived every experimental test thus far of even its most unexpected conclusions.

5.1 QUANTUM MECHANICS

Classical mechanics is an approximation of quantum mechanics

The fundamental difference between classical (or Newtonian) mechanics and quantum mechanics lies in what they describe. In classical mechanics, the future history of a particle is completely determined by its initial position and momentum together with the forces that act upon it. In the everyday world these quantities can all be determined well enough for the predictions of Newtonian mechanics to agree with what we find.

Quantum mechanics also arrives at relationships between observable quantities, but the uncertainty principle suggests that the nature of an observable quantity is different in the atomic realm. Cause and effect are still related in quantum mechanics, but what they concern needs careful interpretation. In quantum mechanics the kind of certainty about the future characteristic of classical mechanics is impossible because the initial state of a particle cannot be established with sufficient accuracy. As we saw in Sec. 3.7, the more we know about the position of a particle now, the less we know about its momentum and hence about its position later.

The quantities whose relationships quantum mechanics explores are *probabilities.* Instead of asserting, for example, that the radius of the electron's orbit in a ground-state hydrogen atom is always exactly 5.3×10^{-11} m, as the Bohr theory does, quantum mechanics states that this is the *most probable* radius. In a suitable experiment most trials will yield a different value, either larger or smaller, but the value most likely to be found will be 5.3×10^{-11} m.

Quantum mechanics might seem a poor substitute for classical mechanics. However, classical mechanics turns out to be just an approximate version of quantum mechanics. The certainties of classical mechanics are illusory, and their apparent agreement with experiment occurs because ordinary objects consist of so many individual atoms that departures from average behavior are unnoticeable. Instead of two sets of physical principles, one for the macroworld and one for the microworld, there is only the single set included in quantum mechanics.

Wave Function

As mentioned in Chap. 3, the quantity with which quantum mechanics is concerned is the **wave function** Ψ of a body. While Ψ itself has no physical interpretation, the square of its absolute magnitude $|\Psi|^2$ evaluated at a particular place at a particular time is proportional to the probability of finding the body there at that time. The linear momentum, angular momentum, and energy of the body are other quantities that can be established from Ψ. The problem of quantum mechanics is to determine Ψ for a body when its freedom of motion is limited by the action of external forces.

Wave functions are usually complex with both real and imaginary parts. A probability, however, must be a positive real quantity. The probability density $|\Psi|^2$ for a complex Ψ is therefore taken as the product $\Psi^*\Psi$ of Ψ and its **complex conjugate** Ψ^*. The complex conjugate of any function is obtained by replacing $i(=\sqrt{-1})$ by $-i$ wherever it appears in the function. Every complex function Ψ can be written in the form

Wave function $$\Psi = A + iB$$

where A and B are real functions. The complex conjugate Ψ^* of Ψ is

Complex conjugate $$\Psi^* = A - iB$$

and so $$|\Psi|^2 = \Psi^*\Psi = A^2 - i^2B^2 = A^2 + B^2$$

since $i^2 = -1$. Hence $|\Psi|^2 = \Psi^*\Psi$ is always a positive real quantity, as required.

Normalization

Even before we consider the actual calculation of Ψ, we can establish certain requirements it must always fulfill. For one thing, since $|\Psi|^2$ is proportional to the probability density P of finding the body described by Ψ, the integral of $|\Psi|^2$ over all space must be finite—the body is *somewhere,* after all. If

$$\int_{-\infty}^{\infty} |\Psi|^2 \, dV = 0$$

the particle does not exist, and the integral obviously cannot be ∞ and still mean anything. Furthermore, $|\Psi|^2$ cannot be negative or complex because of the way it is defined. The only possibility left is that the integral be a finite quantity if Ψ is to describe properly a real body.

It is usually convenient to have $|\Psi|^2$ be *equal* to the probability density P of finding the particle described by Ψ, rather than merely be proportional to P. If $|\Psi|^2$ is to equal P, then it must be true that

Normalization
$$\int_{-\infty}^{\infty} |\Psi|^2 \, dV = 1 \tag{5.1}$$

since if the particle exists somewhere at all times,

$$\int_{-\infty}^{\infty} P \, dV = 1$$

A wave function that obeys Eq. (5.1) is said to be **normalized.** Every acceptable wave function can be normalized by multiplying it by an appropriate constant; we shall shortly see how this is done.

Well-Behaved Wave Functions

Besides being normalizable, Ψ must be single-valued, since P can have only one value at a particular place and time, and continuous. Momentum considerations (see Sec. 5.6) require that the partial derivatives $\partial\Psi/\partial x$, $\partial\Psi/\partial y$, $\partial\Psi/\partial z$ be finite, continuous, and single-valued. Only wave functions with all these properties can yield physically meaningful results when used in calculations, so only such "well-behaved" wave functions are admissible as mathematical representations of real bodies. To summarize:

1 Ψ must be continuous and single-valued everywhere.
2 $\partial\Psi/\partial x$, $\partial\Psi/\partial y$, $\partial\Psi/\partial z$ must be continuous and single-valued everywhere.
3 Ψ must be normalizable, which means that Ψ must go to 0 as $x \to \pm\infty$, $y \to \pm\infty$, $z \to \pm\infty$ in order that $\int |\Psi|^2 \, dV$ over all space be a finite constant.

These rules are not always obeyed by the wave functions of particles in model situations that only approximate actual ones. For instance, the wave functions of a particle in a box with infinitely hard walls do not have continuous derivatives at the walls, since $\Psi = 0$ outside the box (see Fig. 5.4). But in the real world, where walls are never infinitely hard, there is no sharp change in Ψ at the walls (see Fig. 5.7) and the derivatives are continuous. Exercise 7 gives another example of a wave function that is not well-behaved.

Given a normalized and otherwise acceptable wave function Ψ, the probability that the particle it describes will be found in a certain region is simply the integral of the probability density $|\Psi|^2$ over that region. Thus for a particle restricted to motion in the x direction, the probability of finding it between x_1 and x_2 is given by

Probability
$$P_{x_1x_2} = \int_{x_1}^{x_2} |\Psi|^2 \, dx \tag{5.2}$$

We will see examples of such calculations later in this chapter and in Chap. 6.

5.2 THE WAVE EQUATION

It can have a variety of solutions, including complex ones

Schrödinger's equation, which is the fundamental equation of quantum mechanics in the same sense that the second law of motion is the fundamental equation of Newtonian mechanics, is a wave equation in the variable Ψ.

Before we tackle Schrödinger's equation, let us review the wave equation

Wave equation
$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y}{\partial t^2} \tag{5.3}$$

which governs a wave whose variable quantity is y that propagates in the x direction with the speed $\boldsymbol{v}$. In the case of a wave in a stretched string, y is the displacement of the string from the x axis; in the case of a sound wave, y is the pressure difference; in the case of a light wave, y is either the electric or the magnetic field magnitude. Equation (5.3) can be derived from the second law of motion for mechanical waves and from Maxwell's equations for electromagnetic waves.

Partial Derivatives

Suppose we have a function $f(x, y)$ of two variables, x and y, and we want to know how f varies with only one of them, say x. To find out, we differentiate f with respect to x while treating the other variable y as a constant. The result is the **partial derivative** of f with respect to x, which is written $\partial f/\partial x$

$$\frac{\partial f}{\partial x} = \left(\frac{df}{dx}\right)_{y=\text{constant}}$$

The rules for ordinary differentiation hold for partial differentiation as well. For instance, if $f = cx^2$,

$$\frac{df}{dx} = 2cx$$

and so, if $f = yx^2$,

$$\frac{\partial f}{\partial x} = \left(\frac{df}{dx}\right)_{y=\text{constant}} = 2yx$$

The partial derivative of $f = yx^2$ with respect to the other variable, y, is

$$\frac{\partial f}{\partial y} = \left(\frac{df}{dy}\right)_{x=\text{constant}} = x^2$$

Second order partial derivatives occur often in physics, as in the wave equation. To find $\partial^2 f/\partial x^2$, we first calculate $\partial f/\partial x$ and then differentiate again, still keeping y constant:

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)$$

For $f = yx^2$,

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}(2yx) = 2y$$

Similarly
$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}(x^2) = 0$$

Solutions of the wave equation may be of many kinds, reflecting the variety of waves that can occur—a single traveling pulse, a train of waves of constant amplitude and wavelength, a train of superposed waves of the same amplitudes and wavelengths, a train of superposed waves of different amplitudes and wavelengths,

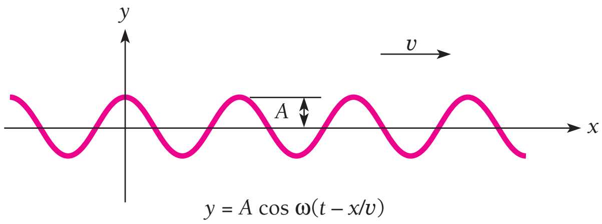


Figure 5.1 Waves in the *xy* plane traveling in the $+x$ direction along a stretched string lying on the *x* axis.

a standing wave in a string fastened at both ends, and so on. All solutions must be of the form

$$y = F\left(t \pm \frac{x}{v}\right) \tag{5.4}$$

where F is any function that can be differentiated. The solutions $F(t - x/v)$ represent waves traveling in the $+x$ direction, and the solutions $F(t + x/v)$ represent waves traveling in the $-x$ direction.

Let us consider the wave equivalent of a "free particle," which is a particle that is not under the influence of any forces and therefore pursues a straight path at constant speed. This wave is described by the general solution of Eq. (5.3) for undamped (that is, constant amplitude A), monochromatic (constant angular frequency ω) harmonic waves in the $+x$ direction, namely

$$y = Ae^{-i\omega(t-x/v)} \tag{5.5}$$

In this formula y is a complex quantity, with both real and imaginary parts.

Because

$$e^{-i\theta} = \cos\theta - i\sin\theta$$

Eq. (5.5) can be written in the form

$$y = A\cos\omega\left(t - \frac{x}{v}\right) - iA\sin\omega\left(t - \frac{x}{v}\right) \tag{5.6}$$

Only the real part of Eq. (5.6) [which is the same as Eq. (3.5)] has significance in the case of waves in a stretched string. There y represents the displacement of the string from its normal position (Fig. 5.1), and the imaginary part of Eq. (5.6) is discarded as irrelevant.

Example 5.1

Verify that Eq. (5.5) is a solution of the wave equation.

Solution

The derivative of an exponential function e^u is

$$\frac{d}{dx}(e^u) = e^u\frac{du}{dx}$$

The partial derivative of y with respect to x (which means t is treated as a constant) from Eq. (5.5) is therefore

$$\frac{\partial y}{\partial x} = \frac{i\omega}{v}y$$

and the second partial derivative is

$$\frac{\partial^2 y}{\partial x^2} = \frac{i^2\omega^2}{v^2}y = -\frac{\omega^2}{v^2}y$$

since $i^2 = -1$. The partial derivative of y with respect to t (now holding x constant) is

$$\frac{\partial y}{\partial t} = -i\omega y$$

and the second partial derivative is

$$\frac{\partial^2 y}{\partial t^2} = i^2\omega^2 y = -\omega^2 y$$

Combining these results gives

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y}{\partial t^2}$$

which is Eq. (5.3). Hence Eq. (5.5) is a solution of the wave equation.

5.3 SCHRÖDINGER'S EQUATION: TIME-DEPENDENT FORM

A basic physical principle that cannot be derived from anything else

In quantum mechanics the wave function Ψ corresponds to the wave variable y of wave motion in general. However, Ψ, unlike y, is not itself a measurable quantity and may therefore be complex. For this reason we assume that Ψ for a particle moving freely in the $+x$ direction is specified by

$$\Psi = Ae^{-i\omega(t-x/v)} \tag{5.7}$$

Replacing ω in the above formula by $2\pi\nu$ and v by $\lambda\nu$ gives

$$\Psi = Ae^{-2\pi i(\nu t - x/\lambda)} \tag{5.8}$$

This is convenient since we already know what ν and λ are in terms of the total energy E and momentum p of the particle being described by Ψ. Because

$$E = h\nu = 2\pi\hbar\nu \qquad \text{and} \qquad \lambda = \frac{h}{p} = \frac{2\pi\hbar}{p}$$

we have

Free particle

$$\Psi = Ae^{-(i/\hbar)(Et-px)} \tag{5.9}$$

Equation (5.9) describes the wave equivalent of an unrestricted particle of total energy E and momentum p moving in the $+x$ direction, just as Eq. (5.5) describes, for example, a harmonic displacement wave moving freely along a stretched string.

The expression for the wave function Ψ given by Eq. (5.9) is correct only for freely moving particles. However, we are most interested in situations where the motion of a particle is subject to various restrictions. An important concern, for example, is an electron bound to an atom by the electric field of its nucleus. What we must now do is obtain the fundamental differential equation for Ψ, which we can then solve for Ψ in a specific situation. This equation, which is Schrödinger's equation, can be arrived at in various ways, but it *cannot* be rigorously derived from existing physical principles:

the equation represents something new. What will be done here is to show one route to the wave equation for Ψ and then to discuss the significance of the result.

We begin by differentiating Eq. (5.9) for Ψ twice with respect to x, which gives

$$\frac{\partial^2\Psi}{\partial x^2} = -\frac{p^2}{\hbar^2}\Psi$$

$$p^2\Psi = -\hbar^2\frac{\partial^2\Psi}{\partial x^2} \tag{5.10}$$

Differentiating Eq. (5.9) once with respect to t gives

$$\frac{\partial\Psi}{\partial t} = -\frac{iE}{\hbar}\Psi$$

$$E\Psi = -\frac{\hbar}{i}\frac{\partial\Psi}{\partial t} \tag{5.11}$$

At speeds small compared with that of light, the total energy E of a particle is the sum of its kinetic energy $p^2/2m$ and its potential energy U, where U is in general a function of position x and time t:

$$E = \frac{p^2}{2m} + U(x, t) \tag{5.12}$$

The function U represents the influence of the rest of the universe on the particle. Of course, only a small part of the universe interacts with the particle to any extent; for

Erwin Schrödinger (1887–1961) was born in Vienna to an Austrian father and a half-English mother and received his doctorate at the university there. After World War I, during which he served as an artillery officer, Schrödinger had appointments at several German universities before becoming professor of physics in Zurich, Switzerland. Late in November, 1925, Schrödinger gave a talk on de Broglie's notion that a moving particle has a wave character. A colleague remarked to him afterward that to deal properly with a wave, one needs a wave equation. Schrödinger took this to heart, and a few weeks later he was "struggling with a new atomic theory. If only I knew more mathematics! I am very optimistic about this thing and expect that if I can only . . . solve it, it will be *very* beautiful." (Schrödinger was not the only physicist to find the mathematics he needed difficult; the eminent mathematician David Hilbert said at about this time, "Physics is much too hard for physicists.")

The struggle was successful, and in January 1926 the first of four papers on "Quantization as an Eigenvalue Problem" was completed. In this epochal paper Schrödinger introduced the equation that bears his name and solved it for the hydrogen atom, thereby opening wide the door to the modern view of the atom which others had only pushed ajar. By June Schrödinger had applied wave mechanics to the harmonic oscillator, the diatomic molecule, the hydrogen atom in an electric field, the absorption and emission of radiation, and the scattering of radiation by atoms and molecules. He had also shown that his wave mechanics was mathematically equivalent to the more abstract Heisenberg-Born-Jordan matrix mechanics.

The significance of Schrödinger's work was at once realized. In 1927 he succeeded Planck at the University of Berlin but left Germany in 1933, the year he received the Nobel Prize, when the Nazis came to power. He was at Dublin's Institute for Advanced Study from 1939 until his return to Austria in 1956. In Dublin, Schrödinger became interested in biology, in particular the mechanism of heredity. He seems to have been the first to make definite the idea of a genetic code and to identify genes as long molecules that carry the code in the form of variations in how their atoms are arranged. Schrödinger's 1944 book *What Is Life?* was enormously influential, not only by what it said but also by introducing biologists to a new way of thinking—that of the physicist—about their subject. *What Is Life?* started James Watson on his search for "the secret of the gene," which he and Francis Crick (a physicist) discovered in 1953 to be the structure of the DNA molecule.

instance, in the case of the electron in a hydrogen atom, only the electric field of the nucleus must be taken into account.

Multiplying both sides of Eq. (5.12) by the wave function Ψ gives

$$E\Psi = \frac{p^2\Psi}{2m} + U\Psi \tag{5.13}$$

Now we substitute for $E\Psi$ and $p^2\Psi$ from Eqs. (5.10) and (5.11) to obtain the **time-dependent form of Schrödinger's equation:**

Time-dependent Schrödinger equation in one dimension

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + U\Psi \tag{5.14}$$

In three dimensions the time-dependent form of Schrödinger's equation is

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\left(\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2}\right) + U\Psi \tag{5.15}$$

where the particle's potential energy U is some function of x, y, z, and t.

Any restrictions that may be present on the particle's motion will affect the potential-energy function U. Once U is known, Schrödinger's equation may be solved for the wave function Ψ of the particle, from which its probability density $|\Psi|^2$ may be determined for a specified x, y, z, t.

Validity of Schrödinger's Equation

Schrödinger's equation was obtained here using the wave function of a freely moving particle (potential energy U = constant). How can we be sure it applies to the general case of a particle subject to arbitrary forces that vary in space and time [$U = U(x, y, z, t)$]? Substituting Eqs. (5.10) and (5.11) into Eq. (5.13) is really a wild leap with no formal justification; this is true for all other ways in which Schrödinger's equation can be arrived at, including Schrödinger's own approach.

What we must do is postulate Schrödinger's equation, solve it for a variety of physical situations, and compare the results of the calculations with the results of experiments. If both sets of results agree, the postulate embodied in Schrödinger's equation is valid. If they disagree, the postulate must be discarded and some other approach would then have to be explored. In other words,

> Schrödinger's equation cannot be derived from other basic principles of physics; it is a basic principle in itself.

What has happened is that Schrödinger's equation has turned out to be remarkably accurate in predicting the results of experiments. To be sure, Eq. (5.15) can be used only for nonrelativistic problems, and a more elaborate formulation is needed when particle speeds near that of light are involved. But because it is in accord with experience within its range of applicability, we must consider Schrödinger's equation as a valid statement concerning certain aspects of the physical world.

It is worth noting that Schrödinger's equation does not increase the number of principles needed to describe the workings of the physical world. Newton's second law of motion $F = ma$, the basic principle of classical mechanics, can be derived from Schrödinger's equation provided the quantities it relates are understood to be averages rather than precise values. (Newton's laws of motion were also not derived from any other principles. Like Schrödinger's equation, these laws are considered valid in their range of applicability because of their agreement with experiment.)

5.4 LINEARITY AND SUPERPOSITION

Wave functions add, not probabilities

An important property of Schrödinger's equation is that it is linear in the wave function Ψ. By this is meant that the equation has terms that contain Ψ and its derivatives but no terms independent of Ψ or that involve higher powers of Ψ or its derivatives. As a result, a linear combination of solutions of Schrödinger's equation for a given system is also itself a solution. If Ψ_1 and Ψ_2 are two solutions (that is, two wave functions that satisfy the equation), then

$$\Psi = a_1\Psi_1 + a_2\Psi_2$$

is also a solution, where a_1 and a_2 are constants (see Exercise 8). Thus the wave functions Ψ_1 and Ψ_2 obey the superposition principle that other waves do (see Sec. 2.1) and we conclude that interference effects can occur for wave functions just as they can for light, sound, water, and electromagnetic waves. In fact, the discussions of Secs. 3.4 and 3.7 assumed that de Broglie waves are subject to the superposition principle.

Let us apply the superposition principle to the diffraction of an electron beam. Figure 5.2*a* shows a pair of slits through which a parallel beam of monoenergetic electrons pass on their way to a viewing screen. If slit 1 only is open, the result is the intensity variation shown in Fig. 5.2*b* that corresponds to the probability density

$$P_1 = |\Psi_1|^2 = \Psi_1^*\Psi_1$$

If slit 2 only is open, as in Fig. 5.2*c*, the corresponding probability density is

$$P_2 = |\Psi_2|^2 = \Psi_2^*\Psi_2$$

We might suppose that opening both slits would give an electron intensity variation described by $P_1 + P_2$, as in Fig. 5.2*d*. However, this is not the case because in quantum

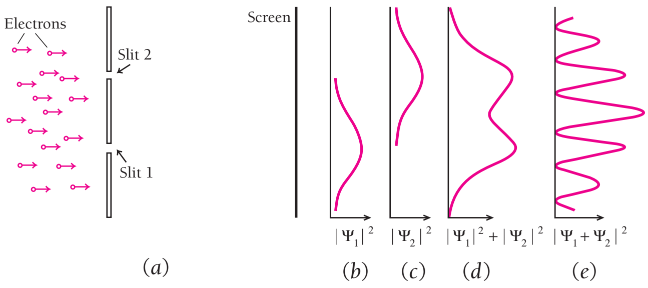


Figure 5.2 (*a*) Arrangement of double-slit experiment. (*b*) The electron intensity at the screen with only slit 1 open. (*c*) The electron intensity at the screen with only slit 2 open. (*d*) The sum of the intensities of (*b*) and (*c*). (*e*) The actual intensity at the screen with slits 1 and 2 both open. The wave functions Ψ_1 and Ψ_2 add to produce the intensity at the screen, not the probability densities $|\Psi_1|^2$ and $|\Psi_2|^2$.

mechanics wave functions add, *not* probabilities. Instead the result with both slits open is as shown in Fig. 5.2*e*, the same pattern of alternating maxima and minima that occurs when a beam of monochromatic light passes through the double slit of Fig. 2.4.

The diffraction pattern of Fig. 5.2*e* arises from the superposition Ψ of the wave functions Ψ_1 and Ψ_2 of the electrons that have passed through slits 1 and 2:

$$\Psi = \Psi_1 + \Psi_2$$

The probability density at the screen is therefore

$$\begin{aligned} P = |\Psi|^2 &= |\Psi_1 + \Psi_2|^2 = (\Psi_1^* + \Psi_2^*)(\Psi_1 + \Psi_2) \\ &= \Psi_1^*\Psi_1 + \Psi_2^*\Psi_2 + \Psi_1^*\Psi_2 + \Psi_2^*\Psi_1 \\ &= P_1 + P_2 + \Psi_1^*\Psi_2 + \Psi_2^*\Psi_1 \end{aligned}$$

The two terms at the right of this equation represent the difference between Fig. 5.2*d* and *e* and are responsible for the oscillations of the electron intensity at the screen. In Sec. 6.8 a similar calculation will be used to investigate why a hydrogen atom emits radiation when it undergoes a transition from one quantum state to another of lower energy.

5.5 EXPECTATION VALUES

How to extract information from a wave function

Once Schrödinger's equation has been solved for a particle in a given physical situation, the resulting wave function $\Psi(x, y, z, t)$ contains all the information about the particle that is permitted by the uncertainty principle. Except for those variables that are quantized this information is in the form of probabilities and not specific numbers.

As an example, let us calculate the **expectation value** $\langle x \rangle$ of the position of a particle confined to the x axis that is described by the wave function $\Psi(x, t)$. This is the value of x we would obtain if we measured the positions of a great many particles described by the same wave function at some instant t and then averaged the results.

To make the procedure clear, we first answer a slightly different question: What is the average position $\bar{x}$ of a number of identical particles distributed along the x axis in such a way that there are N_1 particles at x_1, N_2 particles at x_2, and so on? The average position in this case is the same as the center of mass of the distribution, and so

$$\bar{x} = \frac{N_1x_1 + N_2x_2 + N_3x_3 + \cdots}{N_1 + N_2 + N_3 + \cdots} = \frac{\Sigma N_i x_i}{\Sigma N_i} \tag{5.16}$$

When we are dealing with a single particle, we must replace the number N_i of particles at x_i by the probability P_i that the particle be found in an interval dx at x_i. This probability is

$$P_i = |\Psi_i|^2 \, dx \tag{5.17}$$

where Ψ_i is the particle wave function evaluated at $x = x_i$. Making this substitution and changing the summations to integrals, we see that the expectation value of the position of the single particle is

$$\langle x \rangle = \frac{\int_{-\infty}^{\infty} x|\Psi|^2 \, dx}{\int_{-\infty}^{\infty} |\Psi|^2 \, dx} \tag{5.18}$$

If Ψ is a normalized wave function, the denominator of Eq. (5.18) equals the probability that the particle exists somewhere between $x = -\infty$ and $x = \infty$ and therefore has the value 1. In this case

Expectation value for position

$$\langle x \rangle = \int_{-\infty}^{\infty} x|\Psi|^2 \, dx \tag{5.19}$$

Example 5.2

A particle limited to the x axis has the wave function $\Psi = ax$ between $x = 0$ and $x = 1$; $\Psi = 0$ elsewhere. (*a*) Find the probability that the particle can be found between $x = 0.45$ and $x = 0.55$. (*b*) Find the expectation value $\langle x \rangle$ of the particle's position.

Solution

(*a*) The probability is

$$\int_{x_1}^{x_2} |\Psi|^2 \, dx = a^2 \int_{0.45}^{0.55} x^2 dx = a^2 \left[\frac{x^3}{3} \right]_{0.45}^{0.55} = 0.0251a^2$$

(*b*) The expectation value is

$$\langle x \rangle = \int_0^1 x|\Psi|^2 \, dx = a^2 \int_0^1 x^3 dx = a^2 \left[\frac{x^4}{4} \right]_0^1 = \frac{a^2}{4}$$

The same procedure as that followed above can be used to obtain the expectation value $\langle G(x) \rangle$ of any quantity—for instance, potential energy $U(x)$—that is a function of the position x of a particle described by a wave function Ψ. The result is

Expectation value

$$\langle G(x) \rangle = \int_{-\infty}^{\infty} G(x)|\Psi|^2 \, dx \tag{5.20}$$

The expectation value $\langle p \rangle$ for momentum cannot be calculated this way because, according to the uncertainty principles, no such function as $p(x)$ can exist. If we specify x, so that $\Delta x = 0$, we cannot specify a corresponding p since $\Delta x \, \Delta p \geq \hbar/2$. The same problem occurs for the expectation value $\langle E \rangle$ for energy because $\Delta E \Delta t \geq \hbar/2$ means that, if we specify t, the function $E(t)$ is impossible. In Sec. 5.6 we will see how $\langle p \rangle$ and $\langle E \rangle$ can be determined.

In classical physics no such limitation occurs, because the uncertainty principle can be neglected in the macroworld. When we apply the second law of motion to the motion of a body subject to various forces, we expect to get $p(x, t)$ and $E(x, t)$ from the solution as well as $x(t)$. Solving a problem in classical mechanics gives us the entire future course of the body's motion. In quantum physics, on the other hand, all we get directly by applying Schrödinger's equation to the motion of a particle is the wave function Ψ, and the future course of the particle's motion—like its initial state—is a matter of probabilities instead of certainties.

5.6 OPERATORS

Another way to find expectation values

A hint as to the proper way to evaluate $\langle p \rangle$ and $\langle E \rangle$ comes from differentiating the free-particle wave function $\Psi = Ae^{-(i/\hbar)(Et-px)}$ with respect to x and to t. We find that

$$\frac{\partial \Psi}{\partial x} = \frac{i}{\hbar} p \Psi$$

$$\frac{\partial \Psi}{\partial t} = -\frac{i}{\hbar} E \Psi$$

which can be written in the suggestive forms

$$p\Psi = \frac{\hbar}{i} \frac{\partial}{\partial x} \Psi \tag{5.21}$$

$$E\Psi = i\hbar \frac{\partial}{\partial t} \Psi \tag{5.22}$$

Evidently the dynamical quantity p in some sense corresponds to the differential operator $(\hbar/i)\ \partial/\partial x$ and the dynamical quantity E similarly corresponds to the differential operator $i\hbar\ \partial/\partial t$.

An **operator** tells us what operation to carry out on the quantity that follows it. Thus the operator $i\hbar\ \partial/\partial t$ instructs us to take the partial derivative of what comes after it with respect to t and multiply the result by $i\hbar$. Equation (5.22) was on the postmark used to cancel the Austrian postage stamp issued to commemorate the 100th anniversary of Schrödinger's birth.

It is customary to denote operators by using a caret, so that $\hat{p}$ is the operator that corresponds to momentum p and $\hat{E}$ is the operator that corresponds to total energy E. From Eqs. (5.21) and (5.22) these operators are

Momentum operator

$$\hat{p} = \frac{\hbar}{i} \frac{\partial}{\partial x} \tag{5.23}$$

Total-energy operator

$$\hat{E} = i\hbar \frac{\partial}{\partial t} \tag{5.24}$$

Though we have only shown that the correspondences expressed in Eqs. (5.23) and (5.24) hold for free particles, they are entirely general results whose validity is the same as that of Schrödinger's equation. To support this statement, we can replace the equation $E = \mathrm{KE} + U$ for the total energy of a particle with the operator equation

$$\hat{E} = \widehat{\mathrm{KE}} + \hat{U} \tag{5.25}$$

The operator $\hat{U}$ is just $U(\Psi)$. The kinetic energy KE is given in terms of momentum p by

$$\mathrm{KE} = \frac{p^2}{2m}$$

and so we have

Kinetic-energy operator

$$\hat{KE} = \frac{\hat{p}^2}{2m} = \frac{1}{2m}\left(\frac{\hbar}{i}\frac{\partial}{\partial x}\right)^2 = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} \tag{5.26}$$

Equation (5.25) therefore reads

$$i\hbar\frac{\partial}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + U \tag{5.27}$$

Now we multiply the identity $\Psi = \Psi$ by Eq. (5.27) and obtain

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + U\Psi$$

which is Schrödinger's equation. Postulating Eqs. (5.23) and (5.24) is equivalent to postulating Schrödinger's equation.

Operators and Expectation Values

Because p and E can be replaced by their corresponding operators in an equation, we can use these operators to obtain expectation values for p and E. Thus the expectation value for p is

$$\langle p\rangle = \int_{-\infty}^{\infty}\Psi^*\hat{p}\Psi\,dx = \int_{-\infty}^{\infty}\Psi^*\left(\frac{\hbar}{i}\frac{\partial}{\partial x}\right)\Psi\,dx = \frac{\hbar}{i}\int_{-\infty}^{\infty}\Psi^*\frac{\partial\Psi}{\partial x}\,dx \tag{5.28}$$

and the expectation value for E is

$$\langle E\rangle = \int^{\infty}\Psi^*\hat{E}\Psi\,dx = \int^{\infty}\Psi^*\left(i\hbar\frac{\partial}{\partial t}\right)\Psi\,dx = i\hbar\int^{\infty}\Psi^*\frac{\partial\Psi}{\partial t}\,dx \tag{5.29}$$

Both Eqs. (5.28) and (5.29) can be evaluated for any acceptable wave function $\Psi(x, t)$.

Let us see why expectation values involving operators have to be expressed in the form

$$\langle p\rangle = \int_{-\infty}^{\infty}\Psi^*\hat{p}\Psi\,dx$$

The other alternatives are

$$\int_{-\infty}^{\infty}\hat{p}\Psi^*\Psi\,dx = \frac{\hbar}{i}\int_{-\infty}^{\infty}\frac{\partial}{\partial x}(\Psi^*\Psi)\,dx = \frac{\hbar}{i}\Big[\Psi^*\Psi\Big]_{-\infty}^{\infty} = 0$$

since Ψ^* and Ψ must be 0 at $x = \pm\infty$, and

$$\int_{-\infty}^{\infty}\Psi^*\Psi\hat{p}\,dx = \frac{\hbar}{i}\int_{-\infty}^{\infty}\Psi^*\Psi\frac{\partial}{\partial x}\,dx$$

which makes no sense. In the case of algebraic quantities such as x and $V(x)$, the order of factors in the integrand is unimportant, but when differential operators are involved, the correct order of factors must be observed.

Every observable quantity G characteristic of a physical system may be represented by a suitable quantum-mechanical operator $\hat{G}$. To obtain this operator, we express G in terms of x and p and then replace p by $(\hbar/i)\ \partial/\partial x$. If the wave function Ψ of the system is known, the expectation value of $G(x, p)$ is

Expectation value of an operator

$$\langle G(x, p)\rangle = \int_{-\infty}^{\infty} \Psi^* \hat{G} \Psi \, dx \tag{5.30}$$

In this way all the information about a system that is permitted by the uncertainty principle can be obtained from its wave function Ψ.

5.7 SCHRÖDINGER'S EQUATION: STEADY-STATE FORM

Eigenvalues and eigenfunctions

In a great many situations the potential energy of a particle does not depend on time explicitly; the forces that act on it, and hence U, vary with the position of the particle only. When this is true, Schrödinger's equation may be simplified by removing all reference to t.

We begin by noting that the one-dimensional wave function Ψ of an unrestricted particle may be written

$$\Psi = Ae^{-(i/\hbar)(Et-px)} = Ae^{-(iE/\hbar)t}e^{+(ip/\hbar)x} = \psi e^{-(iE/\hbar)t} \tag{5.31}$$

Evidently Ψ is the product of a time-dependent function $e^{-(iE/\hbar)t}$ and a position-dependent function ψ. As it happens, the time variations of *all* wave functions of particles acted on by forces independent of time have the same form as that of an unrestricted particle. Substituting the Ψ of Eq. (5.31) into the time-dependent form of Schrödinger's equation, we find that

$$E\psi e^{-(iE/\hbar)t} = -\frac{\hbar^2}{2m} e^{-(iE/\hbar)t} \frac{\partial^2 \psi}{\partial x^2} + U\psi e^{-(iE/\hbar)t}$$

Dividing through by the common exponential factor gives

Steady-state Schrödinger equation in one dimension

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{2m}{\hbar^2}(E - U)\psi = 0 \tag{5.32}$$

Equation (5.32) is the **steady-state form of Schrödinger's equation.** In three dimensions it is

Steady-state Schrödinger equation in three dimensions

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} + \frac{2m}{\hbar^2}(E - U)\psi = 0 \tag{5.33}$$

An important property of Schrödinger's steady-state equation is that, if it has one or more solutions for a given system, each of these wave functions corresponds to a specific value of the energy E. Thus energy quantization appears in wave mechanics as a natural element of the theory, and energy quantization in the physical world is revealed as a universal phenomenon characteristic of *all* stable systems.

A familiar and quite close analogy to the manner in which energy quantization occurs in solutions of Schrödinger's equation is with standing waves in a stretched string of length L that is fixed at both ends. Here, instead of a single wave propagating indefinitely in one direction, waves are traveling in both the $+x$ and $-x$ directions simultaneously. These waves are subject to the condition (called a **boundary condition**) that the displacement y always be zero at both ends of the string. An acceptable function $y(x, t)$ for the displacement must, with its derivatives (except at the ends), be as well-behaved as ψ and its derivatives—that is, be continuous, finite, and single-valued. In this case y must be real, not complex, as it represents a directly measurable quantity. The only solutions of the wave equation, Eq. (5.3), that are in accord with these various limitations are those in which the wavelengths are given by

$$\lambda_n = \frac{2L}{n+1} \qquad n = 0, 1, 2, 3, \ldots$$

as shown in Fig. 5.3. It is the *combination* of the wave equation and the restrictions placed on the nature of its solution that leads us to conclude that $y(x, t)$ can exist only for certain wavelengths λ_n.

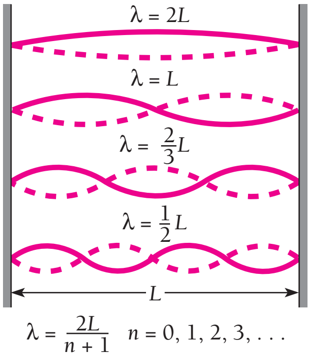


Figure 5.3 Standing waves in a stretched string fastened at both ends.

Eigenvalues and Eigenfunctions

The values of energy E_n for which Schrödinger's steady-state equation can be solved are called **eigenvalues** and the corresponding wave functions ψ_n are called **eigenfunctions.** (These terms come from the German *Eigenwert,* meaning "proper or characteristic value," and *Eigenfunktion,* "proper or characteristic function.") The discrete energy levels of the hydrogen atom

$$E_n = -\frac{me^4}{32\pi^2\epsilon_0^2\hbar^2}\left(\frac{1}{n^2}\right) \qquad n = 1, 2, 3, \ldots$$

are an example of a set of eigenvalues. We shall see in Chap. 6 why these particular values of E are the only ones that yield acceptable wave functions for the electron in the hydrogen atom.

An important example of a dynamical variable other than total energy that is found to be quantized in stable systems is angular momentum **L**. In the case of the hydrogen atom, we shall find that the eigenvalues of the magnitude of the total angular momentum are specified by

$$L = \sqrt{l(l+1)}\,\hbar \qquad l = 0, 1, 2, \ldots, (n-1)$$

Of course, a dynamical variable G may not be quantized. In this case measurements of G made on a number of identical systems will not yield a unique result but instead a spread of values whose average is the expectation value

$$\langle G\rangle = \int_{-\infty}^{\infty} G|\psi|^2\,dx$$

In the hydrogen atom, the electron's position is not quantized, for instance, so that we must think of the electron as being present in the vicinity of the nucleus with a certain probability $|\psi|^2$ per unit volume but with no predictable position or even orbit in the classical sense. This probabilistic statement does not conflict with the fact that experiments performed on hydrogen atoms always show that each one contains a whole electron, not 27 percent of an electron in a certain region and 73 percent elsewhere. The probability is one of *finding* the electron, and although this probability is smeared out in space, the electron itself is not.

Operators and Eigenvalues

The condition that a certain dynamical variable G be restricted to the discrete values G_n—in other words, that G be quantized—is that the wave functions ψ_n of the system be such that

Eigenvalue equation

$$\hat{G}\psi_n = G_n\psi_n \tag{5.34}$$

where $\hat{G}$ is the operator that corresponds to G and each G_n is a real number. When Eq. (5.34) holds for the wave functions of a system, it is a fundamental postulate of quantum mechanics that any measurement of G can only yield one of the values G_n. If measurements of G are made on a number of identical systems all in states described by the particular eigenfunction ψ_k, each measurement will yield the single value G_k.

Example 5.3

An eigenfunction of the operator d^2/dx^2 is $\psi = e^{2x}$. Find the corresponding eigenvalue.

Solution

Here $\hat{G} = d^2/dx^2$, so

$$\hat{G}\psi = \frac{d^2}{dx^2}(e^{2x}) = \frac{d}{dx}\left[\frac{d}{dx^2}(e^{2x})\right] = \frac{d}{dx}(2e^{2x}) = 4e^{2x}$$

But $e^{2x} = \psi$, so

$$\hat{G}\psi = 4\psi$$

From Eq. (5.34) we see that the eigenvalue G here is just $G = 4$.

In view of Eqs. (5.25) and (5.26) the total-energy operator $\hat{E}$ of Eq. (5.24) can also be written as

Hamiltonian operator

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + U \tag{5.35}$$

and is called the **Hamiltonian operator** because it is reminiscent of the Hamiltonian function in advanced classical mechanics, which is an expression for the total energy of a system in terms of coordinates and momenta only. Evidently the steady-state Schrödinger equation can be written simply as

Schrödinger's equation

$$\hat{H}\psi_{\mathrm{n}} = E_n\psi_n \tag{5.36}$$

Table 5.1 Operators Associated with Various Observable Quantities

Quantity	Operator
Position, x	x
Linear momentum, p	$\frac{\hbar}{i}\frac{\partial}{\partial x}$
Potential energy, $U(x)$	$U(x)$
Kinetic energy, KE $= \frac{p^2}{2m}$	$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$
Total energy, E	$i\hbar\frac{\partial}{\partial t}$
Total energy (Hamiltonian form), H	$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + U(x)$

so we can say that the various E_n are the eigenvalues of the Hamiltonian operator $\hat{H}$. This kind of association between eigenvalues and quantum-mechanical operators is quite general. Table 5.1 lists the operators that correspond to various observable quantities.

5.8 PARTICLE IN A BOX

How boundary conditions and normalization determine wave functions

To solve Schrödinger's equation, even in its simpler steady-state form, usually requires elaborate mathematical techniques. For this reason the study of quantum mechanics has traditionally been reserved for advanced students who have the required proficiency in mathematics. However, since quantum mechanics is the theoretical structure whose results are closest to experimental reality, we must explore its methods and applications to understand modern physics. As we shall see, even a modest mathematical background is enough for us to follow the trains of thought that have led quantum mechanics to its greatest achievements.

The simplest quantum-mechanical problem is that of a particle trapped in a box with infinitely hard walls. In Sec. 3.6 we saw how a quite simple argument yields the energy levels of the system. Let us now tackle the same problem in a more formal way, which will give us the wave function ψ_n that corresponds to each energy level.

We may specify the particle's motion by saying that it is restricted to traveling along the x axis between $x = 0$ and $x = L$ by infintely hard walls. A particle does not lose energy when it collides with such walls, so that its total energy stays constant. From a formal point of view the potential energy U of the particle is infinite on both sides of the box, while U is a constant—say 0 for convenience—on the inside (Fig. 5.4). Because the particle cannot have an infinite amount of energy, it cannot exist outside the box, and so its wave function ψ is 0 for $x \leq 0$ and $x \geq L$. Our task is to find what ψ is within the box, namely, between $x = 0$ and $x = L$.

Within the box Schrödinger's equation becomes

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}E\psi = 0 \tag{5.37}$$

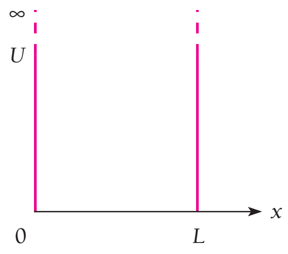


Figure 5.4 A square potential well with infinitely high barriers at each end corresponds to a box with infinitely hard walls.

since $U = 0$ there. (The total derivative $d^2\psi/dx^2$ is the same as the partial derivative $\partial^2\psi/\partial x^2$ because ψ is a function only of x in this problem.) Equation (5.37) has the solution

$$\psi = A \sin \frac{\sqrt{2mE}}{\hbar} x + B \cos \frac{\sqrt{2mE}}{\hbar} x \tag{5.38}$$

which we can verify by substitution back into Eq. (5.37). A and B are constants to be evaluated.

This solution is subject to the boundary conditions that $\psi = 0$ for $x = 0$ and for $x = L$. Since $\cos 0 = 1$, the second term cannot describe the particle because it does not vanish at $x = 0$. Hence we conclude that $B = 0$. Since $\sin 0 = 0$, the sine term always yields $\psi = 0$ at $x = 0$, as required, but ψ will be 0 at $x = L$ only when

$$\frac{\sqrt{2mE}}{\hbar} L = n\pi \qquad n = 1, 2, 3, \ldots \tag{5.39}$$

This result comes about because the sines of the angles $\pi, 2\pi, 3\pi, \ldots$ are all 0.

From Eq. (5.39) it is clear that the energy of the particle can have only certain values, which are the eigenvalues mentioned in the previous section. These eigenvalues, constituting the **energy levels** of the system, are found by solving Eq. (5.39) for E_n, which gives

Particle in a box

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2} \qquad n = 1, 2, 3, \ldots \tag{5.40}$$

Equation (5.40) is the same as Eq. (3.18) and has the same interpretation [see the discussion that follows Eq. (3.18) in Sec. 3.6].

Wave Functions

The wave functions of a particle in a box whose energies are E_n are, from Eq. (5.38) with $B = 0$,

$$\psi_n = A \sin \frac{\sqrt{2mE_n}}{\hbar} x \tag{5.41}$$

Substituting Eq. (5.40) for E_n gives

$$\psi_n = A \sin \frac{n\pi x}{L} \tag{5.42}$$

for the eigenfunctions corresponding to the energy eigenvalues E_n.

It is easy to verify that these eigenfunctions meet all the requirements discussed in Sec. 5.1: for each quantum number n, ψ_n is a finite, single-valued function of x, and ψ_n and $\partial\psi_n/\partial x$ are continuous (except at the ends of the box). Furthermore, the integral of $|\psi_n|^2$ over all space is finite, as we can see by integrating $|\psi_n|^2\, dx$ from $x = 0$ to $x = L$ (since the particle is confined within these limits). With the help of the trigonometric identity $\sin^2\theta = \frac{1}{2}(1 - \cos 2\theta)$ we find that

$$\begin{aligned} \int_{-\infty}^{\infty} |\psi_n|^2\, dx &= \int_0^L |\psi_n|^2\, dx = A^2 \int_0^L \sin^2\left(\frac{n\pi x}{L}\right) dx \\ &= \frac{A^2}{2}\left[\int_0^L dx - \int_0^L \cos\left(\frac{2n\pi x}{L}\right) dx\right] \\ &= \frac{A^2}{2}\left[x - \left(\frac{L}{2n\pi}\right)\sin\frac{2n\pi x}{L}\right]_0^L = A^2\left(\frac{L}{2}\right) \end{aligned} \tag{5.43}$$

To normalize ψ we must assign a value to A such that $|\psi_n|^2\, dx$ is *equal* to the probability $P\, dx$ of finding the particle between x and $x + dx$, rather than merely proportional to $P\, dx$. If $|\psi_n|^2\, dx$ is to equal $P\, dx$, then it must be true that

$$\int_{-\infty}^{\infty} |\psi_n|^2\, dx = 1 \tag{5.44}$$

Comparing Eqs. (5.43) and (5.44), we see that the wave functions of a particle in a box are normalized if

$$A = \sqrt{\frac{2}{L}} \tag{5.45}$$

The normalized wave functions of the particle are therefore

Particle in a box

$$\psi_n = \sqrt{\frac{2}{L}}\,\sin\frac{n\pi x}{L} \qquad n = 1, 2, 3, \ldots \tag{5.46}$$

The normalized wave functions ψ_1, ψ_2, and ψ_3 together with the probability densities $|\psi_1|^2$, $|\psi_2|^2$, and $|\psi_3|^2$ are plotted in Fig. 5.5. Although ψ_n may be negative as well as positive, $|\psi_n|^2$ is never negative and, since ψ_n is normalized, its value at a given x is equal to the probability density of finding the particle there. In every case $|\psi_n|^2 = 0$ at $x = 0$ and $x = L$, the boundaries of the box.

At a particular place in the box the probability of the particle being present may be very different for different quantum numbers. For instance, $|\psi_1|^2$ has its maximum value of $2/L$ in the middle of the box, while $|\psi_2|^2 = 0$ there. A particle in the lowest energy level of $n = 1$ is most likely to be in the middle of the box, while a particle in the next higher state of $n = 2$ is *never* there! Classical physics, of course, suggests the same probability for the particle being anywhere in the box.

The wave functions shown in Fig. 5.5 resemble the possible vibrations of a string fixed at both ends, such as those of the stretched string of Fig. 5.2. This follows from the fact that waves in a stretched string and the wave representing a moving particle are described by equations of the same form, so that when identical restrictions are placed upon each kind of wave, the formal results are identical.

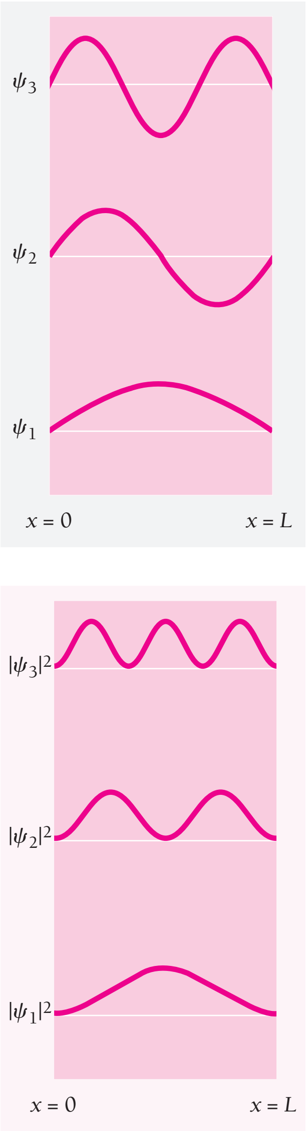


Figure 5.5 Wave functions and probability densities of a particle confined to a box with rigid walls.

Example 5.4

Find the probability that a particle trapped in a box L wide can be found between $0.45L$ and $0.55L$ for the ground and first excited states.

Solution

This part of the box is one-tenth of the box's width and is centered on the middle of the box (Fig. 5.6). Classically we would expect the particle to be in this region 10 percent of the time. Quantum mechanics gives quite different predictions that depend on the quantum number of the particle's state. From Eqs. (5.2) and (5.46) the probability of finding the particle between x_1 and x_2 when it is in the nth state is

$$P_{x_1,x_2} = \int_{x_1}^{x_2} |\psi_n|^2 \, dx = \frac{2}{L}\int_{x_1}^{x_2} \sin^2 \frac{n\pi x}{L}\, dx$$

$$= \left[\frac{x}{L} - \frac{1}{2n\pi}\sin\frac{2n\pi x}{L}\right]_{x_1}^{x_2}$$

Here $x_1 = 0.45L$ and $x_2 = 0.55L$. For the ground state, which corresponds to $n = 1$, we have

$$P_{x_1,x_2} = 0.198 = 19.8 \text{ percent}$$

This is about twice the classical probability. For the first excited state, which corresponds to $n = 2$, we have

$$P_{x_1,x_2} = 0.0065 = 0.65 \text{ percent}$$

This low figure is consistent with the probability density of $|\psi_n|^2 = 0$ at $x = 0.5L$.

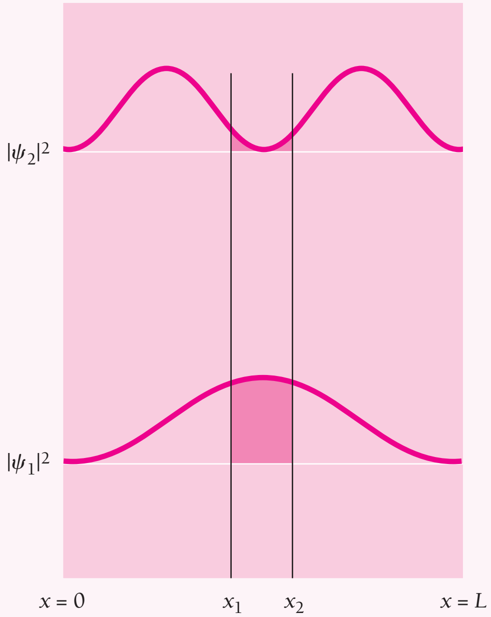


Figure 5.6 The probability P_{x_1,x_2} of finding a particle in the box of Fig. 5.5 between $x_1 = 0.45L$ and $x_2 = 0.55L$ is equal to the area under the $|\psi|^2$ curves between these limits.

Example 5.5

Find the expectation value $\langle x \rangle$ of the position of a particle trapped in a box L wide.

Solution

From Eqs. (5.19) and (5.46) we have

$$\langle x \rangle = \int_{-\infty}^{\infty} x|\psi|^2\, dx = \frac{2}{L}\int_0^L x \sin^2 \frac{n\pi x}{L}\, dx$$

$$= \frac{2}{L}\left[\frac{x^2}{4} - \frac{x \sin(2n\pi x/L)}{4n\pi/L} - \frac{\cos(2n\pi x/L)}{8(n\pi/L)^2}\right]_0^L$$

Since $\sin n\pi = 0$, $\cos 2n\pi = 1$, and $\cos 0 = 1$, for all the values of n the expectation value of x is

$$\langle x \rangle = \frac{2}{L}\left(\frac{L^2}{4}\right) = \frac{L}{2}$$

This result means that the average position of the particle is the middle of the box in all quantum states. There is no conflict with the fact that $|\psi|^2 = 0$ at $L/2$ in the $n = 2, 4, 6, \ldots$ states because $\langle x \rangle$ is an ***average***, not a probability, and it reflects the symmetry *of* $|\psi|^2$ about the middle of the box.

Momentum

Finding the momentum of a particle trapped in a one-dimensional box is not as straightforward as finding $\langle x \rangle$. Here

$$\psi^* = \psi_n = \sqrt{\frac{2}{L}}\sin\frac{n\pi x}{L}$$

$$\frac{d\psi}{dx} = \sqrt{\frac{2}{L}}\frac{n\pi}{L}\cos\frac{n\pi x}{L}$$

and so, from Eq. (5.30),

$$\langle p \rangle = \int_{-\infty}^{\infty} \psi^* \hat{p}\psi\, dx = \int_{-\infty}^{\infty} \psi^*\left(\frac{\hbar}{i}\frac{d}{dx}\right)\psi\, dx$$

$$= \frac{\hbar}{i}\frac{2}{L}\frac{n\pi}{L}\int_0^L \sin\frac{n\pi x}{L}\cos\frac{n\pi x}{L}\, dx$$

We note that

$$\int \sin ax \cos ax\, dx = \frac{1}{2a}\sin^2 ax$$

With $a = n\pi/L$ we have

$$\langle p \rangle = \frac{\hbar}{iL}\left[\sin^2 \frac{n\pi x}{L}\right]_0^L = 0$$

since $$\sin^2 0 = \sin^2 n\pi = 0 \qquad n = 1, 2, 3, \ldots$$

The expectation value $\langle p \rangle$ of the particle's momentum is 0.

At first glance this conclusion seems strange. After all, $E = p^2/2m$, and so we would anticipate that

Momentum eigenvalues for trapped particle

$$p_n = \pm\sqrt{2mE_n} = \pm\frac{n\pi\hbar}{L} \tag{5.47}$$

The ± sign provides the explanation: The particle is moving back and forth, and so its *average* momentum for any value of n is

$$p_{\text{av}} = \frac{(+n\pi\hbar/L) + (-n\pi\hbar/L)}{2} = 0$$

which is the expectation value.

According to Eq. (5.47) there should be two momentum eigenfunctions for every energy eigenfunction, corresponding to the two possible directions of motion. The general procedure for finding the eigenvalues of a quantum-mechanical operator, here $\hat{p}$, is to start from the eigenvalue equation

$$\hat{p}\psi_n = p_n\psi_n \tag{5.48}$$

where each p_n is a real number. This equation holds only when the wave functions ψ_n are eigenfunctions of the momentum operator $\hat{p}$, which here is

$$\hat{p} = \frac{\hbar}{i}\frac{d}{dx}$$

We can see at once that the energy eigenfunctions

$$\psi_n = \sqrt{\frac{2}{L}}\sin\frac{n\pi x}{L}$$

are not also momentum eigenfunctions, because

$$\frac{\hbar}{i}\frac{d}{dx}\left(\sqrt{\frac{2}{L}}\sin\frac{n\pi x}{L}\right) = \frac{\hbar}{i}\frac{n\pi}{L}\sqrt{\frac{2}{L}}\cos\frac{n\pi x}{L} \neq p_n\psi_n$$

To find the correct momentum eigenfunctions, we note that

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} = \frac{1}{2i}e^{i\theta} - \frac{1}{2i}e^{-i\theta}$$

Hence each energy eigenfunction can be expressed as a linear combination of the two wave functions

Momentum eigenfunctions for trapped particle

$$\psi_n^+ = \frac{1}{2i}\sqrt{\frac{2}{L}}\,e^{in\pi x/L} \tag{5.49}$$

$$\psi_n^- = \frac{1}{2i}\sqrt{\frac{2}{L}}\,e^{-in\pi x/L} \tag{5.50}$$

Inserting the first of these wave functions in the eigenvalue equation, Eq. (5.48), we have

$$\hat{p}\psi_n^+ = p_n^+\psi_n^+$$

$$\frac{\hbar}{i}\frac{d}{dx}\psi_n^+ = \frac{\hbar}{i}\frac{1}{2i}\sqrt{\frac{2}{L}}\,\frac{in\pi}{L}e^{in\pi x/L} = \frac{n\pi\hbar}{L}\psi_n^+ = p_n^+\psi_n^+$$

so that

$$p_n^+ = +\frac{n\pi\hbar}{L} \tag{5.51}$$

Similarly the wave function ψ_n^- leads to the momentum eigenvalues

$$p_n^- = -\frac{n\pi\hbar}{L} \tag{5.52}$$

We conclude that ψ_n^+ and ψ_n^- are indeed the momentum eigenfunctions for a particle in a box, and that Eq. (5.47) correctly states the corresponding momentum eigenvalues.

5.9 FINITE POTENTIAL WELL

The wave function penetrates the walls, which lowers the energy levels

Potential energies are never infinite in the real world, and the box with infinitely hard walls of the previous section has no physical counterpart. However, potential wells with barriers of finite height certainly do exist. Let us see what the wave functions and energy levels of a particle in such a well are.

Figure 5.7 shows a potential well with square corners that is U high and L wide and contains a particle whose energy E is less than U. According to classical mechanics, when the particle strikes the sides of the well, it bounces off without entering regions I and III. In quantum mechanics, the particle also bounces back and forth, but now it has a certain probability of penetrating into regions I and III even though $E < U$.

In regions I and III Schrödinger's steady-state equation is

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}(E - U)\psi = 0$$

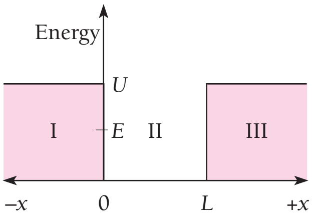


Figure 5.7 A square potential well with finite barriers. The energy E of the trapped particle is less than the height U of the barriers.

which we can rewrite in the more convenient form

$$\frac{d^2\psi}{dx^2} - a^2\psi = 0 \qquad \begin{matrix} x < 0 \\ x > L \end{matrix} \tag{5.53}$$

where

$$a = \frac{\sqrt{2m(U - E)}}{\hbar} \tag{5.54}$$

The solutions to Eq. (5.53) are real exponentials:

$$\psi_{\mathrm{I}} = Ce^{ax} + De^{-ax} \tag{5.55}$$

$$\psi_{\mathrm{III}} = Fe^{ax} + Ge^{-ax} \tag{5.56}$$

Both ψ_{I} and ψ_{III} must be finite everywhere. Since $e^{-ax} \to \infty$ as $x \to -\infty$ and $e^{ax} \to \infty$ as $x \to \infty$, the coefficients D and F must therefore be 0. Hence we have

$$\psi_{\mathrm{I}} = Ce^{ax} \tag{5.57}$$

$$\psi_{\mathrm{III}} = Ge^{-ax} \tag{5.58}$$

These wave functions decrease exponentially inside the barriers at the sides of the well.

Within the well Schrödinger's equation is the same as Eq. (5.37) and its solution is again

$$\psi_{\mathrm{II}} = A \sin \frac{\sqrt{2mE}}{\hbar} x + B \cos \frac{\sqrt{2mE}}{\hbar} x \tag{5.59}$$

In the case of a well with infinitely high barriers, we found that $B = 0$ in order that $\psi = 0$ at $x = 0$ and $x = L$. Here, however, $\psi_{\mathrm{II}} = C$ at $x = 0$ and $\psi_{\mathrm{II}} = G$ at $x = L$, so both the sine and cosine solutions of Eq. (5.59) are possible.

For either solution, both ψ and $d\psi/dx$ must be continuous at $x = 0$ and $x = L$: the wave functions inside and outside each side of the well must not only have the same value where they join but also the same slopes, so they match up perfectly. When these boundary conditions are taken into account, the result is that exact matching only occurs for certain specific values E_n of the particle energy. The complete wave functions and their probability densities are shown in Fig. 5.8.

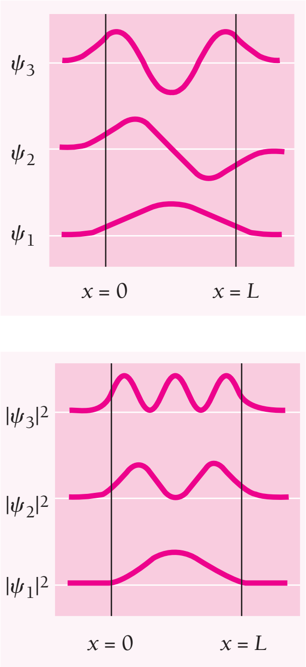


Figure 5.8 Wave functions and probability densities of a particle in a finite potential well. The particle has a certain probability of being found outside the wall.

Because the wavelengths that fit into the well are longer than for an infinite well of the same width (see Fig. 5.5), the corresponding particle momenta are lower (we recall that $\lambda = h/p$). Hence the energy levels E_n are lower for each n than they are for a particle in an infinite well.

5.10 TUNNEL EFFECT

A particle without the energy to pass over a potential barrier may still tunnel through it

Although the walls of the potential well of Fig. 5.7 were of finite height, they were assumed to be infinitely thick. As a result the particle was trapped forever even though it could penetrate the walls. We next look at the situation of a particle that strikes a potential barrier of height U, again with $E <$ U, but here the barrier has a finite width (Fig. 5.9). What we will find is that the particle has a certain probability—not

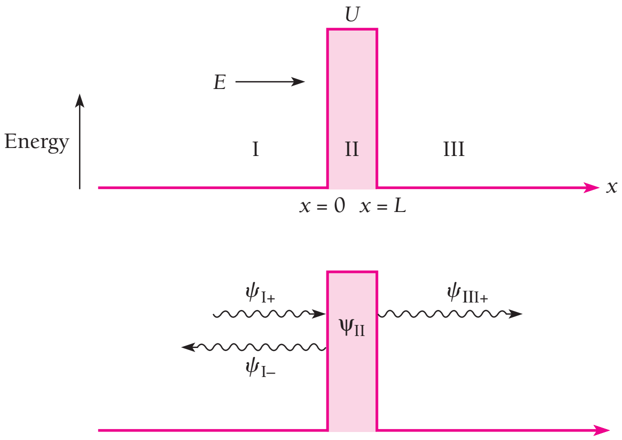


Figure 5.9 When a particle of energy $E < U$ approaches a potential barrier, according to classical mechanics the particle must be reflected. In quantum mechanics, the de Broglie waves that correspond to the particle are partly reflected and partly transmitted, which means that the particle has a finite chance of penetrating the barrier.

necessarily great, but not zero either—of passing through the barrier and emerging on the other side. The particle lacks the energy to go over the top of the barrier, but it can nevertheless tunnel through it, so to speak. Not surprisingly, the higher the barrier and the wider it is, the less the chance that the particle can get through.

The **tunnel effect** actually occurs, notably in the case of the alpha particles emitted by certain radioactive nuclei. As we shall learn in Chap. 12, an alpha particle whose kinetic energy is only a few MeV is able to escape from a nucleus whose potential wall is perhaps 25 MeV high. The probability of escape is so small that the alpha particle might have to strike the wall 10^{38} or more times before it emerges, but sooner or later it does get out. Tunneling also occurs in the operation of certain semiconductor diodes (Sec. 10.7) in which electrons pass through potential barriers even though their kinetic energies are smaller than the barrier heights.

Let us consider a beam of identical particles all of which have the kinetic energy E. The beam is incident from the left on a potential barrier of height U and width L, as in Fig. 5.9. On both sides of the barrier $U = 0$, which means that no forces act on the particles there. The wave function $\psi_{\mathrm{I}+}$ represents the incoming particles moving to the right and $\psi_{\mathrm{I}-}$ represents the reflected particles moving to the left; ψ_{III} represents the transmitted particles moving to the right. The wave function ψ_{II} represents the particles inside the barrier, some of which end up in region III while the others return to region I. The transmission probability T for a particle to pass through the barrier is equal to the fraction of the incident beam that gets through the barrier. This probability is calculated in the Appendix to this chapter. Its approximate value is given by

Approximate transmission probability

$$T = e^{-2k_2L} \tag{5.60}$$

where

$$k_2 = \frac{\sqrt{2m(U - E)}}{\hbar} \tag{5.61}$$

and L is the width of the barrier.

Example 5.6

Electrons with energies of 1.0 eV and 2.0 eV are incident on a barrier 10.0 eV high and 0.50 nm wide. (*a*) Find their respective transmission probabilities. (*b*) How are these affected if the barrier is doubled in width?

Solution

(*a*) For the 1.0-eV electrons

$$\begin{aligned} k_2 &= \frac{\sqrt{2m(U-E)}}{\hbar} \\ &= \frac{\sqrt{(2)(9.1\times 10^{-31}\text{ kg})[(10.0-1.0)\text{ eV}](1.6\times 10^{-19}\text{ J/eV})}}{1.054\times 10^{-34}\text{ J}\cdot\text{s}} \\ &= 1.6\times 10^{10}\text{ m}^{-1} \end{aligned}$$

Since $L = 0.50$ nm $= 5.0\times 10^{-10}$ m, $2k_2L = (2)(1.6\times 10^{10}\text{ m}^{-1})(5.0\times 10^{-10}\text{ m}) = 16$, and the approximate transmission probability is

$$T_1 = e^{-2k_2L} = e^{-16} = 1.1\times 10^{-7}$$

One 1.0-eV electron out of 8.9 million can tunnel through the 10-eV barrier on the average. For the 2.0-eV electrons a similar calculation gives $T_2 = 2.4\times 10^{-7}$. These electrons are over twice as likely to tunnel through the barrier.

(*b*) If the barrier is doubled in width to 1.0 nm, the transmission probabilities become

$$T_1' = 1.3\times 10^{-14} \qquad T_2' = 5.1\times 10^{-14}$$

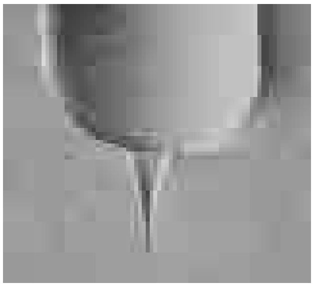
Evidently T is more sensitive to the width of the barrier than to the particle energy here.

Scanning Tunneling Microscope

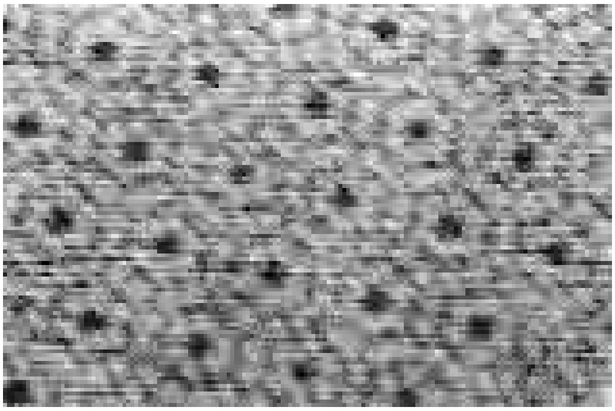
The ability of electrons to tunnel through a potential barner is used in an ingenious way in the **scanning tunneling microscope** (STM) to study surfaces on an atomic scale of size. The STM was invented in 1981 by Gert Binning and Heinrich Rohrer, who shared the 1986 Nobel Prize in physics with Ernst Ruska, the inventor of the electron microscope. In an STM, a metal probe with a point so fine that its tip is a single atom is brought close to the surface of a conducting or semiconducting material. Normally even the most loosely bound electrons in an atom on a surface need several electron-volts of energy to escape—this is the work function discussed in Chap. 2 in connection with the photoelectric effect. However, when a voltage of only 10 mV or so is applied between the probe and the surface, electrons can tunnel across the gap between them if the gap is small enough, a nanometer or two.

According to Eq. (5.60) the electron transmission probability is proportional to e^{-L}, where L is the gap width, so even a small change in L (as little as 0.01 nm, less than a twentieth the diameter of most atoms) means a detectable change in the tunneling current. What is done is to move the probe across the surface in a series of closely spaced back-and-forth scans in about the same way an electron beam traces out an image on the screen of a television picture tube. The height of the probe is continually adjusted to give a constant tunneling current, and the adjustments are recorded so that a map of surface height versus position is built up. Such a map is able to resolve individual atoms on a surface.

How can the position of the probe be controlled precisely enough to reveal the outlines of individual atoms? The thickness of certain ceramics changes when a voltage is applied across them, a property called **piezoelectricity.** The changes might be several tenths of a nanometer per volt. In an STM, piezoelectric controls move the probe in x and y directions across a surface and in the z direction perpendicular to the surface.



The tungsten probe of a scanning tunneling microscope.



Silicon atoms on the surface of a silicon crystal form a regular, repeated pattern in this image produced by an STM.

Actually, the result of an STM scan is not a true topographical map of surface height but a contour map of constant electron density on the surface. This means that atoms of different elements appear differently, which greatly increases the value of the STM as a research tool.

Although many biological materials conduct electricity, they do so by the flow of ions rather than of electrons and so cannot be studied with STMs. A more recent development, the **atomic force microscope** (AFM) can be used on any surface, although with somewhat less resolution than an STM. In an AFM, the sharp tip of a fractured diamond presses gently against the atoms on a surface. A spring keeps the pressure of the tip constant, and a record is made of the deflections of the tip as it moves across the surface. The result is a map showing contours of constant repulsive force between the electrons of the probe and the electrons of the surface atoms. Even relatively soft biological materials can be examined with an AFM and changes in them monitored. For example, the linking together of molecules of the blood protein fibrin, which occurs when blood clots, has been watched with an AFM.

5.11 HARMONIC OSCILLATOR

Its energy levels are evenly spaced

Harmonic motion takes place when a system of some kind vibrates about an equilibrium configuration. The system may be an object supported by a spring or floating in a liquid, a diatomic molecule, an atom in a crystal lattice—there are countless examples on all scales of size. The condition for harmonic motion is the presence of a restoring force that acts to return the system to its equilibrium configuration when it is disturbed. The inertia of the masses involved causes them to overshoot equilibrium, and the system oscillates indefinitely if no energy is lost.

In the special case of simple harmonic motion, the restoring force F on a particle of mass m is linear; that is, F is proportional to the particle's displacement x from its equilibrium position and in the opposite direction. Thus

Hooke's law $$F = -kx$$

This relationship is customarily called Hooke's law. From the second law of motion, $\mathbf{F} = m\mathbf{a}$, we have

$$-kx = m\frac{d^2x}{dt^2}$$

Harmonic oscillator

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \tag{5.62}$$

There are various ways to write the solution to Eq. (5.62). A common one is

$$x = A \cos (2\pi\nu t + \phi) \tag{5.63}$$

where

Frequency of harmonic oscillator

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{5.64}$$

is the frequency of the oscillations and A is their amplitude. The value of ϕ, the phase angle, depends upon what x is at the time $t = 0$ and on the direction of motion then.

The importance of the simple harmonic oscillator in both classical and modern physics lies not in the strict adherence of actual restoring forces to Hooke's law, which is seldom true, but in the fact that these restoring forces reduce to Hooke's law for small displacements x. As a result, any system in which something executes small vibrations about an equilibrium position behaves very much like a simple harmonic oscillator.

To verify this important point, we note that any restoring force which is a function of x can be expressed in a Maclaurin's series about the equilibrium position $x = 0$ as

$$F(x) = F_{x=0} + \left(\frac{dF}{dx}\right)_{x=0} x + \frac{1}{2}\left(\frac{d^2F}{dx^2}\right)_{x=0} x^2 + \frac{1}{6}\left(\frac{d^3F}{dx^3}\right)_{x=0} x^3 + \cdots$$

Since $x = 0$ is the equilibrium position, $F_{x=0} = 0$. For small x the values of $x^2, x^3, \ldots$ are very small compared with x, so the third and higher terms of the series can be neglected. The only term of significance when x is small is therefore the second one. Hence

$$F(x) = \left(\frac{dF}{dx}\right)_{x=0} x$$

which is Hooke's law when $(dF/dx)_{x=0}$ is negative, as of course it is for any restoring force. The conclusion, then, is that *all* oscillations are simple harmonic in character when their amplitudes are sufficiently small.

The potential-energy function $U(x)$ that corresponds to a Hooke's law force may be found by calculating the work needed to bring a particle from $x = 0$ to $x = x$ against such a force. The result is

$$U(x) = -\int_0^x F(x)\, dx = k\int_0^x x\, dx = \frac{1}{2}kx^2 \tag{5.65}$$

which is plotted in Fig. 5.10. The curve of $U(x)$ versus x is a parabola. If the energy of the oscillator is E, the particle vibrates back and forth between $x = -A$ and $x = +A$, where E and A are related by $E = \frac{1}{2}kA^2$. Figure 8.18 shows how a nonparabolic potential energy curve can be approximated by a parabola for small displacements.

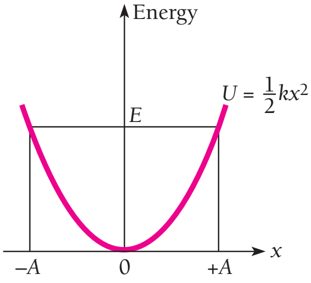


Figure 5.10 The potential energy of a harmonic oscillator is proportional to x^2, where x is the displacement from the equilibrium position. The amplitude A of the motion is determined by the total energy E of the oscillator, which classically can have any value.

Even before we make a detailed calculation we can anticipate three quantum-mechanical modifications to this classical picture:

1 The allowed energies will not form a continuous spectrum but instead a discrete spectrum of certain specific values only.
2 The lowest allowed energy will not be $E = 0$ but will be some definite minimum $E = E_0$.
3 There will be a certain probability that the particle can penetrate the potential well it is in and go beyond the limits of $-A$ and $+A$.

Energy Levels

Schrödinger's equation for the harmonic oscillator is, with $U = \frac{1}{2}kx^2$,

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}\left(E - \frac{1}{2}kx^2\right)\psi = 0 \tag{5.66}$$

It is convenient to simplify Eq. (5.75) by introducing the dimensionless quantities

$$y = \left(\frac{1}{\hbar}\sqrt{km}\right)^{1/2} x = \sqrt{\frac{2\pi m\nu}{\hbar}}\, x \tag{5.67}$$

and
$$\alpha = \frac{2E}{\hbar}\sqrt{\frac{m}{k}} = \frac{2E}{h\nu} \tag{5.68}$$

where ν is the classical frequency of the oscillation given by Eq. (5.64). In making these substitutions, what we have done is change the units in which x and E are expressed from meters and joules, respectively, to dimensionless units.

In terms of y and α Schrödinger's equation becomes

$$\frac{d^2\psi}{dy^2} + (\alpha - y^2)\psi = 0 \tag{5.69}$$

The solutions to this equation that are acceptable here are limited by the condition that $\psi \to 0$ as $y \to \infty$ in order that

$$\int_{-\infty}^{\infty} |\psi|^2 \, dy = 1$$

Otherwise the wave function cannot represent an actual particle. The mathematical properties of Eq. (5.69) are such that this condition will be fulfilled only when

$$\alpha = 2n + 1 \qquad n = 0, 1, 2, 3, \ldots$$

Since $\alpha = 2E/h\nu$ according to Eq. (5.68), the energy levels of a harmonic oscillator whose classical frequency of oscillation is ν are given by the formula

Energy levels of harmonic oscillator

$$E_n = (n + \tfrac{1}{2})h\nu \qquad n = 0, 1, 2, 3, \ldots \tag{5.70}$$

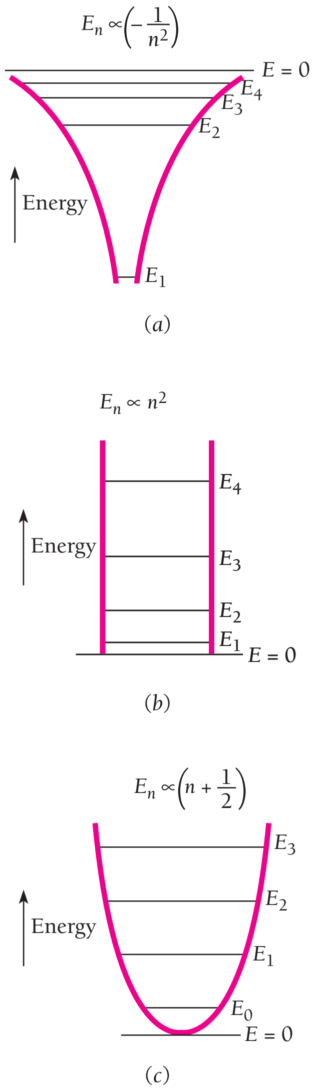


Figure 5.11 Potential wells and energy levels of (*a*) a hydrogen atom, (*b*) a particle in a box, and (*c*) a harmonic oscillator. In each case the energy levels depend in a different way on the quantum number *n*. Only for the harmonic oscillator are the levels equally spaced. The symbol ∝ means "is proportional to."

The energy of a harmonic oscillator is thus quantized in steps of $h\nu$.

We note that when $n = 0$,

Zero-point energy

$$E_0 = \tfrac{1}{2}h\nu \qquad (5.71)$$

which is the lowest value the energy of the oscillator can have. This value is called the **zero-point energy** because a harmonic oscillator in equilibrium with its surroundings would approach an energy of $E = E_0$ and not $E = 0$ as the temperature approaches 0 K.

Figure 5.11 is a comparison of the energy levels of a harmonic oscillator with those of a hydrogen atom and of a particle in a box with infinitely hard walls. The shapes of the respective potential-energy curves are also shown. The spacing of the energy levels is constant only for the harmonic oscillator.

Wave Functions

For each choice of the parameter α_n there is a different wave function ψ_n. Each function consists of a polynomial $H_n(y)$ (called a **Hermite polynomial**) in either odd or even powers of y, the exponential factor $e^{-y^2/2}$, and a numerical coefficient which is needed for ψ_n to meet the normalization condition

$$\int_{-\infty}^{\infty} |\psi_n|^2 \, dy = 1 \qquad n = 0, 1, 2 \ldots$$

The general formula for the nth wave function is

Harmonic oscillator

$$\psi_n = \left(\frac{2m\nu}{\hbar}\right)^{1/4} (2^n n!)^{-1/2} H_n(y) e^{-y^2/2} \qquad (5.72)$$

The first six Hermite polynomials $H_n(y)$ are listed in Table 5.2.

The wave functions that correspond to the first six energy levels of a harmonic oscillator are shown in Fig. 5.12. In each case the range to which a particle oscillating classically with the same total energy E_n would be confined is indicated. Evidently the particle is able to penetrate into classically forbidden regions—in other words, to exceed the amplitude A determined by the energy—with an exponentially decreasing probability, just as in the case of a particle in a finite square potential well.

It is interesting and instructive to compare the probability densities of a classical harmonic oscillator and a quantum-mechanical harmonic oscillator of the same energy. The upper curves in Fig. 5.13 show this density for the classical oscillator. The probability P of finding the particle at a given position is greatest at the endpoints of its motion,

Table 5.2 Some Hermite Polynomials

n	$H_n(y)$	α_n	E_n
0	1	1	$\frac{1}{2}h\nu$
1	$2y$	3	$\frac{3}{2}h\nu$
2	$4y^2 - 2$	5	$\frac{5}{2}h\nu$
3	$8y^3 - 12y$	7	$\frac{7}{2}h\nu$
4	$16y^4 - 48y^2 + 12$	9	$\frac{9}{2}h\nu$
5	$32y^5 - 160y^3 + 120y$	11	$\frac{11}{2}h\nu$

where it moves slowly, and least near the equilibrium position ($x = 0$), where it moves rapidly.

Exactly the opposite behavior occurs when a quantum-mechanical oscillator is in its lowest energy state of $n = 0$. As shown, the probability density $|\psi_0|^2$ has its maximum value at $x = 0$ and drops off on either side of this position. However, this disagreement becomes less and less marked with increasing n. The lower graph of Fig. 5.13 corresponds to $n = 10$, and it is clear that $|\psi_{10}|^2$ when averaged over x has approximately the general character of the classical probability P. This is another example of the correspondence principle mentioned in Chap. 4: In the limit of large quantum numbers, quantum physics yields the same results as classical physics.

It might be objected that although $|\psi_{10}|^2$ does indeed approach P when smoothed out, nevertheless $|\psi_{10}|^2$ fluctuates rapidly with x whereas P does not. However, this objection has meaning only if the fluctuations are observable, and the smaller the spacing of the peaks and hollows, the more difficult it is to detect them experimentally. The exponential "tails" of $|\psi_{10}|^2$ beyond $x = \pm A$ also decrease in magnitude with increasing n. Thus the classical and quantum pictures begin to resemble each other more and more the larger the value of n, in agreement with the correspondence principle, although they are very different for small n.

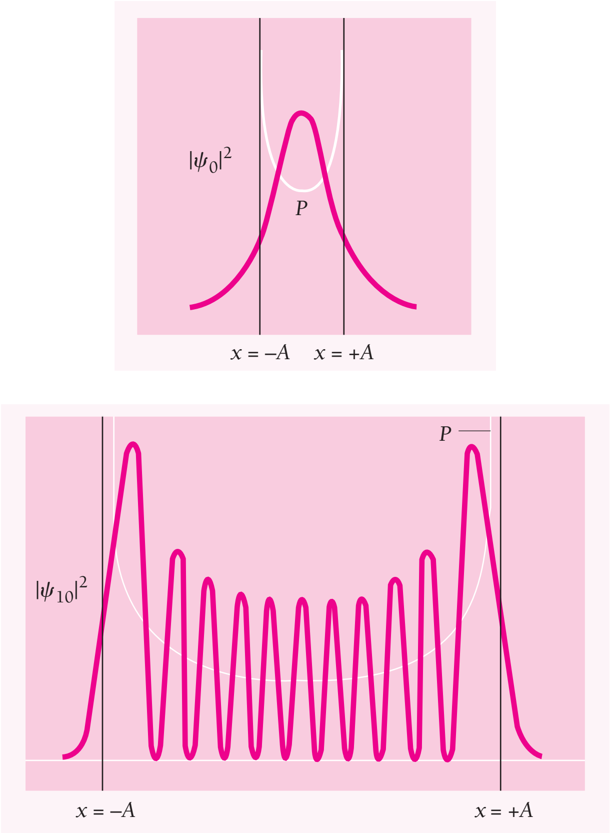


Figure 5.13 Probability densities for the $n = 0$ and $n = 10$ states of a quantum-mechanical harmonic oscillator. The probability densities for classical harmonic oscillators with the same energies are shown in white. In the $n = 10$ state, the wavelength is shortest at $x = 0$ and longest at $x = -A$.

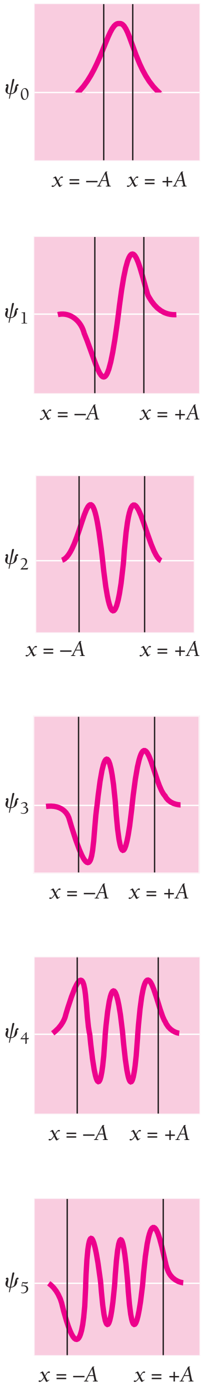


Figure 5.12 The first six harmonic-oscillator wave functions. The vertical lines show the limits $-A$ and $+A$ between which a classical oscillator with the same energy would vibrate.

Example 5.7

Find the expectation value $\langle x \rangle$ for the first two states of a harmonic oscillator.

Solution

The general formula for $\langle x \rangle$ is

$$\langle x \rangle = \int_{-\infty}^{\infty} x\,|\psi|^2\,dx$$

In calculations such as this it is easier to begin with y in place of x and afterward use Eq. (5.67) to change to x. From Eq. (5.72) and Table 5.2,

$$\psi_0 = \left(\frac{2m\nu}{\hbar}\right)^{1/4} e^{-y^2/2}$$

$$\psi_1 = \left(\frac{2m\nu}{\hbar}\right)^{1/4}\left(\frac{1}{2}\right)^{1/2}(2y)\,e^{-y^2/2}$$

The values of $\langle x \rangle$ for $n = 0$ and $n = 1$ will respectively be proportional to the integrals

$$n = 0\colon \int_{-\infty}^{\infty} y|\psi_0|^2\,dy = \int_{-\infty}^{\infty} ye^{-y^2}\,dy = -\left[\frac{1}{2}e^{-y^2}\right]_{-\infty}^{\infty} = 0$$

$$n = 1\colon \int_{-\infty}^{\infty} y|\psi_1|^2\,dy = \int_{-\infty}^{\infty} y^3e^{-y^2}\,dy = -\left[\left(\frac{1}{4} + \frac{y^2}{2}\right)e^{-y^2}\right]_{-\infty}^{\infty} = 0$$

The expectation value $\langle x \rangle$ is therefore 0 in both cases. In fact, $\langle x \rangle = 0$ for *all* states of a harmonic oscillator, which could be predicted since $x = 0$ is the equilibrium position of the oscillator where its potential energy is a minimum.

Appendix to Chapter 5

The Tunnel Effect

We consider the situation that was shown in Fig. 5.9 of a particle of energy $E < U$ that approaches a potential barrier U high and L wide. Outside the barrier in regions I and III Schrödinger's equation for the particle takes the forms

$$\frac{d^2\psi_{\text{I}}}{dx^2} + \frac{2m}{\hbar^2}E\psi_{\text{I}} = 0 \tag{5.73}$$

$$\frac{d^2\psi_{\text{III}}}{dx^2} + \frac{2m}{\hbar^2}E\psi_{\text{III}} = 0 \tag{5.74}$$

The solutions to these equations that are appropriate here are

$$\psi_{\text{I}} = Ae^{ik_1x} + Be^{-ik_1x} \tag{5.75}$$

$$\psi_{\text{III}} = Fe^{ik_1x} + Ge^{-ik_1x} \tag{5.76}$$

where

Wave number outside barrier

$$k_1 = \frac{\sqrt{2mE}}{\hbar} = \frac{p}{\hbar} = \frac{2\pi}{\lambda} \tag{5.77}$$

is the wave number of the de Broglie waves that represent the particles outside the barrier.

Because

$$e^{i\theta} = \cos\theta + i\sin\theta$$

$$e^{-i\theta} = \cos\theta - i\sin\theta$$

these solutions are equivalent to Eq. (5.38)—the values of the coefficients are different in each case, of course—but are in a more suitable form to describe particles that are not trapped.

The various terms in Eqs. (5.75) and (5.76) are not hard to interpret. As was shown schematically in Fig. 5.9, Ae^{ik_1x} is a wave of amplitude A incident from the left on the barrier. Hence we can write

Incoming wave

$$\psi_{\text{I}+} = Ae^{ik_1x} \tag{5.78}$$

This wave corresponds to the incident beam of particles in the sense that $|\psi_{\text{I}+}|^2$ is their probability density. If $v_{\text{I}+}$ is the group velocity of the incoming wave, which equals the velocity of the particles, then

$$S = |\psi_{\text{I}+}|^2 v_{\text{I}+}$$

is the flux of particles that arrive at the barrier. That is, S is the number of particles per second that arrive there.

At $x = 0$ the incident wave strikes the barrier and is partially reflected, with

Reflected wave

$$\psi_{\text{I}-} = Be^{-ik_1x} \tag{5.79}$$

representing the reflected wave. Hence

$$\psi_{\text{I}} = \psi_{\text{I}+} + \psi_{\text{I}-} \tag{5.80}$$

On the far side of the barrier ($x > L$) there can only be a wave

Transmitted wave

$$\psi_{\text{III}+} = Fe^{ik_1x} \tag{5.81}$$

traveling in the $+x$ direction at the velocity $\boldsymbol{v}_{\text{III}+}$ since region III contains nothing that could reflect the wave. Hence $G = 0$ and

$$\psi_{\text{III}} = \psi_{\text{III}+} = Fe^{ik_1x} \tag{5.82}$$

The transmission probability T for a particle to pass through the barrier is the ratio

Transmission probability

$$T = \frac{|\psi_{\text{III}+}|^2\boldsymbol{v}_{\text{III}+}}{|\psi_{\text{I}+}|^2\boldsymbol{v}_{\text{I}+}} = \frac{FF^*\boldsymbol{v}_{\text{III}+}}{AA^*\boldsymbol{v}_{\text{I}+}} \tag{5.83}$$

between the flux of particles that emerges from the barrier and the flux that arrives at it. In other words, T is the fraction of incident particles that succeed in tunneling through the barrier. Classically $T = 0$ because a particle with $E < U$ cannot exist inside the barrier; let us see what the quantum-mechanical result is.

In region II Schrödinger's equation for the particles is

$$\frac{d^2\psi_{\text{II}}}{dx^2} + \frac{2m}{\hbar^2}(E - U)\psi_{\text{II}} = \frac{d^2\psi_{\text{II}}}{dx^2} - \frac{2m}{\hbar^2}(U - E)\psi_{\text{II}} = 0 \tag{5.84}$$

Since $U > E$ the solution is

Wave function inside barrier

$$\psi_{\text{II}} = Ce^{-k_2x} + De^{k_2x} \tag{5.85}$$

where the wave number inside the barrier is

Wave number inside barrier

$$k_2 = \frac{\sqrt{2m(U - E)}}{\hbar} \tag{5.86}$$

Since the exponents are real quantities, ψ_{II} does not oscillate and therefore does not represent a moving particle. However, the probability density $|\psi_{\text{II}}|^2$ is not zero, so there is a finite probability of finding a particle within the barrier. Such a particle may emerge into region III or it may return to region I.

Applying the Boundary Conditions

In order to calculate the transmission probability T we have to apply the appropriate boundary conditions to ψ_{I}, ψ_{II}, and ψ_{III}. Fig. 5.14 shows the wave functions in regions I, II, and III. As discussed earlier, both ψ and its derivative $\partial\psi/\partial x$ must be continuous everywhere. With reference to Fig. 5.14, these conditions mean that for a perfect fit at each side of the barrier, the wave functions inside and outside must have the same value and the same slope. Hence at the left-hand side of the barrier

Boundary conditions at $x = 0$

$$\left.\begin{aligned} \psi_{\mathrm{I}} &= \psi_{\mathrm{II}} \qquad (5.87) \\ \frac{d\psi_{\mathrm{I}}}{dx} &= \frac{d\psi_{\mathrm{II}}}{dx} \qquad (5.88) \end{aligned}\right\} x = 0$$

and at the right-hand side

Boundary conditions at $x = L$

$$\left.\begin{aligned} \psi_{\mathrm{II}} &= \psi_{\mathrm{III}} \qquad (5.89) \\ \frac{d\psi_{\mathrm{II}}}{dx} &= \frac{d\psi_{\mathrm{III}}}{dx} \qquad (5.90) \end{aligned}\right\} x = L$$

Now we substitute ψ_{I}, ψ_{II}, and ψ_{III} from Eqs. (5.75), (5.81), and (5.85) into the above equations. This yields in the same order

$$A + B = C + D \tag{5.91}$$

$$ik_1A - ik_1B = -k_2C + k_2D \tag{5.92}$$

$$Ce^{-k_2L} + De^{k_2L} = Fe^{ik_1L} \tag{5.93}$$

$$-k_2Ce^{-k_2L} + k_2De^{k_2L} = ik_1Fe^{ik_1L} \tag{5.94}$$

Equations (5.91) to (5.94) may be solved for (A/F) to give

$$\left(\frac{A}{F}\right) = \left[\frac{1}{2} + \frac{i}{4}\left(\frac{k_2}{k_1} - \frac{k_1}{k_2}\right)\right]e^{(ik_1+k_2)L} + \left[\frac{1}{2} - \frac{i}{4}\left(\frac{k_2}{k_1} - \frac{k_1}{k_2}\right)\right]e^{(ik_1-k_2)L} \tag{5.95}$$

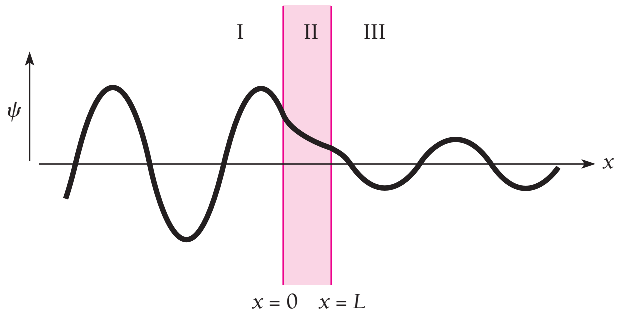


Figure 5.14 At each wall of the barrier, the wave functions inside and outside it must match up perfectly, which means that they must have the same values and slopes there.

Let us assume that the potential barrier U is high relative to the energy E of the incident particles. If this is the case, then $k_2/k_1 > k_1/k_2$ and

$$\frac{k_2}{k_1} - \frac{k_1}{k_2} \approx \frac{k_2}{k_1} \tag{5.96}$$

Let us also assume that the barrier is wide enough for ψ_{II} to be severely weakened between $x = 0$ and $x = L$. This means that $k_2 L \gg 1$ and

$$e^{k_2 L} \gg e^{-k_2 L}$$

Hence Eq. (5.95) can be approximated by

$$\left(\frac{A}{F}\right) = \left(\frac{1}{2} + \frac{ik_2}{4k_1}\right) e^{(ik_1 + k_2)L} \tag{5.97}$$

The complex conjugate of (A/F), which we need to compute the transmission probability T, is found by replacing i by $-i$ wherever it occurs in (A/F):

$$\left(\frac{A}{F}\right)^{*} = \left(\frac{1}{2} - \frac{ik_2}{4k_1}\right) e^{(-ik_1 + k_2)L} \tag{5.98}$$

Now we multiply (A/F) and $(A/F)^*$ to give

$$\frac{AA^*}{FF^*} = \left(\frac{1}{4} + \frac{k_2^2}{16k_1^2}\right) e^{2k_2 L}$$

Here $\boldsymbol{v}_{\mathrm{III}+} = \boldsymbol{v}_{\mathrm{I}+}$ so $\boldsymbol{v}_{\mathrm{III}+}/\boldsymbol{v}_{\mathrm{I}+} = 1$ in Eq. (5.83), which means that the transmission probability is

Transmission probability

$$T = \frac{FF^*\boldsymbol{v}_{\mathrm{III}+}}{AA^*\boldsymbol{v}_{\mathrm{I}+}} = \left(\frac{AA^*}{FF^*}\right)^{-1} = \left[\frac{16}{4 + (k_2/k_1)^2}\right] e^{-2k_2 L} \tag{5.99}$$

From the definitions of k_1, Eq. (5.77), and of k_2, Eq. (5.86), we see that

$$\left(\frac{k_2}{k_1}\right)^2 = \frac{2m(U - E)/\hbar^2}{2mE/\hbar^2} = \frac{U}{E} - 1 \tag{5.100}$$

This formula means that the quantity in brackets in Eq. (5.99) varies much less with E and U than does the exponential. The bracketed quantity, furthermore, always is of the order of magnitude of 1 in value. A reasonable approximation of the transmission probability is therefore

Approximate transmission probability

$$T = e^{-2k_2 L} \tag{5.101}$$

as stated in Sec. 5.10.

EXERCISES

Press on, and faith will catch up with you. —Jean D'Alembert

5.1 Quantum Mechanics

1. Which of the wave functions in Fig. 5.15 cannot have physical significance in the interval shown? Why not?

2. Which of the wave functions in Fig. 5.16 cannot have physical significance in the interval shown? Why not?

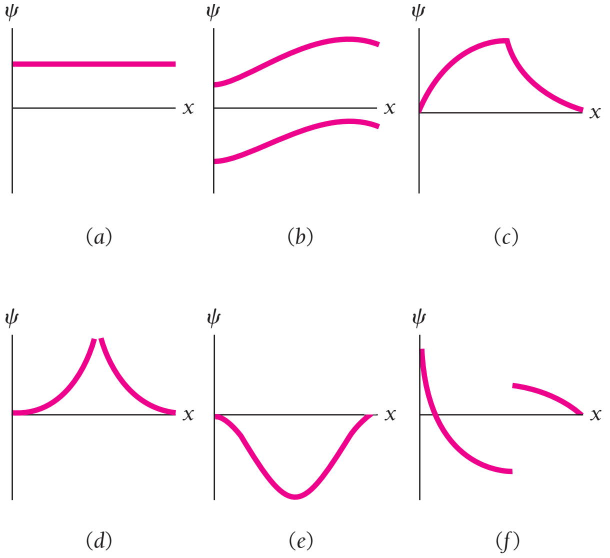


Figure 5.15

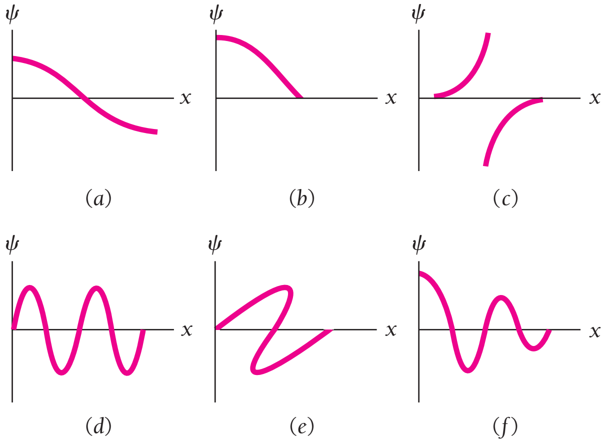


Figure 5.16

3. Which of the following wave functions cannot be solutions of Schrödinger's equation for all values of x? Why not? (*a*) $\psi = A \sec x$; (*b*) $\psi = A \tan x$; (*c*) $\psi = Ae^{x^2}$; (*d*) $\psi = Ae^{-x^2}$.

4. Find the value of the normalization constant A for the wave function $\psi = Axe^{-x^2/2}$.

5. The wave function of a certain particle is $\psi = A \cos^2 x$ for $-\pi/2 < x < \pi/2$. (*a*) Find the value of A. (*b*) Find the probability that the particle be found between $x = 0$ and $x = \pi/4$.

5.2 The Wave Equation

6. The formula $y = A \cos \omega(t - x/v)$, as we saw in Sec. 3.3, describes a wave that moves in the $+x$ direction along a stretched string. Show that this formula is a solution of the wave equation, Eq.(5.3).

7. As mentioned in Sec. 5.1, in order to give physically meaningful results in calculations a wave function and its partial derivatives must be finite, continuous, and single-valued, and in addition must be normalizable. Equation (5.9) gives the wave function of a particle moving freely (that is, with no forces acting on it) in the $+x$ direction as

$$\Psi = Ae^{-(i/\hbar)(Et - px)}$$

where E is the particle's total energy and p is its momentum. Does this wave function meet all the above requirements? If not, could a linear superposition of such wave functions meet these requirements? What is the significance of such a superposition of wave functions?

5.4 Linearity and Superposition

8. Prove that Schrödinger's equation is linear by showing that

$$\Psi = a_1\Psi_1(x, t) + a_2\Psi_2(x, t)$$

is also a solution of Eq. (5.14) if Ψ_1 and Ψ_2 are themselves solutions.

5.6 Operators

9. Show that the expectation values $\langle px \rangle$ and $\langle xp \rangle$ are related by

$$\langle px \rangle - \langle xp \rangle = \frac{\hbar}{i}$$

This result is described by saying that p and x do not **commute** and it is intimately related to the uncertainty principle.

10. An eigenfunction of the operator d^2/dx^2 is $\sin nx$, where $n = 1, 2, 3, \ldots$. Find the corresponding eigenvalues.

5.7 Schrödinger's Equation: Steady-State Form

11. Obtain Schrödinger's steady-state equation from Eq. (3.5) with the help of de Broglie's relationship $\lambda = h/mv$ by letting $y = \psi$ and finding $\partial^2\psi/\partial x^2$.

5.8 Particle in a Box

12. According to the correspondence principle, quantum theory should give the same results as classical physics in the limit of large quantum numbers. Show that as $n \rightarrow \infty$, the probability of finding the trapped particle of Sec. 5.8 between x and $x + \Delta x$ is $\Delta x/L$ and so is independent of x, which is the classical expectation.

13. One of the possible wave functions of a particle in the potential well of Fig. 5.17 is sketched there. Explain why the wavelength and amplitude of ψ vary as they do.

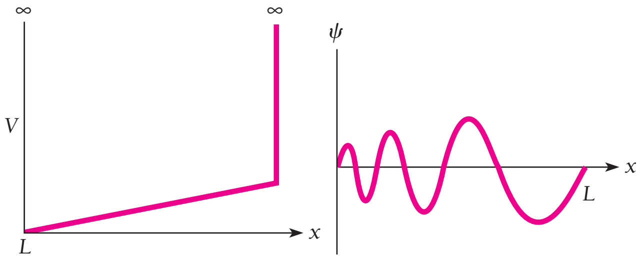


Figure 5.17

14. In Sec. 5.8 a box was considered that extends from $x = 0$ to $x = L$. Suppose the box instead extends from $x = x_0$ to $x = x_0 + L$, where $x_0 \neq 0$. Would the expression for the wave functions of a particle in this box be any different from those in the box that extends from $x = 0$ to $x = L$? Would the energy levels be different?

15. An important property of the eigenfunctions of a system is that they are **orthogonal** to one another, which means that

$$\int_{-\infty}^{\infty} \psi_n \psi_m \, dV = 0 \qquad n \neq m$$

Verify this relationship for the eigenfunctions of a particle in a one-dimensional box given by Eq. (5.46).

16. A rigid-walled box that extends from $-L$ to L is divided into three sections by rigid interior walls at $-x$ and x, where $x < L$. Each section contains one particle in its ground state. (*a*) What is the total energy of the system as a function of x? (*b*) Sketch $E(x)$ versus x. (*c*) At what value of x is $E(x)$ a minimum?

17. As shown in the text, the expectation value $\langle x \rangle$ of a particle trapped in a box L wide is $L/2$, which means that its average position is the middle of the box. Find the expectation value $\langle x^2 \rangle$.

18. As noted in Exercise 8, a linear combination of two wave functions for the same system is also a valid wave function. Find the normalization constant B for the combination

$$\psi = B\left(\sin\frac{\pi x}{L} + \sin\frac{2\pi x}{L}\right)$$

of the wave functions for the $n = 1$ and $n = 2$ states of a particle in a box L wide.

19. Find the probability that a particle in a box L wide can be found between $x = 0$ and $x = L/n$ when it is in the nth state.

20. In Sec. 3.7 the standard deviation σ of a set of N measurements of some quantity x was defined as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x_0)^2}$$

(*a*) Show that, in terms of expectation values, this formula can be written as

$$\sigma = \sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$$

(*b*) If the uncertainty in position of a particle in a box is taken as the standard deviation, find the uncertainty in the expectation value $\langle x \rangle = L/2$ for $n = 1$. (*c*) What is the limit of Δx as n increases?

21. A particle is in a cubic box with infinitely hard walls whose edges are L long (Fig. 5.18). The wave functions of the particle are given by

$$\psi = A \sin\frac{n_x \pi x}{L} \sin\frac{n_y \pi y}{L} \sin\frac{n_z \pi z}{L} \qquad \begin{aligned} n_x &= 1, 2, 3, \ldots \\ n_y &= 1, 2, 3, \ldots \\ n_z &= 1, 2, 3, \ldots \end{aligned}$$

Find the value of the normalization constant A.

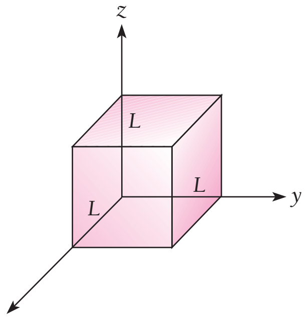


Figure 5.18 A cubic box.

22. The particle in the box of Exercise 21 is in its ground state of $n_x = n_y = n_z = 1$. (*a*) Find the probability that the particle will be found in the volume defined by $0 \le x \le L/4$, $0 \le y \le L/4$, $0 \le z \le L/4$. (*b*) Do the same for $L/2$ instead of $L/4$.

23. (*a*) Find the possible energies of the particle in the box of Exercise 21 by substituting its wave function ψ in Schrödinger's equation and solving for E. (*Hint*: Inside the box $U = 0$.) (*b*) Compare the ground-state energy of a particle in a one-dimensional box of length L with that of a particle in the three-dimensional box.

5.10 Tunnel Effect

24. Electrons with energies of 0.400 eV are incident on a barrier 3.00 eV high and 0.100 nm wide. Find the approximate probability for these electrons to penetrate the barrier.

25. A beam of electrons is incident on a barrier 6.00 eV high and 0.200 nm wide. Use Eq. (5.60) to find the energy they should have if 1.00 percent of them are to get through the barrier.

5.11 Harmonic Oscillator

26. Show that the energy-level spacing of a harmonic oscillator is in accord with the correspondence principle by finding the ratio $\Delta E_n/E_n$ between adjacent energy levels and seeing what happens to this ratio as $n \to \infty$.

27. What bearing would you think the uncertainty principle has on the existence of the zero-point energy of a harmonic oscillator?

28. In a harmonic oscillator, the particle varies in position from $-A$ to $+A$ and in momentum from $-p_0$ to $+p_0$. In such an oscillator, the standard deviations of x and p are $\Delta x = A/\sqrt{2}$ and $\Delta p = p_0/\sqrt{2}$. Use this observation to show that the minimum energy of a harmonic oscillator is $\frac{1}{2}h\nu$.

29. Show that for the $n = 0$ state of a harmonic oscillator whose classical amplitude of motion is A, $y = 1$ at $x = A$, where y is the quantity defined by Eq. (5.67).

30. Find the probability density $|\psi_0|^2\, dx$ at $x = 0$ and at $x = \pm A$ of a harmonic oscillator in its $n = 0$ state (see Fig. 5.13).

31. Find the expectation values $\langle x \rangle$ and $\langle x^2 \rangle$ for the first two states of a harmonic oscillator.

32. The potential energy of a harmonic oscillator is $U = \frac{1}{2}kx^2$. Show that the expectation value $\langle U \rangle$ of U is $E_0/2$ when the oscillator is in the $n = 0$ state. (This is true of all states of the harmonic oscillator, in fact.) What is the expectation value of the oscillator's kinetic energy? How do these results compare with the classical values of $\overline{U}$ and $\overline{KE}$?

33. A pendulum with a 1.00-g bob has a massless string 250 mm long. The period of the pendulum is 1.00 s. (*a*) What is its zero-point energy? Would you expect the zero-point oscillations to be detectable? (*b*) The pendulum swings with a very small amplitude such that its bob rises a maximum of 1.00 mm above its equilibrium position. What is the corresponding quantum number?

34. Show that the harmonic-oscillator wave function ψ_1 is a solution of Schrödinger's equation.

35. Repeat Exercise 34 for ψ_2.

36. Repeat Exercise 34 for ψ_3.

Appendix: The Tunnel Effect

37. Consider a beam of particles of kinetic energy E incident on a potential step at $x = 0$ that is U high, where $E > U$ (Fig. 5.19). (*a*) Explain why the solution $De^{-ik'x}$ (in the notation of appendix) has no physical meaning in this situation, so that $D = 0$. (*b*) Show that the transmission probability here is $T = CC^*\mathbf{v}'/AA^*\mathbf{v}_1 = 4k_1^2/(k_1 + k')^2$. (*c*) A 1.00-mA beam of electrons moving at 2.00×10^6 m/s enters a region with a sharply defined boundary in which the electron speeds are reduced to 1.00×10^6 m/s by a difference in potential. Find the transmitted and reflected currents.

38. An electron and a proton with the same energy E approach a potential barrier whose height U is greater than E. Do they have the same probability of getting through? If not, which has the greater probability?

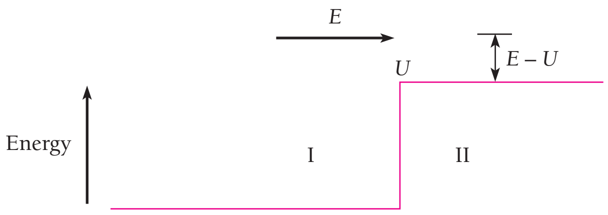


Figure 5.19

Class: B. Tech (Unit III)

I have taken all course materials for Unit III from Book Concept of Modern Physics by Arthur Besier, Shobhit Mahajan & S. Rai Choudhury (McGraw Hill Education).

Students can download this book form given web address;

Web Address : **https://b-ok.cc/book/2700591/864ac0**

All topics of unit III (Special Theory of Relativity) have been taken from **Chapter1** from above said book (**https://b-ok.cc/book/2700591/864ac0**). I am sending pdf file of Chapter 1.

UNIT-III: Special Theory of Relativity **(8 Hours)**

Inertial Frames Of Reference, Galilian And Lorentz Transformations, Postulates of Relativity, Time Dilation, Twin Paradox, Length Contraction, Relativistic Mass, Energy and Momentum, Equivalence of Mass And Energy, Doppler Effect In Light And Its Application in Expanding of Universe, Problems.

CHAPTER 1

Relativity

According to the theory of relativity, nothing can travel faster than light. Although today's spacecraft can exceed 10 km/s, they are far from this ultimate speed limit.

In 1905 a young physicist of twenty-six named Albert Einstein showed how measurements of time and space are affected by motion between an observer and what is being observed. To say that Einstein's theory of relativity revolutionized science is no exaggeration. Relativity connects space and time, matter and energy, electricity and magnetism—links that are crucial to our understanding of the physical universe. From relativity have come a host of remarkable predictions, all of which have been confirmed by experiment. For all their profundity, many of the conclusions of relativity can be reached with only the simplest of mathematics.

1.1 SPECIAL RELATIVITY

All motion is relative; the speed of light in free space is the same for all observers

When such quantities as length, time interval, and mass are considered in elementary physics, no special point is made about how they are measured. Since a standard unit exists for each quantity, who makes a certain determination would not seem to matter—everybody ought to get the same result. For instance, there is no question of principle involved in finding the length of an airplane when we are on board. All we have to do is put one end of a tape measure at the airplane's nose and look at the number on the tape at the airplane's tail.

But what if the airplane is in flight and we are on the ground? It is not hard to determine the length of a distant object with a tape measure to establish a baseline, a surveyor's transit to measure angles, and a knowledge of trigonometry. When we measure the moving airplane from the ground, though, we find it to be shorter than it is to somebody in the airplane itself. To understand how this unexpected difference arises we must analyze the process of measurement when motion is involved.

Frames of Reference

The first step is to clarify what we mean by motion. When we say that something is moving, what we mean is that its position relative to something else is changing. A passenger moves relative to an airplane; the airplane moves relative to the earth; the earth moves relative to the sun; the sun moves relative to the galaxy of stars (the Milky Way) of which it is a member; and so on. In each case a **frame of reference** is part of the description of the motion. To say that something is moving always implies a specific frame of reference.

An **inertial frame of reference** is one in which Newton's first law of motion holds. In such a frame, an object at rest remains at rest and an object in motion continues to move at constant velocity (constant speed and direction) if no force acts on it. Any frame of reference that moves at constant velocity relative to an inertial frame is itself an inertial frame.

All inertial frames are equally valid. Suppose we see something changing its position with respect to us at constant velocity. Is it moving or are we moving? Suppose we are in a closed laboratory in which Newton's first law holds. Is the laboratory moving or is it at rest? These questions are meaningless because all constant-velocity motion is relative. There is no universal frame of reference that can be used everywhere, no such thing as "absolute motion."

The **theory of relativity** deals with the consequences of the lack of a universal frame of reference. **Special relativity,** which is what Einstein published in 1905, treats problems that involve inertial frames of reference. **General relativity,** published by Einstein a decade later, describes the relationship between gravity and the geometrical structure of space and time. The special theory has had an enormous impact on much of physics, and we shall concentrate on it here.

Postulates of Special Relativity

Two postulates underlie special relativity. The first, the **principle of relativity,** states:

> The laws of physics are the same in all inertial frames of reference.

This postulate follows from the absence of a universal frame of reference. If the laws of physics were different for different observers in relative motion, the observers could find from these differences which of them were "stationary" in space and which were "moving." But such a distinction does not exist, and the principle of relativity expresses this fact.

The second postulate is based on the results of many experiments:

> The speed of light in free space has the same value in all inertial frames of reference.

This speed is 2.998×10^8 m/s to four significant figures.

To appreciate how remarkable these postulates are, let us look at a hypothetical experiment basically no different from actual ones that have been carried out in a number of ways. Suppose I turn on a searchlight just as you fly past in a spacecraft at a speed of 2×10^8 m/s (Fig. 1.1). We both measure the speed of the light waves from the searchlight using identical instruments. From the ground I find their speed to be 3×10^8 m/s as usual. "Common sense" tells me that you ought to find a speed of $(3 - 2) \times 10^8$ m/s, or only 1×10^8 m/s, for the same light waves. But you also find their speed to be 3×10^8 m/s, even though to me you seem to be moving parallel to the waves at 2×10^8 m/s.

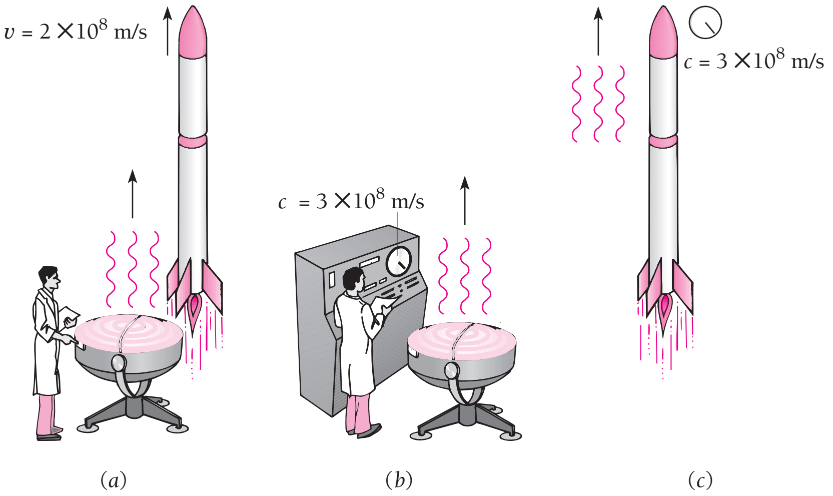


Figure 1.1 The speed of light is the same to all observers.

Albert A. Michelson (1852–1931) was born in Germany but came to the United States at the age of two with his parents, who settled in Nevada. He attended the U.S. Naval Academy at Annapolis where, after two years of sea duty, he became a science instructor. To improve his knowledge of optics, in which he wanted to specialize, Michelson went to Europe and studied in Berlin and Paris. Then he left the Navy to work first at the Case School of Applied Science in Ohio, then at Clark University in Massachusetts, and finally at the University of Chicago, where he headed the physics department from 1892 to 1929. Michelson's speciality was high-precision measurement, and for many decades his successive figures for the speed of light were the best available. He redefined the meter in terms of wavelengths of a particular spectral line and devised an interferometer that could determine the diameter of a star (stars appear as points of light in even the most powerful telescopes).

Michelson's most significant achievement, carried out in 1887 in collaboration with Edward Morley, was an experiment to measure the motion of the earth through the "ether," a hypothetical medium pervading the universe in which light waves were supposed to occur. The notion of the ether was a hangover from the days before light waves were recognized as electromagnetic, but nobody at the time seemed willing to discard the idea that light propagates relative to some sort of universal frame of reference.

To look for the earth's motion through the ether, Michelson and Morley used a pair of light beams formed by a half-silvered mirror, as in Fig. 1.2. One light beam is directed to a mirror along a path perpendicular to the ether current, and the other goes to a mirror along a path parallel to the ether current. Both beams end up at the same viewing screen. The clear glass plate ensures that both beams pass through the same thicknesses of air and glass. If the transit times of the two beams are the same, they will arrive at the screen in phase and will interfere constructively. An ether current due to the earth's motion parallel to one of the beams, however, would cause the beams to have different transit times and the result would be destructive interference at the screen. This is the essence of the experiment.

Although the experiment was sensitive enough to detect the expected ether drift, to everyone's surprise none was found. The negative result had two consequences. First, it showed that the ether does not exist and so there is no such thing as "absolute motion" relative to the ether: all motion is relative to a specified frame of reference, not to a universal one. Second, the result showed that the speed of light is the same for all observers, which is not true of waves that need a material medium in which to occur (such as sound and water waves).

The Michelson-Morley experiment set the stage for Einstein's 1905 special theory of relativity, a theory that Michelson himself was reluctant to accept. Indeed, not long before the flowering of relativity and quantum theory revolutionized physics, Michelson announced that "physical discoveries in the future are a matter of the sixth decimal place." This was a common opinion of the time. Michelson received a Nobel Prize in 1907, the first American to do so.

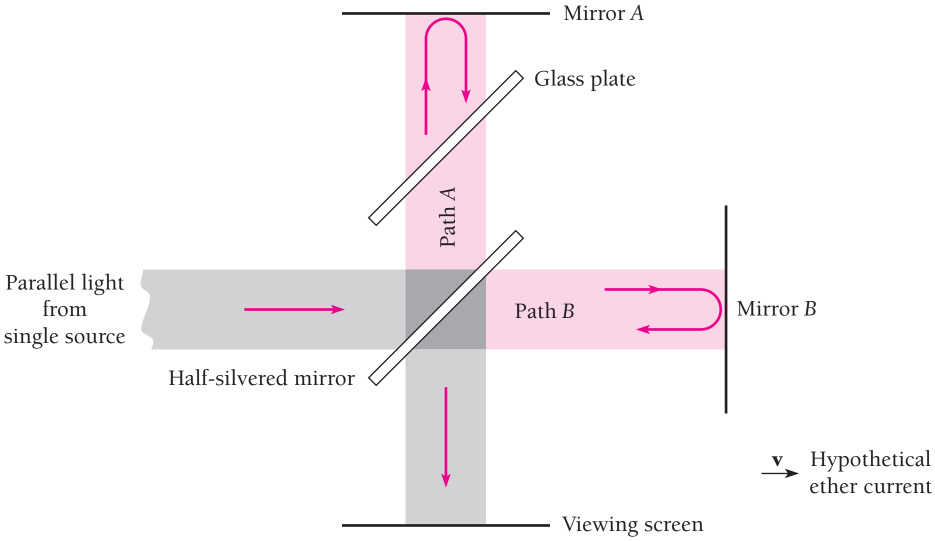


Figure 1.2 The Michelson-Morley experiment.

There is only one way to account for these results without violating the principle of relativity. It must be true that measurements of space and time are not absolute but depend on the relative motion between an observer and what is being observed. If I were to measure from the ground the rate at which your clock ticks and the length of your meter stick, I would find that the clock ticks more slowly than it did at rest on the ground and that the meter stick is shorter in the direction of motion of the spacecraft. To you, your clock and meter stick are the same as they were on the ground before you took off. To me they are different because of the relative motion, different in such a way that the speed of light you measure is the same 3×10^8 m/s I measure. Time intervals and lengths are relative quantities, but the speed of light in free space is the same to all observers.

Before Einstein's work, a conflict had existed between the principles of mechanics, which were then based on Newton's laws of motion, and those of electricity and magnetism, which had been developed into a unified theory by Maxwell. Newtonian mechanics had worked well for over two centuries. Maxwell's theory not only covered all that was then known about electric and magnetic phenomena but had also predicted that electromagnetic waves exist and identified light as an example of them. However, the equations of Newtonian mechanics and those of electromagnetism differ in the way they relate measurements made in one inertial frame with those made in a different inertial frame.

Einstein showed that Maxwell's theory is consistent with special relativity whereas Newtonian mechanics is not, and his modification of mechanics brought these branches of physics into accord. As we will find, relativistic and Newtonian mechanics agree for relative speeds much lower than the speed of light, which is why Newtonian mechanics seemed correct for so long. At higher speeds Newtonian mechanics fails and must be replaced by the relativistic version.

1.2 TIME DILATION

A moving clock ticks more slowly than a clock at rest

Measurements of time intervals are affected by relative motion between an observer and what is observed. As a result, a clock that moves with respect to an observer ticks more slowly than it does without such motion, and all processes (including those of life) occur more slowly to an observer when they take place in a different inertial frame.

If someone in a moving spacecraft finds that the time interval between two events in the spacecraft is t_0, we on the ground would find that the same interval has the longer duration t. The quantity t_0, which is determined by events that occur *at the same place* in an observer's frame of reference, is called the **proper time** of the interval between the events. When witnessed from the ground, the events that mark the beginning and end of the time interval occur at different places, and in consequence the duration of the interval appears longer than the proper time. This effect is called **time dilation** (to dilate is to become larger).

To see how time dilation comes about, let us consider two clocks, both of the particularly simple kind shown in Fig. 1.3. In each clock a pulse of light is reflected back and forth between two mirrors L_0 apart. Whenever the light strikes the lower mirror, an electric signal is produced that marks the recording tape. Each mark corresponds to the tick of an ordinary clock.

One clock is at rest in a laboratory on the ground and the other is in a spacecraft that moves at the speed $\boldsymbol{v}$ relative to the ground. An observer in the laboratory watches both clocks: does she find that they tick at the same rate?

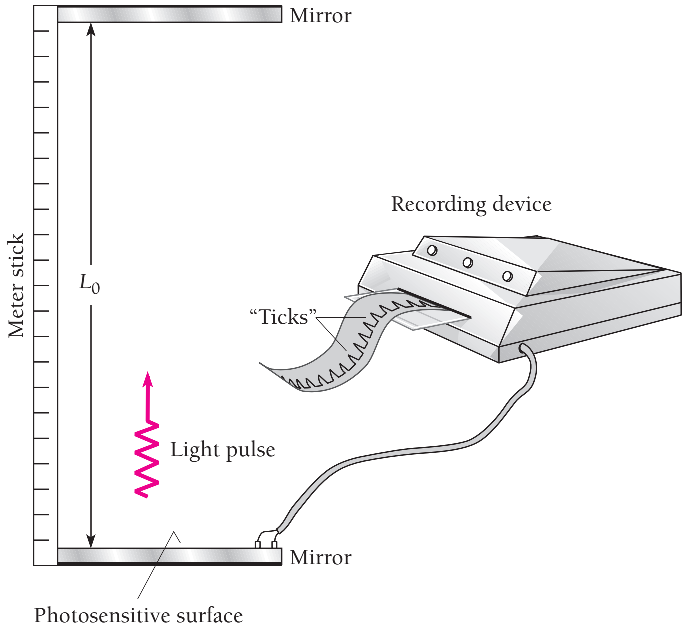


Figure 1.3 A simple clock. Each "tick" corresponds to a round trip of the light pulse from the lower mirror to the upper one and back.

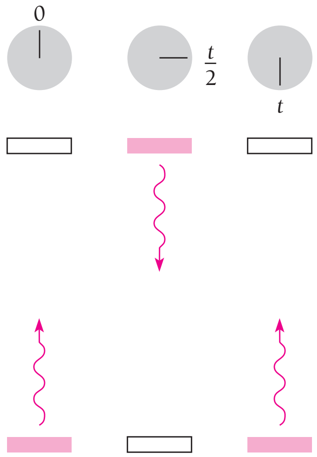


Figure 1.4 A light-pulse clock at rest on the ground as seen by an observer on the ground. The dial represents a conventional clock on the ground.

Figure 1.4 shows the laboratory clock in operation. The time interval between ticks is the proper time t_0 and the time needed for the light pulse to travel between the mirrors at the speed of light c is $t_0/2$. Hence $t_0/2 = L_0/c$ and

$$t_0 = \frac{2L_0}{c} \tag{1.1}$$

Figure 1.5 shows the moving clock with its mirrors perpendicular to the direction of motion relative to the ground. The time interval between ticks is t. Because the clock is moving, the light pulse, as seen from the ground, follows a zigzag path. On its way from the lower mirror to the upper one in the time $t/2$, the pulse travels a horizontal distance of $\boldsymbol{v}(t/2)$ and a total distance of $c(t/2)$. Since L_0 is the vertical distance between the mirrors,

$$\left(\frac{ct}{2}\right)^2 = L_0^2 + \left(\frac{\boldsymbol{v}t}{2}\right)^2$$

$$\frac{t^2}{4}(c^2 - \boldsymbol{v}^2) = L_0^2$$

$$t^2 = \frac{4L_0^2}{c^2 - \boldsymbol{v}^2} = \frac{(2L_0)^2}{c^2(1 - \boldsymbol{v}^2/c^2)}$$

$$t = \frac{2L_0/c}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.2}$$

But $2L_0/c$ is the time interval t_0 between ticks on the clock on the ground, as in Eq. (1.1), and so

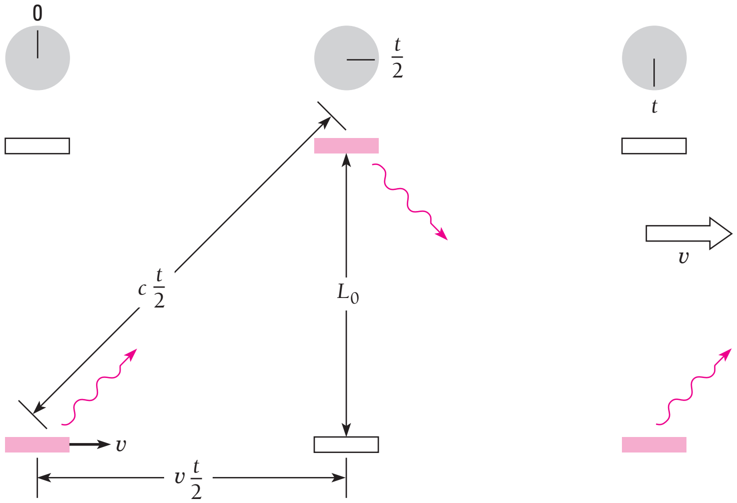


Figure 1.5 A light-pulse clock in a spacecraft as seen by an observer on the ground. The mirrors are parallel to the direction of motion of the spacecraft. The dial represents a conventional clock on the ground.

Time dilation

$$t = \frac{t_0}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.3}$$

Here is a reminder of what the symbols in Eq. (1.4) represent:

t_0 = time interval on clock at rest relative to an observer = proper time
t = time interval on clock in motion relative to an observer
$\boldsymbol{v}$ = speed of relative motion
c = speed of light

Because the quantity $\sqrt{1 - \boldsymbol{v}^2/c^2}$ is always smaller than 1 for a moving object, t is always greater than t_0. The moving clock in the spacecraft appears to tick at a slower rate than the stationary one on the ground, as seen by an observer on the ground.

Exactly the same analysis holds for measurements of the clock on the ground by the pilot of the spacecraft. To him, the light pulse of the ground clock follows a zigzag path that requires a total time t per round trip. His own clock, at rest in the spacecraft, ticks at intervals of t_0. He too finds that

$$t = \frac{t_0}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

so the effect is reciprocal: *every* observer finds that clocks in motion relative to him tick more slowly than clocks at rest relative to him.

Our discussion has been based on a somewhat unusual clock. Do the same conclusions apply to ordinary clocks that use machinery—spring-controlled escapements, tuning forks, vibrating quartz crystals, or whatever—to produce ticks at constant time intervals? The answer must be yes, since if a mirror clock and a conventional clock in the spacecraft agree with each other on the ground but not when in flight, the disagreement between then could be used to find the speed of the spacecraft independently of any outside frame of reference—which contradicts the principle that all motion is relative.

The Ultimate Speed Limit

The earth and the other planets of the solar system seem to be natural products of the evolution of the sun. Since the sun is a rather ordinary star in other ways, it is not surprising that other stars have been found to have planetary systems around them as well. Life developed here on earth, and there is no known reason why it should not also have done so on some of these planets. Can we expect ever to be able to visit them and meet our fellow citizens of the universe? The trouble is that nearly all stars are very far away—thousands or millions of light-years away. (A light-year, the distance light travels in a year, is 9.46×10^{15} m.) But if we can build a spacecraft whose speed is thousands or millions of times greater than the speed of light c, such distances would not be an obstacle.

Alas, a simple argument based on Einstein's postulates shows that nothing can move faster than c. Suppose you are in a spacecraft traveling at a constant speed $\boldsymbol{v}$ relative to the earth that is greater than c. As I watch from the earth, the lamps in the spacecraft suddenly go out. You switch on a flashlight to find the fuse box at the front of the spacecraft and change the blown fuse (Fig. 1.6*a*). The lamps go on again.

From the ground, though, I would see something quite different. To me, since your speed $\boldsymbol{v}$ is greater than c, the light from your flashlight illuminates the *back* of the spacecraft (Fig. 1.6*b*). I can only conclude that the laws of physics are different in your inertial frame from what they are in my inertial frame—which contradicts the principle of relativity. The only way to avoid this contradiction is to assume that nothing can move faster than the speed of light. This assumption has been tested experimentally many times and has always been found to be correct.

The speed of light c in relativity is always its value in free space of 3.00×10^8 m/s. In all material media, such as air, water, or glass, light travels more slowly than this, and atomic particles are able to move faster in such media than does light. When an electrically charged particle moves through a transparent substance at a speed exceeding that of light in the substance, a cone of light waves is emitted that corresponds to the bow wave produced by a ship moving through the water faster than water waves do. These light waves are known as **Cerenkov radiation** and form the basis of a method of determining the speeds of such particles. The minimum speed a particle must have to emit Cerenkov radiation is c/n in a medium whose index of refraction is n. Cerenkov radiation is visible as a bluish glow when an intense beam of particles is involved.

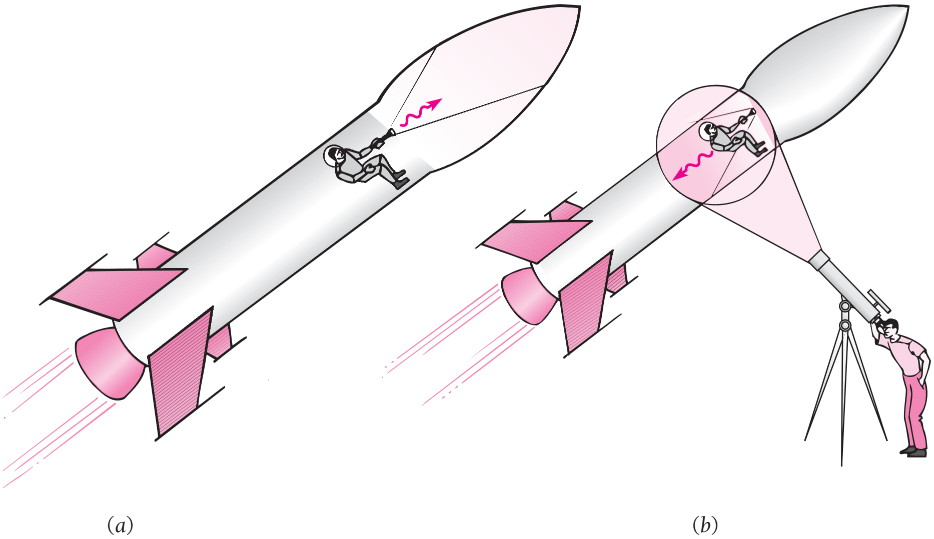
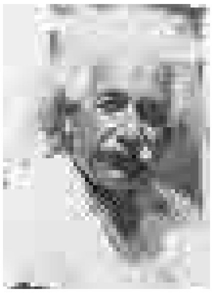


Figure 1.6 A person switches on a flashlight in a spacecraft assumed to be moving relative to the earth faster than light. (*a*) In the spacecraft frame, the light goes to the front of the spacecraft. (*b*) In the earth frame, the light goes to the back of the spacecraft. Because observers in the spacecraft and on the earth would see different events, the principle of relativity would be violated. The conclusion is that the spacecraft cannot be moving faster than light relative to the earth (or relative to anything else).



(AIP Niels Bohr Library)

Albert Einstein (1879–1955), bitterly unhappy with the rigid discipline of the schools of his native Germany, went at sixteen to Switzerland to complete his education, and later got a job examining patent applications at the Swiss Patent Office. Then, in 1905, ideas that had been germinating in his mind for years when he should have been paying attention to other matters (one of his math teachers called Einstein a "lazy dog") blossomed into three short papers that were to change decisively the course not only of physics but of modern civilization as well.

The first paper, on the photoelectric effect, proposed that light has a dual character with both particle and wave properties. The subject of the second paper was Brownian motion, the irregular zigzag movement of tiny bits of suspended matter, such as pollen grains in water. Einstein showed that Brownian motion results from the bombardment of the particles by randomly moving molecules in the fluid in which they are suspended. This provided the long-awaited definite link with experiment that convinced the remaining doubters of the molecular theory of matter. The third paper introduced the special theory of relativity.

Although much of the world of physics was originally either indifferent or skeptical, even the most unexpected of Einstein's conclusions were soon confirmed and the development of what is now called modern physics began in earnest. After university posts in Switzerland and Czechoslovakia, in 1913 he took up an appointment at the Kaiser Wilhelm Institute in Berlin that left him able to do research free of financial worries and routine duties. Einstein's interest was now mainly in gravitation, and he started where Newton had left off more than two centuries earlier.

Einstein's general theory of relativity, published in 1916, related gravity to the structure of space and time. In this theory the force of gravity can be thought of as arising from a warping of spacetime around a body of matter so that a nearby mass tends to move toward it, much as a marble rolls toward the bottom of a saucer-shaped hole. From general relativity came a number of remarkable predictions, such as that light should be subject to gravity, all of which were verified experimentally. The later discovery that the universe is expanding fit neatly into the theory. In 1917 Einstein introduced the idea of stimulated emission of radiation, an idea that bore fruit forty years later in the invention of the laser.

The development of quantum mechanics in the 1920s disturbed Einstein, who never accepted its probabilistic rather than deterministic view of events on an atomic scale. "God does not play dice with the world," he said, but for once his physical intuition seemed to be leading him in the wrong direction.

Einstein, by now a world celebrity, left Germany in 1933 after Hitler came to power and spent the rest of his life at the Institute for Advanced Study in Princeton, New Jersey, thereby escaping the fate of millions of other European Jews at the hands of the Germans. His last years were spent in an unsuccessful search for a theory that would bring gravitation and electromagnetism together into a single picture, a problem worthy of his gifts but one that remains unsolved to this day.

Example 1.1

A spacecraft is moving relative to the earth. An observer on the earth finds that, between 1 P.M. and 2 P.M. according to her clock, 3601 s elapse on the spacecraft's clock. What is the spacecraft's speed relative to the earth?

Solution

Here $t_0 = 3600$ s is the proper time interval on the earth and $t = 3601$ s is the time interval in the moving frame as measured from the earth. We proceed as follows:

$$t = \frac{t_0}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

$$1 - \frac{\boldsymbol{v}^2}{c^2} = \left(\frac{t_0}{t}\right)^2$$

$$\boldsymbol{v} = c\sqrt{1 - \left(\frac{t_0}{t}\right)^2} = (2.998 \times 10^8 \text{ m/s})\sqrt{1 - \left(\frac{3600 \text{ s}}{3601 \text{ s}}\right)^2}$$

$$= 7.1 \times 10^6 \text{ m/s}$$

Today's spacecraft are much slower than this. For instance, the highest speed of the Apollo 11 spacecraft that went to the moon was only 10,840 m/s, and its clocks differed from those on the earth by less than one part in 10^9. Most of the experiments that have confirmed time dilation made use of unstable nuclei and elementary particles which readily attain speeds not far from that of light.

Apollo 11 lifts off its pad to begin the first human visit to the moon. At its highest speed of 10.8 km/s relative to the earth, its clocks differed from those on the earth by less than one part in a billion.

Although time is a relative quantity, not all the notions of time formed by everyday experience are incorrect. Time does not run backward to *any* observer, for instance. A sequence of events that occur at some particular point at t_1, t_2, t_3, . . . will appear in the same order to all observers everywhere, though not necessarily with the same time intervals $t_2 - t_1$, $t_3 - t_2$, . . . between each pair of events. Similarly, no distant observer, regardless of his or her state of motion, can see an event before it happens—more precisely, before a nearby observer sees it—since the speed of light is finite and signals require the minimum period of time L/c to travel a distance L. There is no way to peer into the future, although past events may appear different to different observers.

1.3 DOPPLER EFFECT

Why the universe is believed to be expanding

We are all familiar with the increase in pitch of a sound when its source approaches us (or we approach the source) and the decrease in pitch when the source recedes from us (or we recede from the source). These changes in frequency constitute the **doppler effect,** whose origin is straightforward. For instance, successive waves emitted by a source moving toward an observer are closer together than normal because of the advance of the source; because the separation of the waves is the wavelength of the sound, the corresponding frequency is higher. The relationship between the source frequency ν_0 and the observed frequency ν is

Doppler effect in sound

$$\nu = \nu_0\left(\frac{1 + v/c}{1 - V/c}\right) \tag{1.4}$$

where c = speed of sound
v = speed of observer (+ for motion toward the source, − for motion away from it)
V = speed of the source (+ for motion toward the observer, − for motion away from him)

If the observer is stationary, $v = 0$, and if the source is stationary, $V = 0$.

The doppler effect in sound varies depending on whether the source, or the observer, or both are moving. This appears to violate the principle of relativity: all that should count is the relative motion of source and observer. But sound waves occur only in a material medium such as air or water, and this medium is itself a frame of reference with respect to which motions of source and observer are measurable. Hence there is no contradiction. In the case of light, however, no medium is involved and only relative motion of source and observer is meaningful. The doppler effect in light must therefore differ from that in sound.

We can analyze the doppler effect in light by considering a light source as a clock that ticks ν_0 times per second and emits a wave of light with each tick. We will examine the three situations shown in Fig. 1.7.

1 *Observer moving perpendicular to a line between him and the light source*. The proper time between ticks is $t_0 = 1/\nu_0$, so between one tick and the next the time $t = t_0/\sqrt{1 - v^2/c^2}$ elapses in the reference frame of the observer. The frequency he finds is accordingly

$$\nu(\text{transverse}) = \frac{1}{t} = \frac{\sqrt{1 - v^2/c^2}}{t_0}$$

Transverse doppler effect in light

$$\nu = \nu_0\sqrt{1 - v^2/c^2} \tag{1.5}$$

The observed frequency ν is always lower than the source frequency ν_0.

2 *Observer receding from the light source*. Now the observer travels the distance vt away from the source between ticks, which means that the light wave from a given tick takes

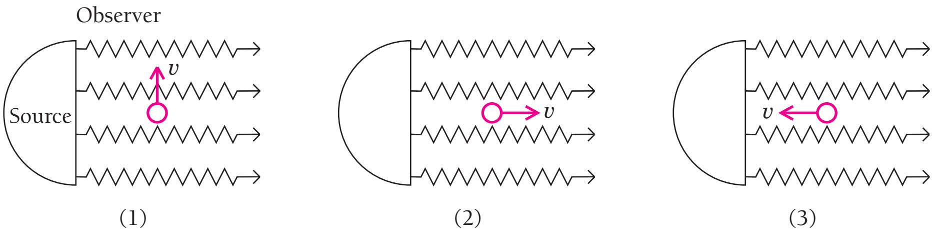


Figure 1.7 The frequency of the light seen by an observer depends on the direction and speed of the observer's motion relative to its source.

$\boldsymbol{v}t/c$ longer to reach him than the previous one. Hence the total time between the arrival of successive waves is

$$T = t + \frac{\boldsymbol{v}t}{c} = t_0 \frac{1 + \boldsymbol{v}/c}{\sqrt{1 - \boldsymbol{v}^2/c^2}} = t_0 \frac{\sqrt{1 + \boldsymbol{v}/c}\sqrt{1 + \boldsymbol{v}/c}}{\sqrt{1 + \boldsymbol{v}/c}\sqrt{1 - \boldsymbol{v}/c}} = t_0 \sqrt{\frac{1 + \boldsymbol{v}/c}{1 - \boldsymbol{v}/c}}$$

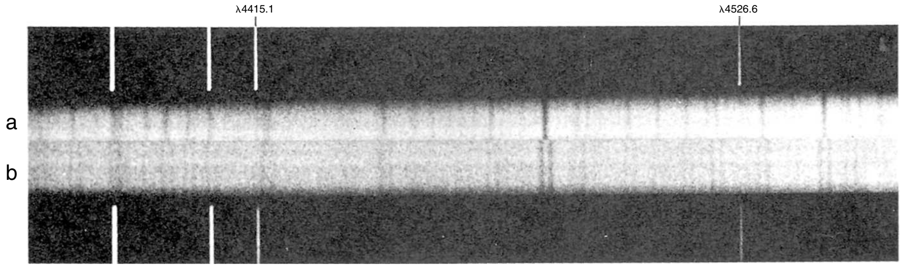
and the observed frequency is

$$\nu(\text{receding}) = \frac{1}{T} = \frac{1}{t_0}\sqrt{\frac{1 - \boldsymbol{v}/c}{1 + \boldsymbol{v}/c}} = \nu_0 \sqrt{\frac{1 - \boldsymbol{v}/c}{1 + \boldsymbol{v}/c}} \tag{1.6}$$

The observed frequency ν is lower than the source frequency ν_0. Unlike the case of sound waves, which propagate relative to a material medium it makes no difference whether the observer is moving away from the source or the source is moving away from the observer.

3 *Observer approaching the light source*. The observer here travels the distance $\boldsymbol{v}t$ toward the source between ticks, so each light wave takes $\boldsymbol{v}t/c$ less time to arrive than the previous one. In this case $T = t - \boldsymbol{v}t/c$ and the result is

$$\nu(\text{approaching}) = \nu_0 \sqrt{\frac{1 + \boldsymbol{v}/c}{1 - \boldsymbol{v}/c}} \tag{1.7}$$



Spectra of the double star Mizar, which consists of two stars that circle their center of mass, taken 2 days apart. In *a* the stars are in line with no motion toward or away from the earth, so their spectral lines are superimposed. In *b* one star is moving toward the earth and the other is moving away from the earth, so the spectral lines of the former are doppler-shifted toward the blue end of the spectrum and those of the latter are shifted toward the red end.

The observed frequency is higher than the source frequency. Again, the same formula holds for motion of the source toward the observer.

Equations (1.6) and (1.7) can be combined in the single formula

Longitudinal doppler effect in light

$$\nu = \nu_0 \sqrt{\frac{1 + \boldsymbol{v}/c}{1 - \boldsymbol{v}/c}} \tag{1.8}$$

by adopting the convention that $\boldsymbol{v}$ is + for source and observer approaching each other and − for source and observer receding from each other.

Example 1.2

A driver is caught going through a red light. The driver claims to the judge that the color she actually saw was green ($\nu = 5.60 \times 10^{14}$ Hz) and not red ($\nu_0 = 4.80 \times 10^{14}$ Hz) because of the doppler effect. The judge accepts this explanation and instead fines her for speeding at the rate of $1 for each km/h she exceeded the speed limit of 80 km/h. What was the fine?

Solution

Solving Eq. (1.8) for v gives

$$v = c\left(\frac{\nu^2 - \nu_0^2}{\nu^2 + \nu_0^2}\right) = (3.00 \times 10^8 \text{ m/s})\left[\frac{(5.60)^2 - (4.80)^2}{(5.60)^2 + (4.80)^2}\right]$$
$$= 4.59 \times 10^7 \text{ m/s} = 1.65 \times 10^8 \text{ km/h}$$

since 1 m/s = 3.6 km/h. The fine is therefore $\$(1.65 \times 10^8 - 80) = \$164{,}999{,}920$.

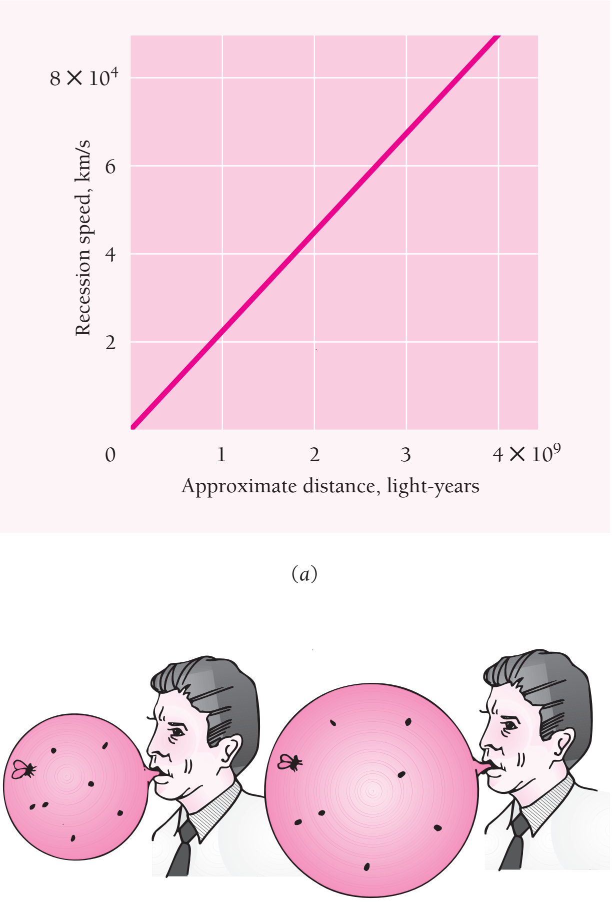
Visible light consists of electromagnetic waves in a frequency band to which the eye is sensitive. Other electromagnetic waves, such as those used in radar and in radio communications, also exhibit the doppler effect in accord with Eq. (1.8). Doppler shifts in radar waves are used by police to measure vehicle speeds, and doppler shifts in the radio waves emitted by a set of earth satellites formed the basis of the highly accurate Transit system of marine navigation.

The Expanding Universe

The doppler effect in light is an important tool in astronomy. Stars emit light of certain characteristic frequencies called spectral lines, and motion of a star toward or away from the earth shows up as a doppler shift in these frequencies. The spectral lines of distant galaxies of stars are all shifted toward the low-frequency (red) end of the spectrum and hence are called "red shifts." Such shifts indicate that the galaxies are receding from us and from one another. The speeds of recession are observed to be

Edwin Hubble (1889–1953) was born in Missouri and, although always interested in astronomy, pursued a variety of other subjects as well at the University of Chicago. He then went as a Rhodes Scholar to Oxford University in England where he concentrated on law, Spanish, and heavyweight boxing. After two years of teaching at an Indiana high school, Hubble realized what his true vocation was and returned to the University of Chicago to study astronomy.

At Mt. Wilson Observatory in California, Hubble made the first accurate measurements of the distances of spiral galaxies which showed that they are far away in space from our own Milky Way galaxy. It had been known for some time that such galaxies have red shifts in their spectra that indicate motion away from the Milky Way, and Hubble joined his distance figures with the observed red shifts to conclude that the recession speeds were proportional to distance. This implies that the universe is expanding, a remarkable discovery that has led to the modern picture of the universe. Hubble was the first to use the 200-inch telescope, for many years the world's largest, at Mt. Palomar in California, in 1949. In his later work Hubble tried to determine the structure of the universe by finding how the concentration of remote galaxies varies with distance, a very difficult task that only today is being accomplished.



(*a*)

(*b*)

Figure 1.8 (*a*) Graph of recession speed versus distance for distant galaxies. The speed of recession averages about 21 km/s per million light-years. (*b*) Two-dimensional analogy of the expanding universe. As the balloon is inflated, the spots on it become farther apart. A bug on the balloon would find that the farther away a spot is from its location, the faster the spot seems to be moving away; this is true no matter where the bug is. In the case of the universe, the more distant a galaxy is from us, the faster it is moving away, which means that the universe is expanding uniformly.

proportional to distance, which suggests that the entire universe is expanding (Fig. 1.8). This proportionality is called **Hubble's law.**

The expansion apparently began about 13 billion years ago when a very small, intensely hot mass of primeval matter exploded, an event usually called the **Big Bang.** As described in Chap. 13, the matter soon turned into the electrons, protons, and neutrons of which the present universe is composed. Individual aggregates that formed during the expansion became the galaxies of today. Present data suggest that the current expansion will continue forever.

Example 1.3

A distant galaxy in the constellation Hydra is receding from the earth at 6.12×10^7 m/s. By how much is a green spectral line of wavelength 500 nm (1 nm = 10^{-9} m) emitted by this galaxy shifted toward the red end of the spectrum?

Solution

Since $\lambda = c/\nu$ and $\lambda_0 = c/\nu_0$, from Eq. (1.6) we have

$$\lambda = \lambda_0 \sqrt{\frac{1 + v/c}{1 - v/c}}$$

Here $v = 0.204c$ and $\lambda_0 = 500$ nm, so

$$\lambda = 500 \text{ nm} \sqrt{\frac{1 + 0.204}{1 - 0.204}} = 615 \text{ nm}$$

which is in the orange part of the spectrum. The shift is $\lambda - \lambda_0 = 115$ nm. This galaxy is believed to be 2.9 billion light-years away.

1.4 LENGTH CONTRACTION

Faster means shorter

Measurements of lengths as well as of time intervals are affected by relative motion. The length L of an object in motion with respect to an observer always appears to the observer to be shorter than its length L_0 when it is at rest with respect to him. This contraction occurs only in the direction of the relative motion. The length L_0 of an object in its rest frame is called its **proper length.** (We note that in Fig. 1.5 the clock is moving perpendicular to $\mathbf{v}$, hence $L = L_0$ there.)

The length contraction can be derived in a number of ways. Perhaps the simplest is based on time dilation and the principle of relativity. Let us consider what happens to unstable particles called muons that are created at high altitudes by fast cosmic-ray particles (largely protons) from space when they collide with atomic nuclei in the earth's atmosphere. A muon has a mass 207 times that of the electron and has a charge of either $+e$ or $-e$; it decays into an electron or a positron after an average lifetime of 2.2 μs (2.2×10^{-6} s).

Cosmic-ray muons have speeds of about 2.994×10^8 m/s ($0.998c$) and reach sea level in profusion—one of them passes through each square centimeter of the earth's surface on the average slightly more often than once a minute. But in $t_0 = 2.2$ μs, their average lifetime, muons can travel a distance of only

$$vt_0 = (2.994 \times 10^8 \text{ m/s})(2.2 \times 10^{-6} \text{ s}) = 6.6 \times 10^2 \text{ m} = 0.66 \text{ km}$$

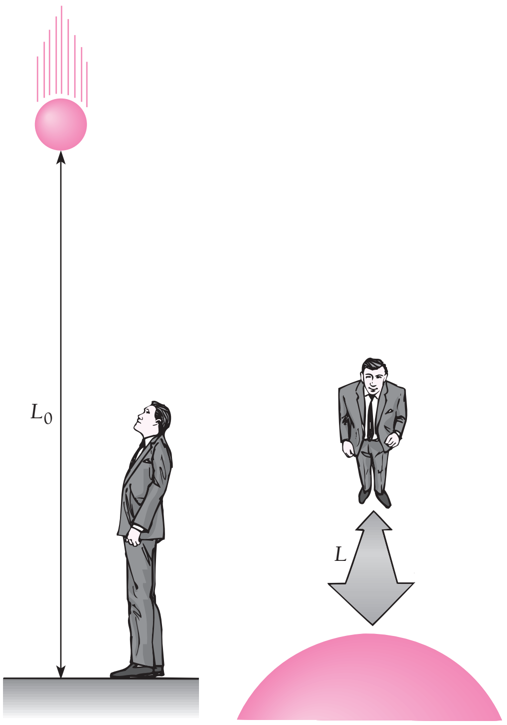
before decaying, whereas they are actually created at altitudes of 6 km or more.

To resolve the paradox, we note that the muon lifetime of $t_0 = 2.2$ μs is what an observer at rest with respect to a muon would find. Because the muons are hurtling toward us at the considerable speed of $0.998c$, their lifetimes are extended in our frame of reference by time dilation to

$$t = \frac{t_0}{\sqrt{1 - v^2/c^2}} = \frac{2.2 \times 10^{-6} \text{ s}}{\sqrt{1 - (0.998c)^2/c^2}} = 34.8 \times 10^{-6} \text{ s} = 34.8 \text{ } \mu\text{s}$$

The moving muons have lifetimes almost 16 times longer than those at rest. In a time interval of 34.8 μs, a muon whose speed is $0.998c$ can cover the distance

$$vt = (2.994 \times 10^8 \text{ m/s})(34.8 \times 10^{-6} \text{ s}) = 1.04 \times 10^4 \text{ m} = 10.4 \text{ km}$$



As found by observer on the ground, the muon altitude is L_0.

As found by an observer moving with the muon, the ground is L below it, which is a shorter distance than L_0.

Figure 1.9 Muon decay as seen by different observers. The muon size is greatly exaggerated here; in fact, the muon seems likely to be a point particle with no extension in space.

Although its lifetime is only $t_0 = 2.2\ \mu\text{s}$ in its own frame of reference, a muon can reach the ground from altitudes of as much as 10.4 km because in the frame in which these altitudes are measured, the muon lifetime is $t = 34.8\ \mu\text{s}$.

What if somebody were to accompany a muon in its descent at $\boldsymbol{v} = 0.998c$, so that to him or her the muon is at rest? The observer and the muon are now in the same frame of reference, and in this frame the muon's lifetime is only 2.2 μs. To the observer, the muon can travel only 0.66 km before decaying. The only way to account for the arrival of the muon at ground level is if the distance it travels, from the point of view of an observer in the moving frame, is shortened by virtue of its motion (Fig. 1.9). The principle of relativity tells us the extent of the shortening—it must be by the same factor of $\sqrt{1 - \boldsymbol{v}^2/c^2}$ that the muon lifetime is extended from the point of view of a stationary observer.

We therefore conclude that an altitude we on the ground find to be h_0 must appear in the muon's frame of reference as the lower altitude

$$h = h_0 \sqrt{1 - \boldsymbol{v}^2/c^2}$$

In our frame of reference the muon can travel $h_0 = 10.4$ km because of time dilation. In the muon's frame of reference, where there is no time dilation, this distance is abbreviated to

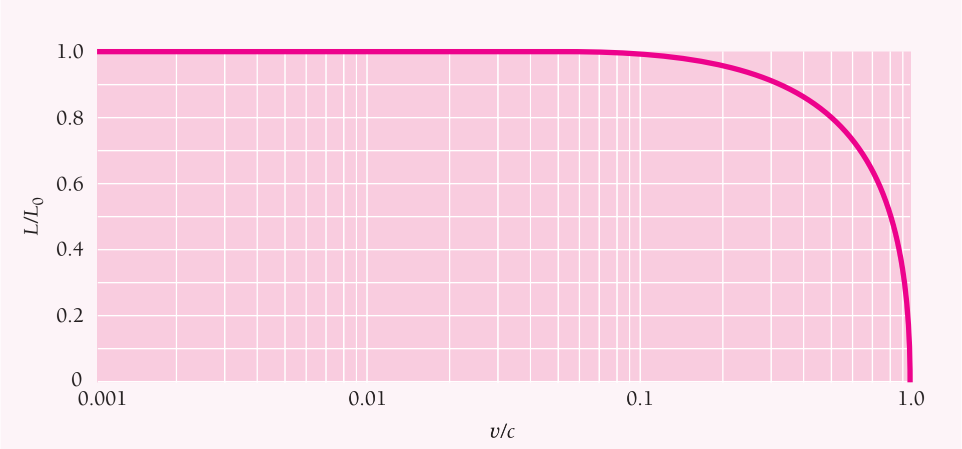


Figure 1.10 Relativistic length contraction. Only lengths in the direction of motion are affected. The horizontal scale is logarithmic.

$$h = (10.4 \text{ km}) \sqrt{1 - (0.998c)^2/c^2} = 0.66 \text{ km}$$

As we know, a muon traveling at $0.998c$ goes this far in 2.2 μs.

The relativistic shortening of distances is an example of the general contraction of lengths in the direction of motion:

Length contraction

$$L = L_0 \sqrt{1 - v^2/c^2} \tag{1.9}$$

Figure 1.10 is a graph of L/L_0 versus v/c. Clearly the length contraction is most significant at speeds near that of light. A speed of 1000 km/s seems fast to us, but it only results in a shortening in the direction of motion to 99.9994 percent of the proper length of an object moving at this speed. On the other hand, something traveling at nine-tenths the speed of light is shortened to 44 percent of its proper length, a significant change.

Like time dilation, the length contraction is a reciprocal effect. To a person in a spacecraft, objects on the earth appear shorter than they did when he or she was on the ground by the same factor of $\sqrt{1 - v^2/c^2}$ that the spacecraft appears shorter to somebody at rest. The proper length L_0 found in the rest frame is the maximum length any observer will measure. As mentioned earlier, only lengths in the direction of motion undergo contraction. Thus to an outside observer a spacecraft is shorter in flight than on the ground, but it is not narrower.

1.5 TWIN PARADOX

A longer life, but it will not seem longer

We are now in a position to understand the famous relativistic effect known as the twin paradox. This paradox involves two identical clocks, one of which remains on the earth while the other is taken on a voyage into space at the speed v and eventually is brought back. It is customary to replace the clocks with the pair of twins Dick and

Jane, a substitution that is perfectly acceptable because the processes of life—heartbeats, respiration, and so on—constitute biological clocks of reasonable regularity.

Dick is 20 y old when he takes off on a space voyage at a speed of 0.80c to a star 20 light-years away. To Jane, who stays behind, the pace of Dick's life is slower than hers by a factor of

$$\sqrt{1 - \boldsymbol{v}^2/c^2} = \sqrt{1 - (0.80c)^2/c^2} = 0.60 = 60\%$$

To Jane, Dick's heart beats only 3 times for every 5 beats of her heart; Dick takes only 3 breaths for every 5 of hers; Dick thinks only 3 thoughts for every 5 of hers. Finally Dick returns after 50 years have gone by according to Jane's calendar, but to Dick the trip has taken only 30 y. Dick is therefore 50 y old whereas Jane, the twin who stayed home, is 70 y old (Fig. 1.11).

Where is the paradox? If we consider the situation from the point of view of Dick in the spacecraft, Jane on the earth is in motion relative to him at a speed of 0.80c. Should not Jane then be 50 y old when the spacecraft returns, while Dick is then 70—the precise opposite of what was concluded above?

But the two situations are not equivalent. Dick changed from one inertial frame to a different one when he started out, when he reversed direction to head home, and when he landed on the earth. Jane, however, remained in the same inertial frame during Dick's whole voyage. The time dilation formula applies to Jane's observations of Dick, but not to Dick's observations of her.

To look at Dick's voyage from his perspective, we must take into account that the distance L he covers is shortened to

$$L = L_0\sqrt{1 - \boldsymbol{v}^2/c^2} = (20 \text{ light-years})\sqrt{1 - (0.80c)^2/c^2} = 12 \text{ light-years}$$

To Dick, time goes by at the usual rate, but his voyage to the star has taken $L/\boldsymbol{v} = 15$ y and his return voyage another 15 y, for a total of 30 y. Of course, Dick's life span has

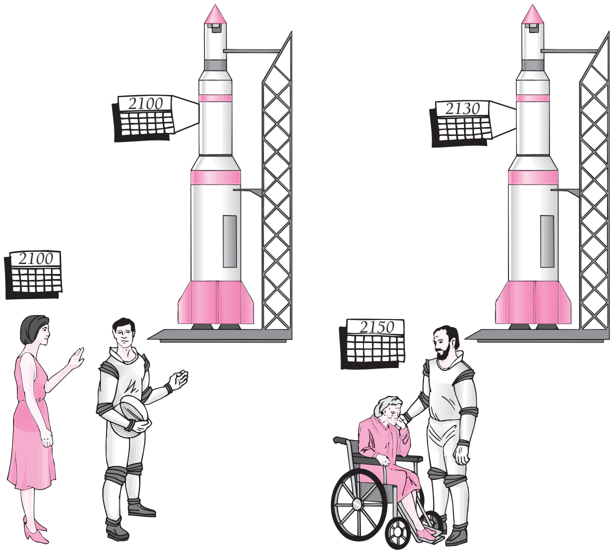


Figure 1.11 An astronaut who returns from a space voyage will be younger than his or her twin who remains on earth. Speeds close to the speed of light (here $\boldsymbol{v} = 0.8c$) are needed for this effect to be conspicuous.

not been extended *to him,* because regardless of Jane's 50-y wait, he has spent only 30 y on the roundtrip.

The nonsymmetric aging of the twins has been verified by experiments in which accurate clocks were taken on an airplane trip around the world and then compared with identical clocks that had been left behind. An observer who departs from an inertial system and then returns after moving relative to that system will always find his or her clocks slow compared with clocks that stayed in the system.

Example 1.4

Dick and Jane each send out a radio signal once a year while Dick is away. How many signals does Dick receive? How many does Jane receive?

Solution

On the outward trip, Dick and Jane are being separated at a rate of $0.80c$. With the help of the reasoning used to analyze the doppler effect in Sec. 1.3, we find that each twin receives signals

$$T_1 = t_0 \sqrt{\frac{1 + v/c}{1 - v/c}} = (1 \text{ y}) \sqrt{\frac{1 + 0.80}{1 - 0.80}} = 3 \text{ y}$$

apart. On the return trip, Dick and Jane are getting closer together at the same rate, and each receives signals more frequently, namely

$$T_2 = t_0 \sqrt{\frac{1 - v/c}{1 + v/c}} = (1 \text{ y}) \sqrt{\frac{1 - 0.80}{1 + 0.80}} = \frac{1}{3} \text{ y}$$

apart.

To Dick, the trip to the star takes 15 y, and he receives $15/3 = 5$ signals from Jane. During the 15 y of the return trip, Dick receives $15/(1/3) = 45$ signals from Jane, for a total of 50 signals. Dick therefore concludes that Jane has aged by 50 y in his absence. Both Dick and Jane agree that Jane is 70 y old at the end of the voyage.

To Jane, Dick needs $L_0/v = 25$ y for the outward trip. Because the star is 20 light-years away. Jane on the earth continues to receive Dick's signals at the original rate of one every 3 y for 20 y after Dick has arrived at the star. Hence Jane receives signals every 3 y for 25 y + 20 y = 45 y to give a total of $45/3 = 15$ signals. (These are the 15 signals Dick sent out on the outward trip.) Then, for the remaining 5 y of what is to Jane a 50-y voyage, signals arrive from Dick at the shorter intervals of $1/3$ y for an additional $5/(1/3) = 15$ signals. Jane thus receives 30 signals in all and concludes that Dick has aged by 30 y during the time he was away—which agrees with Dick's own figure. Dick is indeed 20 y younger than his twin Jane on his return.

1.6 ELECTRICITY AND MAGNETISM

Relativity is the bridge

One of the puzzles that set Einstein on the trail of special relativity was the connection between electricity and magnetism, and the ability of his theory to clarify the nature of this connection is one of its triumphs.

Because the moving charges (usually electrons) whose interactions give rise to many of the magnetic forces familiar to us have speeds far smaller than c, it is not obvious that the operation of an electric motor, say, is based on a relativistic effect. The idea becomes less implausible, however, when we reflect on the strength of electric forces. The electric attraction between the electron and proton in a hydrogen atom, for instance, is 10^{39} times greater than the gravitational attraction between them. Thus even a small change in the character of these forces due to relative motion, which is what magnetic forces represent, may have large consequences. Furthermore, although the effective speed of an individual electron in a current-carrying wire (<1 mm/s) is less than that of a tired caterpillar, there may be 10^{20} or more moving electrons per centimeter in such a wire, so the total effect may be considerable.

Although the full story of how relativity links electricity and magnetism is mathematically complex, some aspects of it are easy to appreciate. An example is the origin of the magnetic force between two parallel currents. An important point is that, like the speed of light,

Electric charge is relativistically invariant.

A charge whose magnitude is found to be Q in one frame of reference is also Q in all other frames.

Let us look at the two idealized conductors shown in Fig. 1.12*a*. They contain equal numbers of positive and negative charges at rest that are equally spaced. Because the conductors are electrically neutral, there is no force between them.

Figure 1.12*b* shows the same conductors when they carry currents i_{I} and i_{II} in the same direction. The positive charges move to the right and the negative charges move to the left, both at the same speed $\boldsymbol{v}$ as seen from the laboratory frame of reference. (Actual currents in metals consist of flows of negative electrons only, of course, but the electrically equivalent model here is easier to analyze and the results are the same.) Because the charges are moving, their spacing is smaller than before by the factor $\sqrt{1 - \boldsymbol{v}^2/c^2}$. Since $\boldsymbol{v}$ is the same for both sets of charges, their spacings shrink by the same amounts, and both conductors remain neutral to an observer in the laboratory. However, the conductors now attract each other. Why?

Let us look at conductor II from the frame of reference of one of the negative charges in conductor I. Because the negative charges in II appear at rest in this frame, their spacing is not contracted, as in Fig. 1.12*c*. On the other hand, the positive charges in II now have the velocity $2\boldsymbol{v}$, and their spacing is accordingly contracted to a greater extent than they are in the laboratory frame. Conductor II therefore appears to have a net positive charge, and an attractive force acts on the negative charge in I.

Next we look at conductor II from the frame of reference of one of the positive charges in conductor I. The positive charges in II are now at rest, and the negative charges there move to the left at the speed $2\boldsymbol{v}$. Hence the negative charges are closer together than the positive ones, as in Fig. 1.12*d*, and the entire conductor appears negatively charged. An attractive force therefore acts on the positive charges in I.

Identical arguments show that the negative and positive charges in II are attracted to I. Thus all the charges in each conductor experience forces directed toward the other conductor. To each charge, the force on it is an "ordinary" electric force that arises because the charges of opposite sign in the other conductor are closer together than the charges of the same sign, so the other conductor appears to have a net charge. From the laboratory frame the situation is less straightforward. Both conductors are electrically neutral in this frame, and it is natural to explain their mutual attraction by attributing it to a special "magnetic" interaction between the currents.

A similar analysis explains the repulsive force between parallel conductors that carry currents in opposite directions. Although it is convenient to think of magnetic forces as being different from electric ones, they both result from a single electromagnetic interaction that occurs between charged particles.

Clearly a current-carrying conductor that is electrically neutral in one frame of reference might not be neutral in another frame. How can this observation be reconciled

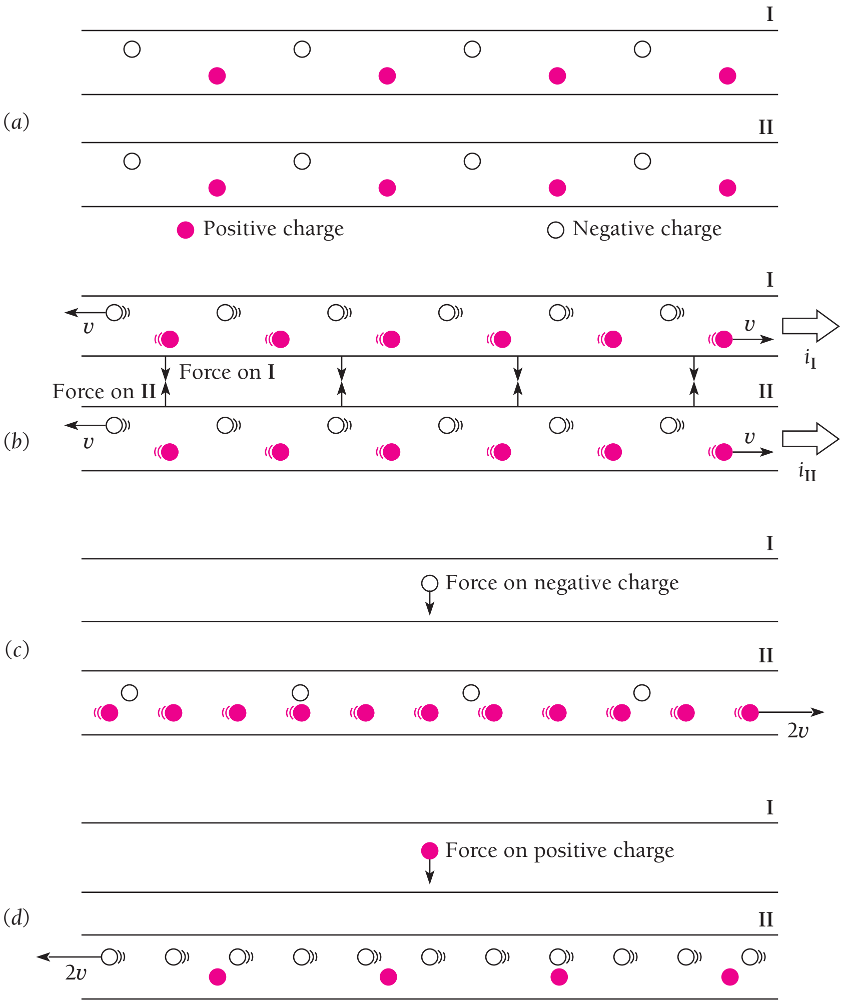


Figure 1.12 How the magnetic attraction between parallel currents arises. (*a*) Idealized parallel conductors that contain equal numbers of positive and negative charges. (*b*) When the conductors carry currents, the spacing of their moving charges undergoes a relativistic contraction as seen from the laboratory. The conductors attract each other when i_I and i_{II} are in the same direction. (*c*) As seen by a negative charge in I, the negative charges in II are at rest whereas the positive charges are in motion. The contracted spacing of the latter leads to a net positive charge in II that attracts the negative charge in I. (*d*) As seen by a positive charges in I, the positive charges in II are at rest whereas the negative charges are in motion. The contracted spacing of the latter leads to a net negative charge on II that attrats the positive charge in I. The contracted spacings in *b*, *c*, and *d* are greatly exaggerated.

with charge invariance? The answer is that we must consider the entire circuit of which the conductor is a part. Because the circuit must be closed for a current to occur in it, for every current element in one direction that a moving observer finds to have, say, a positive charge, there must be another current element in the opposite direction which the same observer finds to have a negative charge. Hence magnetic forces always act between different parts of the same circuit, even though the circuit as a whole appears electrically neutral to all observers.

The preceding discussion considered only a particular magnetic effect. All other magnetic phenomena can also be interpreted on the basis of Coulomb's law, charge invariance, and special relativity, although the analysis is usually more complicated.

1.7 RELATIVISTIC MOMENTUM

Redefining an important quantity

In classical mechanics linear momentum $\mathbf{p} = m\mathbf{v}$ is a useful quantity because it is conserved in a system of particles not acted upon by outside forces. When an event such as a collision or an explosion occurs inside an isolated system, the vector sum of the momenta of its particles before the event is equal to their vector sum afterward. We now have to ask whether $\mathbf{p} = m\mathbf{v}$ is valid as the definition of momentum in inertial frames in relative motion, and if not, what a relativistically correct definition is.

To start with, we require that $\mathbf{p}$ be conserved in a collision for all observers in relative motion at constant velocity. Also, we know that $\mathbf{p} = m\mathbf{v}$ holds in classical mechanics, that is, for $v \ll c$. Whatever the relativistically correct $\mathbf{p}$ is, then, it must reduce to $m\mathbf{v}$ for such velocities.

Let us consider an elastic collision (that is, a collision in which kinetic energy is conserved) between two particles A and B, as witnessed by observers in the reference frames S and S' which are in uniform relative motion. The properties of A and B are identical when determined in reference frames in which they are at rest. The frames S and S' are oriented as in Fig. 1.13, with S' moving in the $+x$ direction with respect to S at the velocity $\mathbf{v}$.

Before the collision, particle A had been at rest in frame S and particle B in frame S'. Then, at the same instant, A was thrown in the $+y$ direction at the speed V_A while B was thrown in the $-y'$ direction at the speed V'_B, where

$$V_A = V'_B \tag{1.10}$$

Hence the behavior of A as seen from S is exactly the same as the behavior of B as seen from S'.

When the two particles collide, A rebounds in the $-y$ direction at the speed V_A, while B rebounds in the $+y'$ direction at the speed V'_B. If the particles are thrown from positions Y apart, an observer in S finds that the collision occurs at $y = \frac{1}{2}Y$ and one in S' finds that it occurs at $y' = y = \frac{1}{2}Y$. The round-trip time T_0 for A as measured in frame S is therefore

$$T_0 = \frac{Y}{V_A} \tag{1.11}$$

and it is the same for B in S':

$$T_0 = \frac{Y}{V'_B}$$

In S the speed V_B is found from

$$V_B = \frac{Y}{T} \tag{1.12}$$

where T is the time required for B to make its round trip *as measured in* S. In S', however, B's trip requires the time T_0, where

$$T = \frac{T_0}{\sqrt{1 - v^2/c^2}} \tag{1.13}$$

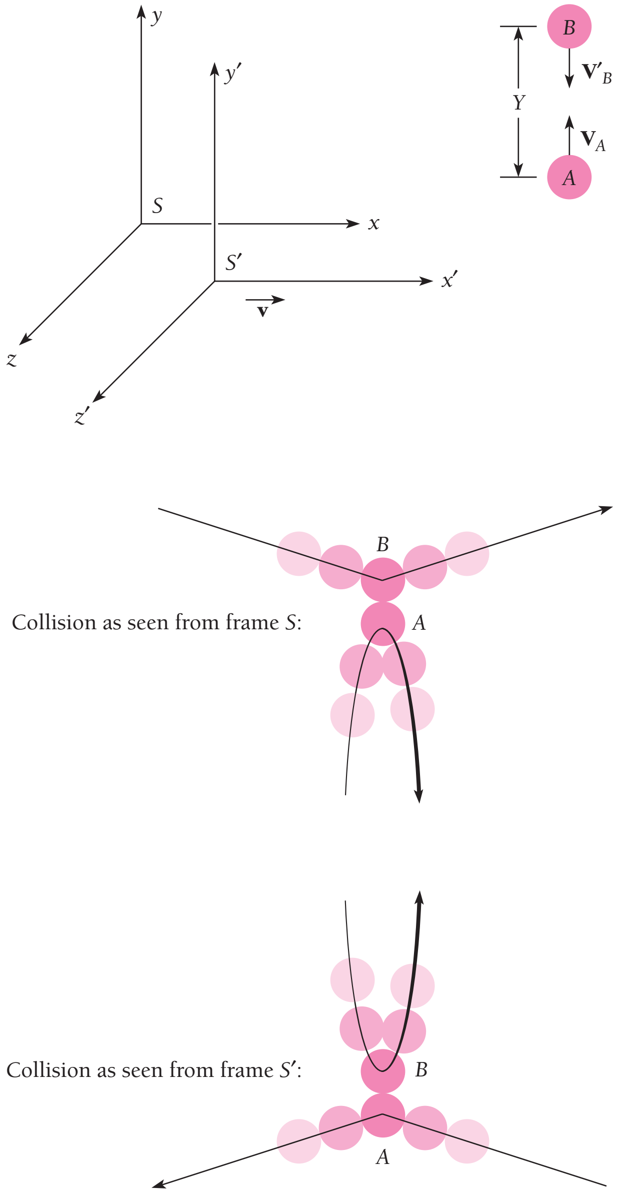

y
y′
S
x
S′
x′
v
z
z′
B
V′_B
Y
V_A
A
B
Collision as seen from frame S:
A
Collision as seen from frame S′:
B
A

Figure 1.13 An elastic collision as observed in two different frames of reference. The balls are initially Y apart, which is the same distance in both frames since S' moves only in the x direction.

according to our previous results. Although observers in both frames see the same event, they disagree about the length of time the particle thrown from the other frame requires to make the collision and return.

Replacing T in Eq. (1.12) with its equivalent in terms of T_0, we have

$$V_B = \frac{Y\sqrt{1 - v^2/c^2}}{T_0}$$

From Eq. (1.11), $$V_A = \frac{Y}{T_0}$$

If we use the classical definition of momentum, $\mathbf{p} = m\mathbf{v}$, then in frame S

$$p_A = m_A V_A = m_A\left(\frac{Y}{T_0}\right)$$

$$p_B = m_B V_B = m_B\sqrt{1 - v^2/c^2}\left(\frac{Y}{T_0}\right)$$

This means that, in this frame, momentum will not be conserved if $m_A = m_B$, where m_A and m_B are the masses as measured in S. However, if

$$m_B = \frac{m_A}{\sqrt{1 - v^2/c^2}} \tag{1.14}$$

then momentum *will* be conserved.

In the collision of Fig. 1.13 both A and B are moving in both frames. Suppose now that V_A and V'_B are very small compared with v, the relative velocity of the two frames. In this case an observer in S will see B approach A with the velocity v, make a glancing collision (since $V'_B \ll v$), and then continue on. In the limit of $V_A = 0$, if m is the mass in S of A when A is at rest, then $m_A = m$. In the limit of $V'_B = 0$, if $m(v)$ is the mass in S of B, which is moving at the velocity v, then $m_B = m(v)$. Hence Eq. (1.14) becomes

$$m(v) = \frac{m}{\sqrt{1 - v^2/c^2}} \tag{1.15}$$

We can see that if linear momentum is defined as

Relativistic momentum

$$\mathbf{p} = \frac{m\mathbf{v}}{\sqrt{1 - v^2/c^2}} \tag{1.16}$$

then conservation of momentum is valid in special relativity. When $v \ll c$, Eq. (1.16) becomes just $\mathbf{p} = m\mathbf{v}$, the classical momentum, as required. Equation (1.16) is often written as

Relativistic momentum

$$\mathbf{p} = \gamma m\mathbf{v} \tag{1.17}$$

where

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} \tag{1.18}$$

In this definition, m is the **proper mass** (or **rest mass**) of an object, its mass when measured at rest relative to an observer. (The symbol γ is the Greek letter gamma.)

“Relativistic Mass”

We could alternatively regard the increase in an object’s momentum over the classical value as being due to an increase in the object’s mass. Then we would call $m_0 = m$ the rest mass of the object and $m = m(v)$ from Eq. (1.17) its relativistic mass, its mass when moving relative to an observer, so that $\mathbf{p} = m\mathbf{v}$. This is the view often taken in the past, at one time even by Einstein. However, as Einstein later wrote, the idea of relativistic mass is “not good” because “no clear definition can be given. It is better to introduce no other mass concept than the ‘rest mass’ m.” In this book the term mass and the symbol m will always refer to proper (or rest) mass, which will be considered relativistically invariant.

Figure 1.14 shows how p varies with v/c for both $\gamma m v$ and mv. When v/c is small, mv and $\gamma m v$ are very nearly the same. (For $v = 0.01c$, the difference is only 0.005 percent; for $v = 0.1c$, it is 0.5 percent, still small). As v approaches c, however, the curve for $\gamma m v$ rises more and more steeply (for $v = 0.9c$, the difference is 229 percent). If $v = c$, $p = \gamma m v = \infty$, which is impossible. We conclude that no material object can travel as fast as light.

But what if a spacecraft moving at $v_1 = 0.5c$ relative to the earth fires a projectile at $v_2 = 0.5c$ in the same direction? We on earth might expect to observe the projectile’s speed as $v_1 + v_2 = c$. Actually, as discussed in Appendix I to this chapter, velocity addition in relativity is not so simple a process, and we would find the projectile’s speed to be only $0.8c$ in such a case.

Relativistic Second Law

In relativity Newton’s second law of motion is given by

Relativistic second law

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} = \frac{d}{dt}(\gamma m \mathbf{v}) \tag{1.19}$$

This is more complicated than the classical formula $\mathbf{F} = m\mathbf{a}$ because γ is a function of v. When $v \ll c$, γ is very nearly equal to 1, and $\mathbf{F}$ is very nearly equal to $m\mathbf{v}$, as it should be.

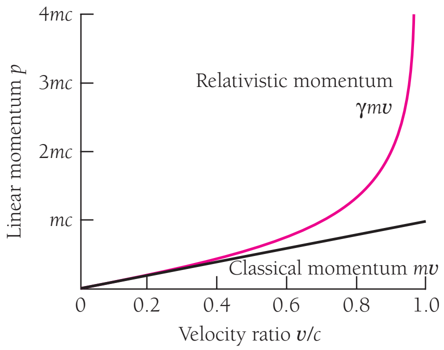


Figure 1.14 The momentum of an object moving at the velocity v relative to an observer. The mass m of the object is its value when it is at rest relative to the observer. The object's velocity can never reach c because its momentum would then be infinite, which is impossible. The relativistic momentum $\gamma m v$ is always correct; the classical momentum mv is valid for velocities much smaller than c.

Example 1.5

Find the acceleration of a particle of mass m and velocity $\mathbf{v}$ when it is acted upon by the constant force $\mathbf{F}$, where $\mathbf{F}$ is parallel to $\mathbf{v}$.

Solution

From Eq. (1.19), since $a = dv/dt$,

$$\begin{aligned} F &= \frac{d}{dt}(\gamma m v) = m\frac{d}{dt}\left(\frac{v}{\sqrt{1 - v^2/c^2}}\right) \\ &= m\left[\frac{1}{\sqrt{1 - v^2/c^2}} + \frac{v^2/c^2}{(1 - v^2/c^2)^{3/2}}\right]\frac{dv}{dt} \\ &= \frac{ma}{(1 - v^2/c^2)^{3/2}} \end{aligned}$$

We note that F is equal to $\gamma^3 ma$, *not* to γma. Merely replacing m by γm in classical formulas does not always give a relativistically correct result.

The acceleration of the particle is therefore

$$a = \frac{F}{m}(1 - v^2/c^2)^{3/2}$$

Even though the force is constant, the acceleration of the particle decreases as its velocity increases. As $v \to c$, $a \to 0$, so the particle can never reach the speed of light, a conclusion we expect.

1.8 MASS AND ENERGY

Where $E_0 = mc^2$ comes from

The most famous relationship Einstein obtained from the postulates of special relativity—how powerful they turn out to be!—concerns mass and energy. Let us see how this relationship can be derived from what we already know.

As we recall from elementary physics, the work W done on an object by a constant force of magnitude F that acts through the distance s, where $\mathbf{F}$ is in the same direction as $\mathbf{s}$, is given by $W = Fs$. If no other forces act on the object and the object starts from rest, all the work done on it becomes kinetic energy KE, so KE $= Fs$. In the general case where F need not be constant, the formula for kinetic energy is the integral

$$\text{KE} = \int_0^s F\, ds$$

In nonrelativistic physics, the kinetic energy of an object of mass m and speed v is KE $= \frac{1}{2}mv^2$. To find the correct relativistic formula for KE we start from the relativistic form of the second law of motion, Eq. (1.19), which gives

$$\text{KE} = \int_0^s \frac{d(\gamma m v)}{dt}\, ds = \int_0^{mv} v\, d(\gamma m v) = \int_0^v v\, d\left(\frac{mv}{\sqrt{1 - v^2/c^2}}\right)$$

Integrating by parts ($\int x\,dy = xy - \int y\,dx$),

$$\begin{aligned} \mathrm{KE} &= \frac{m\boldsymbol{v}^2}{\sqrt{1-\boldsymbol{v}^2/c^2}} - m\int_0^{\boldsymbol{v}} \frac{\boldsymbol{v}\,d\boldsymbol{v}}{\sqrt{1-\boldsymbol{v}^2/c^2}} \\ &= \frac{m\boldsymbol{v}^2}{\sqrt{1-\boldsymbol{v}^2/c^2}} + \left[mc^2\sqrt{1-\boldsymbol{v}^2/c^2}\right]_0^{\boldsymbol{v}} \\ &= \frac{mc^2}{\sqrt{1-\boldsymbol{v}^2/c^2}} - mc^2 \end{aligned}$$

Kinetic energy
$$\mathrm{KE} = \gamma mc^2 - mc^2 = (\gamma - 1)mc^2 \tag{1.20}$$

This result states that the kinetic energy of an object is equal to the difference between γmc^2 and mc^2. Equation (1.20) may be written

Total energy
$$E = \gamma mc^2 = mc^2 + \mathrm{KE} \tag{1.21}$$

If we interpret γmc^2 as the **total energy** E of the object, we see that when it is at rest and KE = 0, it nevertheless possesses the energy mc^2. Accordingly mc^2 is called the **rest energy** E_0 of something whose mass is m. We therefore have

$$E = E_0 + \mathrm{KE}$$

where

Rest energy
$$E_0 = mc^2 \tag{1.22}$$

If the object is moving, its total energy is

Total energy
$$E = \gamma mc^2 = \frac{mc^2}{\sqrt{1-\boldsymbol{v}^2/c^2}} \tag{1.23}$$

Example 1.6

A stationary body explodes into two fragments each of mass 1.0 kg that move apart at speeds of 0.6c relative to the original body. Find the mass of the original body.

Solution

The rest energy of the original body must equal the sum of the total energies of the fragments. Hence

$$E_0 = mc^2 = \gamma m_1c^2 + \gamma m_2c^2 = \frac{m_1c^2}{\sqrt{1-\boldsymbol{v}_1^2/c^2}} + \frac{m_2c^2}{\sqrt{1-\boldsymbol{v}_2^2/c^2}}$$

and

$$m = \frac{E_0}{c^2} = \frac{(2)(1.0\text{ kg})}{\sqrt{1-(0.60)^2}} = 2.5\text{ kg}$$

Since mass and energy are not independent entities, their separate conservation principles are properly a single one—the principle of conservation of mass energy. Mass *can* be created or destroyed, but when this happens, an equivalent amount of energy simultaneously vanishes or comes into being, and vice versa. Mass and energy are different aspects of the same thing.

It is worth emphasizing the difference between a *conserved* quantity, such as total energy, and an *invariant* quantity, such as proper mass. Conservation of E means that, in a given reference frame, the total energy of some isolated system remains the same regardless of what events occur in the system. However, the total energy may be different as measured from another frame. On the other hand, the invariance of m means that m has the same value in all inertial frames.

The conversion factor between the unit of mass (the kilogram, kg) and the unit of energy (the joule, J) is c^2, so 1 kg of matter—the mass of this book is about that—has an energy content of $mc^2 = (1\ \text{kg})(3 \times 10^8\ \text{m/s})^2 = 9 \times 10^{16}$ J. This is enough to send a payload of a million tons to the moon. How is it possible for so much energy to be bottled up in even a modest amount of matter without anybody having been aware of it until Einstein's work?

In fact, processes in which rest energy is liberated are very familiar. It is simply that we do not usually think of them in such terms. In every chemical reaction that evolves energy, a certain amount of matter disappears, but the lost mass is so small a fraction of the total mass of the reacting substances that it is imperceptible. Hence the "law" of conservation of mass in chemistry. For instance, only about 6×10^{-11} kg of matter vanishes when 1 kg of dynamite explodes, which is impossible to measure directly, but the more than 5 million joules of energy that is released is hard to avoid noticing.

Example 1.7

Solar energy reaches the earth at the rate of about 1.4 kW per square meter of surface perpendicular to the direction of the sun (Fig. 1.15). By how much does the mass of the sun decrease per second owing to this energy loss? The mean radius of the earth's orbit is 1.5×10^{11} m.

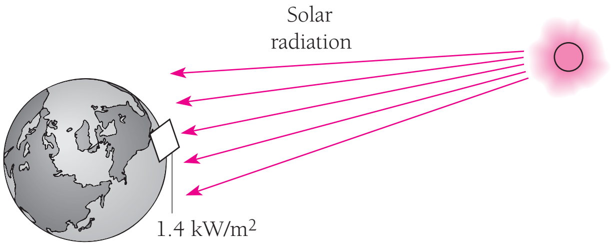


Figure 1.15

Solution

The surface area of a sphere of radius r is $A = 4\pi r^2$. The total power radiated by the sun, which is equal to the power received by a sphere whose radius is that of the earth's orbit, is therefore

$$P = \frac{P}{A}A = \frac{P}{A}(4\pi r^2) = (1.4 \times 10^3\ \text{W/m}^2)(4\pi)(1.5 \times 10^{11}\ \text{m})^2 = 4.0 \times 10^{26}\ \text{W}$$

Thus the sun loses $E_0 = 4.0 \times 10^{26}$ J of rest energy per second, which means that the sun's rest mass decreases by

$$m = \frac{E_0}{c^2} = \frac{4.0 \times 10^{26}\ \text{J}}{(3.0 \times 10^8\ \text{m/s})^2} = 4.4 \times 10^9\ \text{kg}$$

per second. Since the sun's mass is 2.0×10^{30} kg, it is in no immediate danger of running out of matter. The chief energy-producing process in the sun and most other stars is the conversion of hydrogen to helium in its interior. The formation of each helium nucleus is accompanied by the release of 4.0×10^{-11} J of energy, so 10^{37} helium nuclei are produced in the sun per second.

Kinetic Energy at Low Speeds

When the relative speed $\boldsymbol{v}$ is small compared with c, the formula for kinetic energy must reduce to the familiar $\frac{1}{2}m\boldsymbol{v}^2$, which has been verified by experiment at such speeds. Let us see if this is true. The relativistic formula for kinetic energy is

Kinetic energy

$$\mathrm{KE} = \gamma mc^2 - mc^2 = \frac{mc^2}{\sqrt{1 - \boldsymbol{v}^2/c^2}} - mc^2 \qquad (1.20)$$

Since $\boldsymbol{v}^2/c^2 \ll 1$, we can use the binomial approximation $(1 + x)^n \approx 1 + nx$, valid for $|x| \ll 1$, to obtain

$$\frac{1}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \approx 1 + \frac{1}{2}\frac{\boldsymbol{v}^2}{c^2} \qquad \boldsymbol{v} \ll c$$

Thus we have the result

$$\mathrm{KE} \approx \left(1 + \frac{1}{2}\frac{\boldsymbol{v}^2}{c^2}\right)mc^2 - mc^2 \approx \frac{1}{2}m\boldsymbol{v}^2 \qquad \boldsymbol{v} \ll c$$

At low speeds the relativistic expression for the kinetic energy of a moving object does indeed reduce to the classical one. So far as is known, the correct formulation of mechanics has its basis in relativity, with classical mechanics representing an approximation that is valid only when $\boldsymbol{v} \ll c$. Figure 1.16 shows how the kinetic energy of

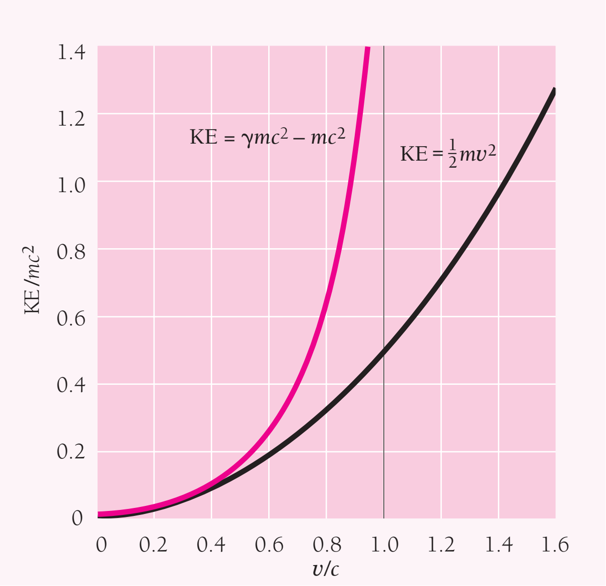


Figure 1.16 A comparison between the classical and relativistic formulas for the ratio between kinetic energy KE of a moving body and its rest energy mc^2. At low speeds the two formulas give the same result, but they diverge at speeds approaching that of light. According to relativistic mechanics, a body would need an infinite kinetic energy to travel with the speed of light, whereas in classical mechanics it would need only a kinetic energy of half its rest energy to have this speed.

a moving object varies with its speed according to both classical and relativistic mechanics.

The degree of accuracy required is what determines whether it is more appropriate to use the classical or to use the relativistic formulas for kinetic energy. For instance, when $\boldsymbol{v} = 10^7$ m/s ($0.033c$), the formula $\frac{1}{2}m\boldsymbol{v}^2$ understates the true kinetic energy by only 0.08 percent; when $\boldsymbol{v} = 3 \times 10^7$ m/s ($0.1c$), it understates the true kinetic energy by 0.8 percent; but when $\boldsymbol{v} = 1.5 \times 10^8$ m/s ($0.5c$), the understatement is a significant 19 percent; and when $\boldsymbol{v} = 0.999c$, the understatement is a whopping 4300 percent. Since 10^7 m/s is about 6310 mi/s, the nonrelativistic formula $\frac{1}{2}m\boldsymbol{v}^2$ is entirely satisfactory for finding the kinetic energies of ordinary objects, and it fails only at the extremely high speeds reached by elementary particles under certain circumstances.

1.9 ENERGY AND MOMENTUM

How they fit together in relativity

Total energy and momentum are conserved in an isolated system, and the rest energy of a particle is invariant. Hence these quantities are in some sense more fundamental than velocity or kinetic energy, which are neither. Let us look into how the total energy, rest energy, and momentum of a particle are related.

We begin with Eq. (1.23) for total energy,

Total energy

$$E = \frac{mc^2}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.23}$$

and square it to give

$$E^2 = \frac{m^2c^4}{1 - \boldsymbol{v}^2/c^2}$$

From Eq. (1.17) for momentum,

Momentum

$$p = \frac{m\boldsymbol{v}}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.17}$$

we find that

$$p^2c^2 = \frac{m^2\boldsymbol{v}^2c^2}{1 - \boldsymbol{v}^2/c^2}$$

Now we subtract p^2c^2 from E^2:

$$\begin{aligned} E^2 - p^2c^2 &= \frac{m^2c^4 - m^2\boldsymbol{v}^2c^2}{1 - \boldsymbol{v}^2/c^2} = \frac{m^2c^4(1 - \boldsymbol{v}^2/c^2)}{1 - \boldsymbol{v}^2/c^2} \\ &= (mc^2)^2 \end{aligned}$$

Hence

Energy and momentum

$$E^2 = (mc^2)^2 + p^2c^2 \tag{1.24}$$

which is the formula we want. We note that, because mc^2 is invariant, so is $E^2 - p^2c^2$: this quantity for a particle has the same value in all frames of reference.

For a system of particles rather than a single particle, Eq. (1.24) holds provided that the rest energy mc^2—and hence mass m—is that of the entire system. If the particles in the system are moving with respect to one another, the sum of their individual rest energies may not equal the rest energy of the system. We saw this in Example 1.7 when a stationary body of mass 2.5 kg exploded into two smaller bodies, each of mass 1.0 kg, that then moved apart. If we were inside the system, we would interpret the difference of 0.5 kg of mass as representing its conversion into kinetic energy of the smaller bodies. But seen as a whole, the system is at rest both before and after the explosion, so the *system* did not gain kinetic energy. Therefore the rest energy of the system includes the kinetic energies of its internal motions and it corresponds to a mass of 2.5 kg both before and after the explosion.

In a given situation, the rest energy of an isolated system may be greater than, the same as, or less than the sum of the rest energies of its members. An important case in which the system rest energy is less than the rest energies of its members is that of a system of particles held together by attractive forces, such as the neutrons and protons in an atomic nucleus. The rest energy of a nucleus (except that of ordinary hydrogen, which is a single proton) is less than the total of the rest energies of its constituent particles. The difference is called the *binding energy* of the nucleus. To break a nucleus up completely calls for an amount of energy at least equal to its binding energy. This topic will be explored in detail in Sec. 11.4. For the moment it is interesting to note how large nuclear binding energies are—nearly 10^{12} kJ per kg of nuclear matter is typical. By comparison, the binding energy of water molecules in liquid water is only 2260 kJ/kg; this is the energy needed to turn 1 kg of water at 100°C to steam at the same temperature.

Massless Particles

Can a massless particle exist? To be more precise, can a particle exist which has no rest mass but which nevertheless exhibits such particlelike properties as energy and momentum? In classical mechanics, a particle must have rest mass in order to have energy and momentum, but in relativistic mechanics this requirement does not hold.

From Eqs. (1.17) and (1.23), when $m = 0$ and $\boldsymbol{v} \ll c$, it is clear that $E = p = 0$. A massless particle with a speed less than that of light can have neither energy nor momentum. However, when $m = 0$ and $\boldsymbol{v} = c$, $E = 0/0$ and $p = 0/0$, which are indeterminate: E and p can have any values. Thus Eqs. (1.17) and (1.23) are consistent with the existence of massless particles that possess energy and momentum *provided that they travel with the speed of light.*

Equation (1.24) gives us the relationship between E and p for a particle with $m = 0$:

Massless particle

$$E = pc \tag{1.25}$$

The conclusion is not that massless particles necessarily occur, only that the laws of physics do not exclude the possibility as long as $\boldsymbol{v} = \boldsymbol{c}$ and $E = pc$ for them. In fact, a massless particle—the photon—indeed exists and its behavior is as expected, as we shall find in Chap. 2.

Electronvolts

In atomic physics the usual unit of energy is the **electronvolt** (eV), where 1 eV is the energy gained by an electron accelerated through a potential difference of 1 volt. Since $W = QV$,

$$1 \text{ eV} = (1.602 \times 10^{-19} \text{ C})(1.000 \text{ V}) = 1.602 \times 10^{-19} \text{ J}$$

Two quantities normally expressed in electronvolts are the ionization energy of an atom (the work needed to remove one of its electrons) and the binding energy of a molecule (the energy needed to break it apart into separate atoms). Thus the ionization energy of nitrogen is 14.5 eV and the binding energy of the hydrogen molecule H_2 is 4.5 eV. Higher energies in the atomic realm are expressed in **kiloelectronvolts** (keV), where 1 keV $= 10^3$ eV.

In nuclear and elementary-particle physics even the keV is too small a unit in most cases, and the **megaelectronvolt** (MeV) and **gigaelectronvolt** (GeV) are more appropriate, where

$$1 \text{ MeV} = 10^6 \text{ eV} \qquad 1 \text{ GeV} = 10^9 \text{ eV}$$

An example of a quantity expressed in MeV is the energy liberated when the nucleus of a certain type of uranium atom splits into two parts. Each such fission event releases about 200 MeV; this is the process that powers nuclear reactors and weapons.

The rest energies of elementary particles are often expressed in MeV and GeV and the corresponding rest masses in MeV/c^2 and GeV/c^2. The advantage of the latter units is that the rest energy equivalent to a rest mass of, say, 0.938 GeV/c^2 (the rest mass of the proton) is just $E_0 = mc^2 = 0.938$ GeV. If the proton's kinetic energy is 5.000 GeV, finding its total energy is simple:

$$E = E_0 + \text{KE} = (0.938 + 5.000) \text{ GeV} = 5.938 \text{ GeV}$$

In a similar way the MeV/c and GeV/c are sometimes convenient units of linear momentum. Suppose we want to know the momentum of a proton whose speed is 0.800c. From Eq. (1.17) we have

$$p = \frac{m\boldsymbol{v}}{\sqrt{1 - \boldsymbol{v}^2/c^2}} = \frac{(0.938 \text{ GeV}/c^2)(0.800c)}{\sqrt{1 - (0.800c)^2/c^2}}$$
$$= \frac{0.750 \text{ GeV}/c}{0.600} = 1.25 \text{ GeV}/c$$

Example 1.8

An electron ($m = 0.511$ MeV/c^2) and a photon ($m = 0$) both have momenta of 2.000 MeV/c. Find the total energy of each.

Solution

(*a*) From Eq. (1.24) the electron's total energy is

$$E = \sqrt{m^2c^4 + p^2c^2} = \sqrt{(0.511 \text{ MeV}/c^2)^2c^4 + (2.000 \text{ MeV}/c)^2c^2}$$
$$= \sqrt{(0.511 \text{ MeV})^2 + (2.000 \text{ MeV})^2} = 2.064 \text{ MeV}$$

(*b*) From Eq. (1.25) the photon's total energy is

$$E = pc = (2.000 \text{ MeV}/c)c = 2.000 \text{ MeV}$$

1.10 GENERAL RELATIVITY

Gravity is a warping of spacetime

Special relativity is concerned only with inertial frames of reference, that is, frames that are not accelerated. Einstein's 1916 **general theory of relativity** goes further by including the effects of accelerations on what we observe. Its essential conclusion is that the force of gravity arises from a warping of spacetime around a body of matter (Fig. 1.17). As a result, an object moving through such a region of space in general follows a curved path rather than a straight one, and may even be trapped there.

The **principle of equivalence** is central to general relativity:

> An observer in a closed laboratory cannot distinguish between the effects produced by a gravitational field and those produced by an acceleration of the laboratory.

This principle follows from the experimental observation (to better than 1 part in 10^{12}) that the inertial mass of an object, which governs the object's acceleration when a force acts on it, is always equal to its gravitational mass, which governs the gravitational force another object exerts on it. (The two masses are actually proportional; the constant of proportionality is set equal to 1 by an appropriate choice of the constant of gravitation *G*.)

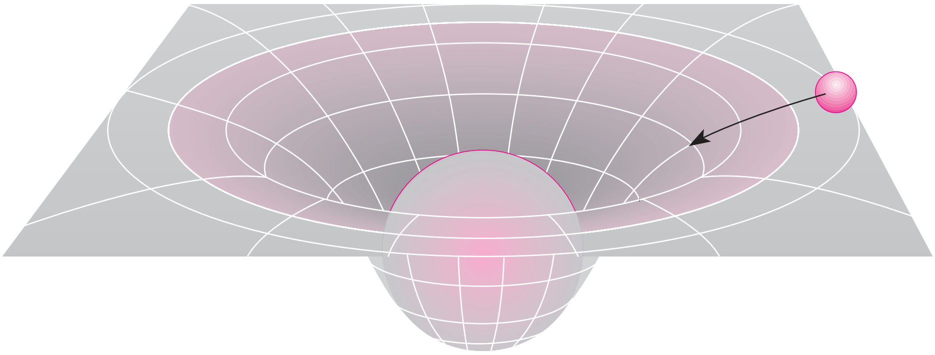
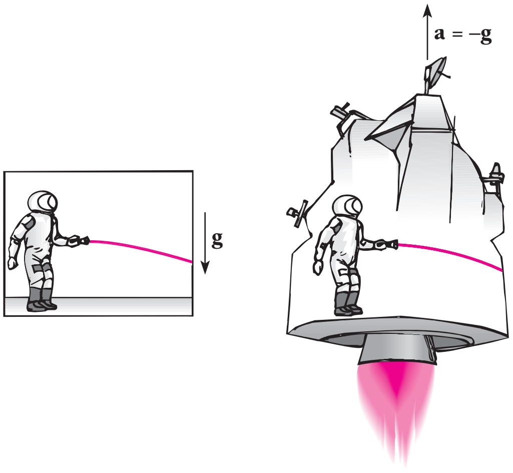


Figure 1.17 General relativity pictures gravity as a warping of spacetime due to the presence of a body of matter. An object nearby experiences an attractive force as a result of this distortion, much as a marble rolls toward the bottom of a depression in a rubber sheet. To paraphrase J. A. Wheeler, spacetime tells mass how to move, and mass tells spacetime how to curve.



Laboratory in gravitational field

Accelerated laboratory

Figure 1.18 According to the principle of equivalence, events that take place in an accelerated laboratory cannot be distinguished from those which take place in a gravitational field. Hence the deflection of a light beam relative to an observer in an accelerated laboratory means that light must be similarly deflected in a gravitational field.

Gravity and Light

It follows from the principle of equivalence that light should be subject to gravity. If a light beam is directed across an accelerated laboratory, as in Fig. 1.18, its path relative to the laboratory will be curved. This means that, if the light beam is subject to the gravitational field to which the laboratory's acceleration is equivalent, the beam would follow the same curved path.

According to general relativity, light rays that graze the sun should have their paths bent toward it by 0.005°—the diameter of a dime seen from a mile away. This prediction was first confirmed in 1919 by photographs of stars that appeared in the sky near the sun during an eclipse, when they could be seen because the sun's disk was covered by the moon. The photographs were then compared with other photographs of the same part of the sky taken when the sun was in a distant part of the sky (Fig. 1.19). Einstein became a world celebrity as a result.

Because light is deflected in a gravitational field, a dense concentration of mass—such as a galaxy of stars—can act as a lens to produce multiple images of a distant light source located behind it (Fig. 1.20). A **quasar,** the nucleus of a young galaxy, is brighter than 100 billion stars but is no larger than the solar system. The first observation of gravitational lensing was the discovery in 1979 of what seemed to be a pair of nearby quasars but was actually a single one whose light was deviated by an intervening massive object. Since then a number of other gravitational lenses have been found; the effect occurs in radio waves from distant sources as well as in light waves.

The interaction between gravity and light also gives rise to the gravitational red shift and to black holes, topics that are considered in Chap. 2.

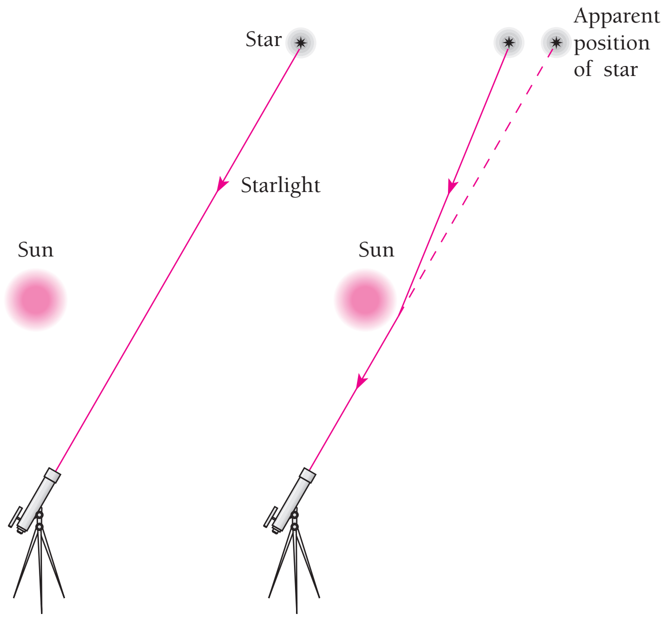


Figure 1.19 Starlight passing near the sun is deflected by its strong gravitational field. The deflection can be measured during a solar eclipse when the sun's disk is obscured by the moon.

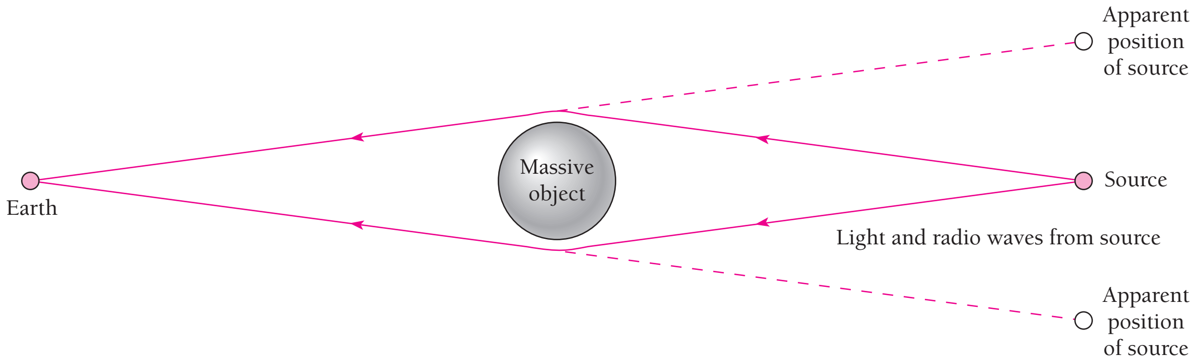

Apparent position of source
Massive object
Earth
Source
Light and radio waves from source
Apparent position of source

Figure 1.20 A gravitational lens. Light and radio waves from a source such as a quasar are deviated by a massive object such as a galaxy so that they seem to come from two or more identical sources. A number of such gravitational lenses have been identified.

Other Findings of General Relativity

A further success of general relativity was the clearing up of a long-standing puzzle in astronomy. The perihelion of a planetary orbit is the point in the orbit nearest the sun. Mercury's orbit has the peculiarity that its perihelion shifts (precesses) about 1.6° per century (Fig. 1.21). All but 43″ (1″ = 1 arc second = $\frac{1}{3600}$ of a degree) of this shift is due to the attractions of other planets, and for a while the discrepancy was used as evidence for an undiscovered planet called Vulcan whose orbit was supposed to lie inside that of Mercury. When gravity is weak, general relativity gives very nearly the same results as Newton's formula $F = Gm_1m_2/r^2$. But Mercury is close to the sun and so moves in a strong gravitational field, and Einstein was able to show from general relativity that a precession of 43″ per century was to be expected for its orbit.

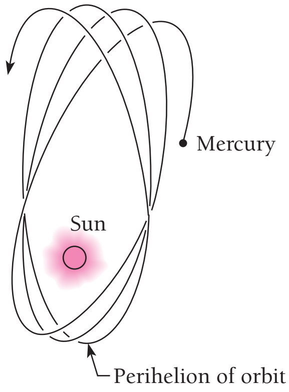


Figure 1.21 The precession of the perihelion of Mercury's orbit.

The existence of **gravitational waves** that travel with the speed of light was the prediction of general relativity that had to wait the longest to be verified. To visualize gravitational waves, we can think in terms of the model of Fig. 1.17 in which two-dimensional space is represented by a rubber sheet distorted by masses embedded in it. If one of the masses vibrates, waves will be sent out in the sheet that set other masses in vibration. A vibrating electric charge similarly sends out electromagnetic waves that excite vibrations in other charges.

A big difference between the two kinds of waves is that gravitational waves are extremely weak, so that despite much effort none have as yet been directly detected. However, in 1974 strong evidence for gravitational waves was found in the behavior of a system of two nearby stars, one a pulsar, that revolve around each other. A **pulsar** is a very small, dense star, composed mainly of neutrons, that spins rapidly and sends out flashes of light and radio waves at a regular rate, much as the rotating beam of a lighthouse does (see Sec. 9.11). The pulsar in this particular binary system emits pulses every 59 milliseconds (ms), and it and its companion (probably another neutron star) have an orbital period of about 8 h. According to general relativity, such a system should give off gravitational waves and lose energy as a result, which would reduce the orbital period as the stars spiral in toward each other. A change in orbital period means a change in the arrival times of the pulsar's flashes, and in the case of the observed binary system the orbital period was found to be decreasing at 75 ms per year. This is so close to the figure that general relativity predicts for the system that there seems to be no doubt that gravitational radiation is responsible. The 1993 Nobel Prize in physics was awarded to Joseph Taylor and Russell Hulse for this work.

Much more powerful sources of gravitational waves ought to be such events as two black holes colliding and supernova explosions in which the remnant star cores collapse into neutron stars (again, see Sec. 9.11). A gravitational wave that passes through a body of matter will cause distortions to ripple through it due to fluctuations in the gravitational field. Because gravitational forces are feeble—the electric attraction between a proton and an electron is over 10^{39} times greater than the gravitational attraction between them—such distortions at the earth induced by gravitational waves from a supernova in our galaxy (which occurs an average of once every 30 years or so) would amount to only about 1 part in 10^{18}, even less for a more distant supernova. This corresponds to a change in, say, the height of a person by well under the diameter of an atomic nucleus, yet it seems to be detectable—just—with current technology.

In one method, a large metal bar cooled to a low temperature to minimize the random thermal motions of its atoms is monitored by sensors for vibrations due to gravitational waves. In another method, an interferometer similar to the one shown in Fig. 1.2 with a laser as the light source is used to look for changes in the lengths of the arms to which the mirrors are attached. Instruments of both kinds are operating, thus far with no success.

A really ambitious scheme has been proposed that would use six spacecraft in orbit around the sun placed in pairs at the corners of a triangle whose sides are 5 million kilometers (km) long. Lasers, mirrors, and sensors in the spacecraft would detect changes in their spacings resulting from the passing of a gravitational wave. It may only be a matter of time before gravitational waves will be providing information about a variety of cosmic disturbances on the largest scale.

Appendix I to Chapter 1

The Lorentz Transformation

Suppose we are in an inertial frame of reference S and find the coordinates of some event that occurs at the time t are x, y, z. An observer located in a different inertial frame S' which is moving with respect to S at the constant velocity $\mathbf{v}$ will find that the same event occurs at the time t' and has the coordinates x', y', z'. (In order to simplify our work, we shall assume that $\mathbf{v}$ is in the $+x$ direction, as in Fig. 1.22.) How are the measurements x, y, z, t related to x', y', z', t'?

Galilean Transformation

Before special relativity, transforming measurements from one inertial system to another seemed obvious. If clocks in both systems are started when the origins of S and S' coincide, measurements in the x direction made is S will be greater than those made in S' by the amount $\boldsymbol{v}t$, which is the distance S' has moved in the x direction. That is,

$$x' = x - \boldsymbol{v}t \tag{1.26}$$

There is no relative motion in the y and z directions, and so

$$y' = y \tag{1.27}$$

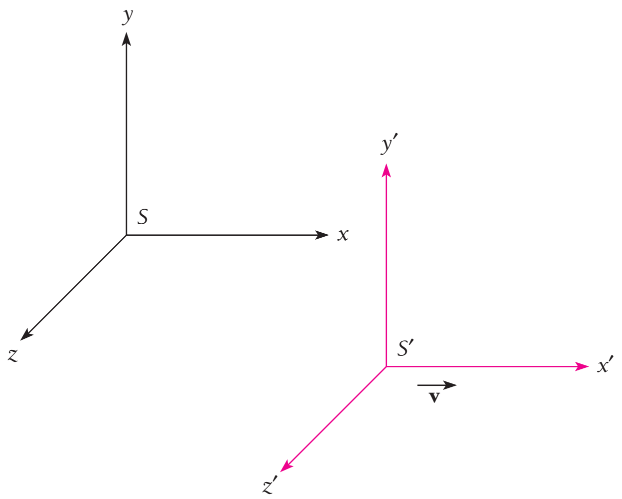


Figure 1.22 Frame S' moves in the $+x$ direction with the speed $\boldsymbol{v}$ relative to frame S. The Lorentz transformation must be used to convert measurements made in one of these frames to their equivalents in the other.

$$z' = z \tag{1.28}$$

In the absence of any indication to the contrary in our everyday experience, we further assume that

$$t' = t \tag{1.29}$$

The set of Eqs. (1.26) to (1.29) is known as the **Galilean transformation.**

To convert velocity components measured in the S frame to their equivalents in the S' frame according to the Galilean transformation, we simply differentiate x', y', and z' with respect to time:

$$\boldsymbol{v}'_x = \frac{dx'}{dt'} = \boldsymbol{v}_x - \boldsymbol{v} \tag{1.30}$$

$$\boldsymbol{v}'_y = \frac{dy'}{dt'} = \boldsymbol{v}_y \tag{1.31}$$

$$\boldsymbol{v}'_z = \frac{dz'}{dt'} = \boldsymbol{v}_z \tag{1.32}$$

Although the Galilean transformation and the corresponding velocity transformation seem straightforward enough, they violate both of the postulates of special relativity. The first postulate calls for the same equations of physics in both the S and S' inertial frames, but the equations of electricity and magnetism become very different when the Galilean transformation is used to convert quantities measured in one frame into their equivalents in the other. The second postulate calls for the same value of the speed of light c whether determined in S or S'. If we measure the speed of light in the x direction in the S system to be c, however, in the S' system it will be

$$c' = c - \boldsymbol{v}$$

according to Eq. (1.30). Clearly a different transformation is required if the postulates of special relativity are to be satisfied. We would expect both time dilation and length contraction to follow naturally from this new transformation.

Lorentz Transformation

A reasonable guess about the nature of the correct relationship between x and x' is

$$x' = k(x - \boldsymbol{v}t) \tag{1.33}$$

Here k is a factor that does not depend upon either x or t but may be a function of $\boldsymbol{v}$. The choice of Eq. (1.33) follows from several considerations:

1 It is linear in x and x', so that a single event in frame S corresponds to a single event in frame S', as it must.
2 It is simple, and a simple solution to a problem should always be explored first.
3 It has the possibility of reducing to Eq. (1.26), which we know to be correct in ordinary mechanics.

Because the equations of physics must have the same form in both S and S', we need only change the sign of $\boldsymbol{v}$ (in order to take into account the difference in the direction of relative motion) to write the corresponding equation for x in terms of x' and t':

$$x = k(x' + \boldsymbol{v}t') \tag{1.34}$$

The factor k must be the same in both frames of reference since there is no difference between S and S' other than in the sign of $\boldsymbol{v}$.

As in the case of the Galilean transformation, there is nothing to indicate that there might be differences between the corresponding coordinates y, y' and z, z' which are perpendicular to the direction of $\boldsymbol{v}$. Hence we again take

$$y' = y \tag{1.35}$$

$$z' = z \tag{1.36}$$

The time coordinates t and t', however, are *not* equal. We can see this by substituting the value of x' given by Eq. (1.33) into Eq. (1.34). This gives

$$x = k^2(x - \boldsymbol{v}t) + k\boldsymbol{v}t'$$

from which we find that

$$t' = kt + \left(\frac{1 - k^2}{k\boldsymbol{v}}\right)x \tag{1.37}$$

Equations (1.33) and (1.35) to (1.37) constitute a coordinate transformation that satisfies the first postulate of special relativity.

The second postulate of relativity gives us a way to evaluate k. At the instant $t = 0$, the origins of the two frames of reference S and S' are in the same place, according to our initial conditions, and $t' = 0$ then also. Suppose that a flare is set off at the common origin of S and S' at $t = t' = 0$, and the observers in each system measure the speed with which the flare's light spreads out. Both observers must find the same speed c (Fig. 1.23), which means that in the S frame

$$x = ct \tag{1.38}$$

and in the S' frame

$$x' = ct' \tag{1.39}$$

Substituting for x' and t' in Eq. (1.39) with the help of Eqs. (1.33) and (1.37) gives

$$k(x - \boldsymbol{v}t) = ckt + \left(\frac{1 - k^2}{k\boldsymbol{v}}\right)cx$$

and solving for x,

$$x = \frac{ckt + \boldsymbol{v}kt}{k - \left(\frac{1 - k^2}{k\boldsymbol{v}}\right)c} = ct\left[\frac{k + \frac{\boldsymbol{v}}{c}k}{k - \left(\frac{1 - k^2}{k\boldsymbol{v}}\right)c}\right] = ct\left[\frac{1 + \frac{\boldsymbol{v}}{c}}{1 - \left(\frac{1}{k^2} - 1\right)\frac{c}{\boldsymbol{v}}}\right]$$

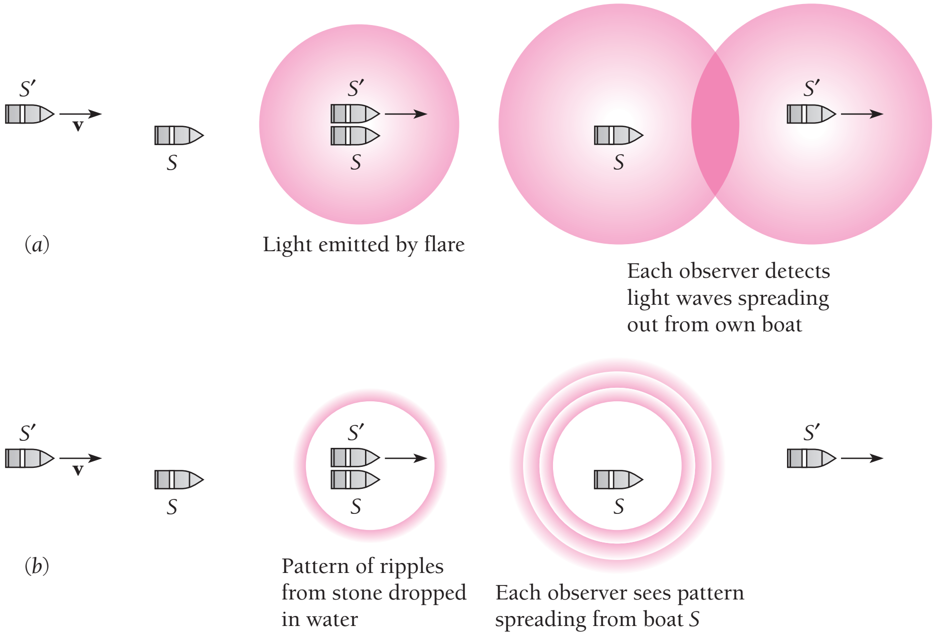

S′
v
S
S′
S
S′
S
(a)
Light emitted by flare
Each observer detects light waves spreading out from own boat
S′
v
S
S′
S
S
S′
(b)
Pattern of ripples from stone dropped in water
Each observer sees pattern spreading from boat S

Figure 1.23 (*a*) Inertial frame S' is a boat moving at speed $\boldsymbol{v}$ in the $+x$ direction relative to another boat, which is the inertial frame S. When $t = t_0 = 0$, S' is next to S, and $x = x_0 = 0$. At this moment a flare is fired from one of the boats. An observer on boat S detects light waves spreading out at speed c from his boat. An observer on boat S' also detects light waves spreading out at speed c from her boat, even though S' is moving to the right relative to S. (*b*) If instead a stone were dropped in the water at $t = t_0 = 0$, the observers would find a pattern of ripples spreading out around S at different speeds relative to their boats. The difference between (*a*) and (*b*) is that water, in which the ripples move, is itself a frame of reference whereas space, in which light moves, is not.

This expression for x will be the same as that given by Eq. (1.38), namely, $x = ct$, provided that the quantity in the brackets equals 1. Therefore

$$\frac{1 + \dfrac{\boldsymbol{v}}{c}}{1 - \left(\dfrac{1}{k^2} - 1\right)\dfrac{c}{\boldsymbol{v}}} = 1$$

and

$$k = \frac{1}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.40}$$

Finally we put this value of k in Eqs. (1.36) and (1.40). Now we have the complete transformation of measurements of an event made in S to the corresponding measurements made in S':

Lorentz transformation

$$x' = \frac{x - \boldsymbol{v}t}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.41}$$

$$y' = y \tag{1.42}$$

$$z' = z \tag{1.43}$$

$$t' = \frac{t - \dfrac{vx}{c^2}}{\sqrt{1 - v^2/c^2}} \tag{1.44}$$

These equations comprise the **Lorentz transformation.** They were first obtained by the Dutch physicist H.A. Lorentz, who showed that the basic formulas of electromagnetism are the same in all inertial frames only when Eqs. (1.41) to (1.44) are used. It was not until several years later that Einstein discovered their full significance. It is obvious that the Lorentz transformation reduces to the Galilean transformation when the relative velocity v is small compared with the velocity of light c.

Example 1.9

Derive the relativistic length contraction using the Lorentz transformation.

Solution

Let us consider a rod lying along the x' axis in the moving frame S'. An observer in this frame determines the coordinates of its ends to be x_1' and x_2', and so the proper length of the rod is

$$L_0 = x_2' - x_1'$$

Hendrik A. Lorentz (1853–1928) was born in Arnhem, Holland, and studied at the University of Leyden. At nineteen he returned to Arnhem and taught at the high school there while preparing a doctoral thesis that extended Maxwell's theory of electromagnetism to cover the details of the refraction and reflection of light. In 1878 he became professor of theoretical physics at Leyden, the first such post in Holland, where he remained for thirty-four years until he moved to Haarlem. Lorentz went on to reformulate and simplify Maxwell's theory and to introduce the idea that electromagnetic fields are created by electric charges on the atomic level. He proposed that the emission of light by atoms and various optical phenomena could be traced to the motions and interactions of atomic electrons. The discovery in 1896 by Pieter Zeeman, a student of his, that the spectral lines of atoms that radiate in a magnetic field are split into components of slightly different frequency confirmed Lorentz's work and led to a Nobel Prize for both of them in 1902.

The set of equations that enables electromagnetic quantities in one frame of reference to be transformed into their values in another frame of reference moving relative to the first were found by Lorentz in 1895, although their full significance was not realized until Einstein's theory of special relativity ten years afterward. Lorentz (and, independently, the Irish physicist G. F. Fitzgerald) suggested that the negative result of the Michelson-Morley experiment could be understood if lengths in the direction of motion relative to an observer were contracted. Subsequent experiments showed that although such contractions do occur, they are not the real reason for the Michelson-Morley result, which is that there is no "ether" to serve as a universal frame of reference.

In order to find $L = x_2 - x_1$, the length of the rod as measured in the stationary frame S at the time t, we make use of Eq. (1.41) to give

$$x'_1 = \frac{x_1 - \boldsymbol{v}t}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \qquad x'_2 = \frac{x_2 - \boldsymbol{v}t}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

Hence $$L = x_2 - x_1 = (x'_2 - x'_1)\sqrt{1 - \boldsymbol{v}^2/c^2} = L_0\sqrt{1 - \boldsymbol{v}^2/c^2}$$

This is the same as Eq. (1.9)

Inverse Lorentz Transformation

In Example 1.9 the coordinates of the ends of the moving rod were measured in the stationary frame S at the same time t, and it was easy to use Eq. (1.41) to find L in terms of L_0 and $\boldsymbol{v}$. If we want to examine time dilation, though, Eq. (1.44) is not convenient, because t_1 and t_2, the start and finish of the chosen time interval, must be measured when the moving clock is at the respective *different* positions x_1 and x_2. In situations of this kind it is easier to use the **inverse Lorentz transformation,** which converts measurements made in the moving frame S' to their equivalents in S.

To obtain the inverse transformation, primed and unprimed quantities in Eqs. (1.41) to (1.44) are exchanged, and $\boldsymbol{v}$ is replaced by $-\boldsymbol{v}$:

Inverse Lorentz transformation

$$x = \frac{x' + \boldsymbol{v}t'}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.45}$$

$$y = y' \tag{1.46}$$

$$z' = z' \tag{1.47}$$

$$t = \frac{t' + \dfrac{\boldsymbol{v}x'}{c^2}}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \tag{1.48}$$

Example 1.10

Derive the formula for time dilation using the inverse Lorentz transformation.

Solution

Let us consider a clock at the point x' in the moving frame S'. When an observer in S' finds that the time is t'_1, an observer in S will find it to be t_1, where, from Eq. (1.48),

$$t_1 = \frac{t'_1 + \dfrac{\boldsymbol{v}x'}{c^2}}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

After a time interval of t_0 (to him), the observer in the moving system finds that the time is now t'_2 according to his clock. That is,

$$t_0 = t'_2 - t'_1$$

The observer in S, however, measures the end of the same time interval to be

$$t_2 = \frac{t'_2 + \dfrac{\boldsymbol{v}x'}{c^2}}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

so to her the duration of the interval t is

$$t = t_2 - t_1 = \frac{t'_2 - t'_1}{\sqrt{1 - \boldsymbol{v}^2/c^2}} = \frac{t_0}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

This is what we found earlier with the help of a light-pulse clock.

Velocity Addition

Special relativity postulates that the speed of light c in free space has the same value for all observers, regardless of their relative motion."Common sense" (which means here the Galilean transformation) tells us that if we throw a ball forward at 10 m/s from a car moving at 30 m/s, the ball's speed relative to the road will be 40 m/s, the sum of the two speeds. What if we switch on the car's headlights when its speed is $\boldsymbol{v}$? The same reasoning suggests that their light, which is emitted from the reference frame S' (the car) in the direction of its motion relative to another frame S (the road), ought to have a speed of $c + \boldsymbol{v}$ as measured in S. But this violates the above postulate, which has had ample experimental verification. Common sense is no more reliable as a guide in science than it is elsewhere, and we must turn to the Lorentz transformation equations for the correct scheme of velocity addition.

Suppose something is moving relative to both S and S'. An observer in S measures its three velocity components to be

$$V_x = \frac{dx}{dt} \qquad V_y = \frac{dy}{dt} \qquad V_z = \frac{dz}{dt}$$

while to an observer in S' they are

$$V'_x = \frac{dx'}{dt'} \qquad V'_y = \frac{dy'}{dt'} \qquad V'_z = \frac{dz'}{dt'}$$

By differentiating the inverse Lorentz transformation equations for x, y, z, and t, we obtain

$$dx = \frac{dx' + \boldsymbol{v}\,dt'}{\sqrt{1 - \boldsymbol{v}^2/c^2}} \qquad dy = dy' \qquad dz = dz' \qquad dt = \frac{dt' + \dfrac{\boldsymbol{v}\,dz'}{c^2}}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

and so

$$V_x = \frac{dx}{dt} = \frac{dx' + \boldsymbol{v}\,dt'}{dt' + \dfrac{\boldsymbol{v}\,dx'}{c^2}} = \frac{\dfrac{dx'}{dt'} + \boldsymbol{v}}{1 + \dfrac{\boldsymbol{v}}{c^2}\dfrac{dx'}{dt'}}$$

Relativistic velocity transformation

$$V_x = \frac{V'_x + v}{1 + \dfrac{vV'_x}{c^2}} \tag{1.49}$$

Similarly,

$$V_y = \frac{V'_y\sqrt{1 - v^2/c^2}}{1 + \dfrac{vV'_x}{c^2}} \tag{1.50}$$

$$V_z = \frac{V'_z\sqrt{1 - v^2/c^2}}{1 + \dfrac{vV'_x}{c^2}} \tag{1.51}$$

If $V'_x = c$, that is, if light is emitted in the moving frame S' in its direction of motion relative to S, an observer in frame S will measure the speed

$$V_x = \frac{V'_x + v}{1 + \dfrac{vV'_x}{c^2}} = \frac{c + v}{1 + \dfrac{vc}{c^2}} = \frac{c(c + v)}{c + v} = c$$

Thus observers in the car and on the road both find the same value for the speed of light, as they must.

Example 1.11

Spacecraft Alpha is moving at $0.90c$ with respect to the earth. If spacecraft Beta is to pass Alpha at a relative speed of $0.50c$ in the same direction, what speed must Beta have with respect to the earth?

Solution

According to the Galilean transformation, Beta would need a speed relative to the earth of $0.90c + 0.50c = 1.40c$, which we know is impossible. According to Eq. (1.49), however, with $V'_x = 0.50c$ and $v = 0.90c$, the required speed is only

$$V_x = \frac{V'_x + v}{1 + \dfrac{vV'_x}{c^2}} = \frac{0.50c + 0.90c}{1 + \dfrac{(0.90c)(0.50c)}{c^2}} = 0.97c$$

which is less than c. It is necessary to go less than 10 percent faster than a spacecraft traveling at $0.90c$ in order to pass it at a relative speed of $0.50c$.

Simultaneity

The relative character of time as well as space has many implications. Notably, events that seem to take place simultaneously to one observer may not be simultaneous to another observer in relative motion, and vice versa.

Let us examine two events—the setting off of a pair of flares, say—that occur at the same time t_0 to somebody on the earth but at the different locations x_1 and x_2. What does the pilot of a spacecraft in flight see? To her, the flare at x_1 and t_0 appears at the time

$$t_1' = \frac{t_0 - \boldsymbol{v}x_1/c^2}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

according to Eq. (1.44), while the flare at x_2 and t_0 appears at the time

$$t_2' = \frac{t_0 - \boldsymbol{v}x_2/c^2}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

Hence two events that occur simultaneously to one observer are separated by a time interval of

$$t_2' - t_1' = \frac{\boldsymbol{v}(x_1 - x_2)/c^2}{\sqrt{1 - \boldsymbol{v}^2/c^2}}$$

to an observer moving at the speed $\boldsymbol{v}$ relative to the other observer. Who is right? The question is, of course, meaningless: both observers are “right” since each simply measures what he or she sees.

Because simultaneity is a relative concept and not an absolute one, physical theories that require simultaneity in events at different locations cannot be valid. For instance, saying that total energy is conserved in an isolated system does not rule out a process in which an amount of energy ΔE vanishes at one place while an equal amount of energy ΔE comes into being somewhere else with no actual transport of energy from one place to the other. Because simultaneity is relative, some observers of the process will find energy not being conserved. To rescue conservation of energy in the light of special relativity, then, we have to say that, when energy disappears somewhere and appears elsewhere, it has actually flowed from the first location to the second. Thus energy is conserved *locally* everywhere, not merely when an isolated system is considered—a much stronger statement of this principle.

Appendix II to Chapter 1

Spacetime

As we have seen, the concepts of space and time are inextricably mixed in nature. A length that one observer can measure with only a meter stick may have to be measured with both a meter stick and a clock by another observer.

A convenient and elegant way to express the results of special relativity is to regard events as occurring in a four-dimensional **spacetime** in which the usual three coordinates x, y, z refer to space and a fourth coordinate ict refers to time, where $i = \sqrt{-1}$. Although we cannot visualize spacetime, it is no harder to deal with mathematically than three-dimensional space.

The reason that ict is chosen as the time coordinate instead of just t is that the quantity

$$s^2 = x^2 + y^2 + z^2 - (ct)^2 \tag{1.52}$$

is **invariant** under a Lorentz transformation. That is, if an event occurs at x, y, z, t in an inertial frame S and at x', y', z', t' in another inertial frame S', then

$$s^2 = x^2 + y^2 + z^2 - (ct)^2 = x'^2 + y'^2 + z'^2 - (ct')^2$$

Because s^2 is invariant, we can think of a Lorentz transformation merely as a rotation in spacetime of the coordinate axes x, y, z, ict (Fig. 1.24).

The four coordinates x, y, z, ict define a vector in spacetime, and this *four-vector* remains fixed in spacetime regardless of any rotation of the coordinate system—that is, regardless of any shift in point of view from one inertial frame S to another S'.

Another four-vector whose magnitude remains constant under Lorentz transformations has the components $p_x, p_y, p_z, iE/c$. Here p_x, p_y, p_z are the usual components of the linear momentum of a body whose total energy is E. Hence the value of

$$p_x^2 + p_y^2 + p_z^2 - \frac{E^2}{c}$$

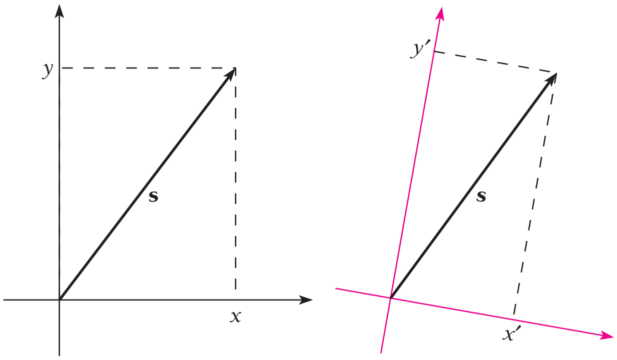


Figure 1.24 Rotating a two-dimensional coordinate system does not change the quantity $s^2 = x^2 + y^2 = x'^2 + y'^2$, where s is the length of the vector $\mathbf{s}$. This result can be generalized to the four-dimensional spacetime coordinate system x, y, z, ict.

is the same in all inertial frames even though p_x, p_y, p_z and E separately may be different. This invariance was noted earlier in connection with Eq. (1.24); we note that $p^2 = p_x^2 + p_y^2 + p_z^2$.

A more mathematically elaborate formulation brings together the electric and magnetic fields **E** and **B** into an invariant quantity called a tensor. This approach to incorporating special relativity into physics has led both to a deeper understanding of natural laws and to the discovery of new phenomena and relationships.

Spacetime Intervals

The statements made at the end of Sec. 1.2 (P. 10) are easy to confirm using the idea of spacetime. Figure 1.25 shows two events plotted on the axes x and ct. Event 1 occurs at $x = 0$, $t = 0$ and event 2 occurs at $x = \Delta x$, $t = \Delta t$. The *spacetime interval* Δs between them is defined by

Spacetime interval between events

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 \tag{1.53}$$

The virtue of this definition is that $(\Delta s)^2$, like the s^2 of Eq. 1.52, is invariant under Lorentz transformations. If Δx and Δt are the differences in space and time between two events measured in the S frame and $\Delta x'$ and $\Delta t'$ are the same quantities measured in the S' frame,

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 = (c\Delta t')^2 - (\Delta x')^2$$

Therefore whatever conclusions we arrive at in the S frame in which event 1 is at the origin hold equally well in any other frame in relative motion at constant velocity.

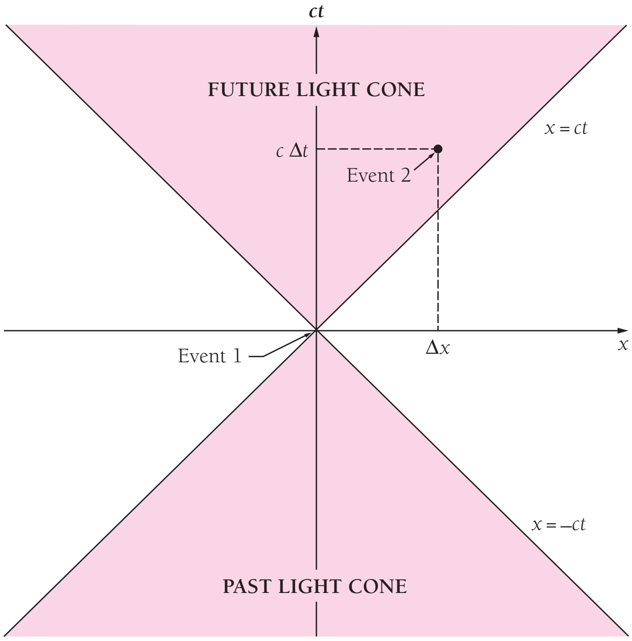


Figure 1.25 The past and future light cones in spacetime of event 1.

Now let us look into the possible relationships between events 1 and 2. Event 2 can be related causally in some way to event 1 provided that a signal traveling slower than the speed of light can connect these events, that is, provided that

$$c\Delta t > |\Delta x|$$

or

Timelike interval $$(\Delta s)^2 > 0 \tag{1.53}$$

An interval in which $(\Delta s)^2 > 0$ is said to be *timelike*. Every timelike interval that connects event 1 with another event lies within the *light cones* bounded by $x = \pm ct$ in Fig. 1.25. All events that could have affected event 1 lie in the past light cone; all events that event 1 is able to affect lie in the future light cone. (Events connected by timelike intervals need not *necessarily* be related, of course, but it is *possible* for them to be related.)

Conversely, the criterion for there being no causal relationship between events 1 and 2 is that

$$c\Delta t < |\Delta x|$$

or

Spacelike interval $$(\Delta s)^2 < 0 \tag{1.54}$$

An interval in which $(\Delta s)^2 < 0$ is said to be *spacelike*. Every event that is connected with event 1 by a spacelike interval lies outside the light cones of event 1 and neither has interacted with event 1 in the past nor is capable of interacting with it in the future; the two events must be entirely unrelated.

When events 1 and 2 can be connected with a light signal only,

$$c\Delta t = |\Delta x|$$

or

Lightlike interval $$\Delta s = 0 \tag{1.55}$$

An interval in which $\Delta s = 0$ is said to be *lightlike*. Events that can be connected with event 1 by lightlike intervals lie on the boundaries of the light cones.

These conclusions hold in terms of the light cones of event 2 because $(\Delta s)^2$ is invariant; for example, if event 2 is inside the past light cone of event 1, event 1 is inside the future light cone of event 2. In general, events that lie in the future of an event as seen in one frame of reference S lie in its future in every other frame S', and events that lie in the past of an event in S lie in its past in every other frame S'. Thus "future" and "past" have invariant meanings. However, "simultaneity" is an ambiguous concept, because all events that lie outside the past and future light cones of event 1 (that is, all events connected by spacelike intervals with event 1) can appear to occur simultaneously with event 1 in some particular frame of reference.

The path of a particle in spacetime is called its *world line* (Fig. 1.26). The world line of a particle must lie within its light cones.

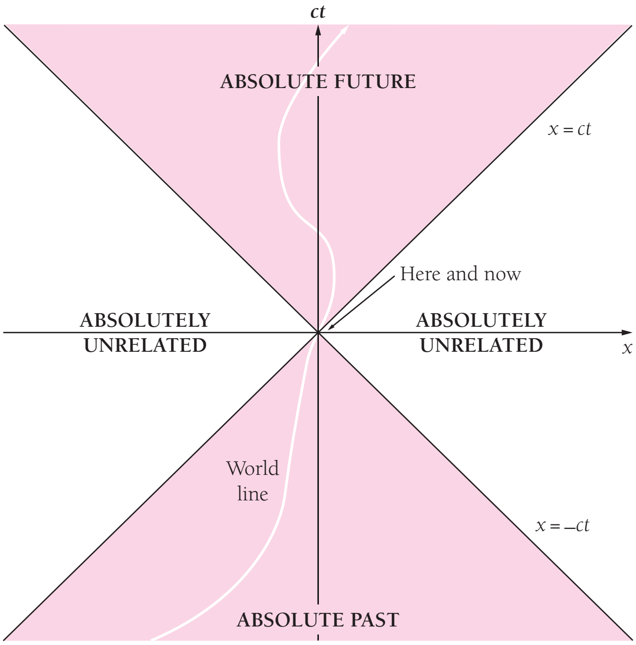


Figure 1.26 The world line of a particle in spacetime.

EXERCISES

But be ye doers of the word, and not hearers only, deceiving your own selves. —James I:22

1.1 Special Relativity

1. If the speed of light were smaller than it is, would relativistic phenomena be more or less conspicuous than they are now?

2. It is possible for the electron beam in a television picture tube to move across the screen at a speed faster than the speed of light. Why does this not contradict special relativity?

1.2 Time Dilation

3. An athlete has learned enough physics to know that if he measures from the earth a time interval on a moving spacecraft, what he finds will be greater than what somebody on the spacecraft would measure. He therefore proposes to set a world record for the 100-m dash by having his time taken by an observer on a moving spacecraft. Is this a good idea?

4. An observer on a spacecraft moving at $0.700c$ relative to the earth finds that a car takes 40.0 min to make a trip. How long does the trip take to the driver of the car?

5. Two observers, A on earth and B in a spacecraft whose speed is 2.00×10^8 m/s, both set their watches to the same time when the ship is abreast of the earth. (*a*) How much time must elapse by A's reckoning before the watches differ by 1.00 s? (*b*) To A, B's watch seems to run slow. To B, does A's watch seem to run fast, run slow, or keep the same time as his own watch?

6. An airplane is flying at 300 m/s (672 mi/h). How much time must elapse before a clock in the airplane and one on the ground differ by 1.00 s?

7. How fast must a spacecraft travel relative to the earth for each day on the spacecraft to correspond to 2 d on the earth?

8. The Apollo 11 spacecraft that landed on the moon in 1969 traveled there at a speed relative to the earth of 1.08×10^4 m/s. To an observer on the earth, how much longer than his own day was a day on the spacecraft?

9. A certain particle has a lifetime of 1.00×10^{-7} s when measured at rest. How far does it go before decaying if its speed is $0.99c$ when it is created?

1.3 Doppler Effect

10. A spacecraft receding from the earth at $0.97c$ transmits data at the rate of 1.00×10^4 pulses/s. At what rate are they received?

11. A galaxy in the constellation Ursa Major is receding from the earth at 15,000 km/s. If one of the characteristic wavelengths of the light the galaxy emits is 550 nm, what is the corresponding wavelength measured by astronomers on the earth?

12. The frequencies of the spectral lines in light from a distant galaxy are found to be two-thirds as great as those of the same lines in light from nearby stars. Find the recession speed of the distant galaxy.

13. A spacecraft receding from the earth emits radio waves at a constant frequency of 10^9 Hz. If the receiver on earth can measure frequencies to the nearest hertz, at what spacecraft speed can the difference between the relativistic and classical doppler effects be detected? For the classical effect, assume the earth is stationary.

14. A car moving at 150 km/h (93 mi/h) is approaching a stationary police car whose radar speed detector operates at a frequency of 15 GHz. What frequency change is found by the speed detector?

15. If the angle between the direction of motion of a light source of frequency ν_0 and the direction from it to an observer is θ, the frequency ν the observer finds is given by

$$\nu = \nu_0 \frac{\sqrt{1 - v^2/c^2}}{1 - (v/c)\cos\theta}$$

where v is the relative speed of the source. Show that this formula includes Eqs. (1.5) to (1.7) as special cases.

16. (*a*) Show that when $v \ll c$, the formulas for the doppler effect both in light and in sound for an observer approaching a source, and vice versa, all reduce to $\nu \approx \nu_0(1 + v/c)$, so that $\Delta\nu/\nu \approx v/c$. [*Hint*: For $x \ll 1$, $1/(1 + x) \approx 1 - x$.] (*b*) What do the formulas for an observer receding from a source, and vice versa, reduce to when $v \ll c$?

1.4 Length Contraction

17. An astronaut whose height on the earth is exactly 6 ft is lying parallel to the axis of a spacecraft moving at 0.90*c* relative to the earth. What is his height as measured by an observer in the same spacecraft? By an observer on the earth?

18. An astronaut is standing in a spacecraft parallel to its direction of motion. An observer on the earth finds that the spacecraft speed is 0.60*c* and the astronaut is 1.3 m tall. What is the astronaut's height as measured in the spacecraft?

19. How much time does a meter stick moving at 0.100*c* relative to an observer take to pass the observer? The meter stick is parallel to its direction of motion.

20. A meter stick moving with respect to an observer appears only 500 mm long to her. What is its relative speed? How long does it take to pass her? The meter stick is parallel to its direction of motion.

21. A spacecraft antenna is at an angle of 10° relative to the axis of the spacecraft. If the spacecraft moves away from the earth at a speed of 0.70*c*, what is the angle of the antenna as seen from the earth?

1.5 Twin Paradox

22. Twin *A* makes a round trip at 0.6*c* to a star 12 light-years away, while twin *B* stays on the earth. Each twin sends the other a signal once a year by his own reckoning. (*a*) How many signals does *A* send during the trip? How many does *B* send? (*b*) How many signals does *A* receive? How many does *B* receive?

23. A woman leaves the earth in a spacecraft that makes a round trip to the nearest star, 4 light-years distant, at a speed of 0.9*c*. How much younger is she upon her return than her twin sister who remained behind?

1.7 Relativistic Momentum

24. (*a*) An electron's speed is doubled from 0.2*c* to 0.4*c*. By what ratio does its momentum increase? (*b*) What happens to the momentum ratio when the electron's speed is doubled again from 0.4*c* to 0.8*c*?

25. All definitions are arbitrary, but some are more useful than others. What is the objection to defining linear momentum as $\mathbf{p} = m\mathbf{v}$ instead of the more complicated $\mathbf{p} = \gamma m\mathbf{v}$?

26. Verify that

$$\frac{1}{\sqrt{1 - v^2/c^2}} = 1 + \frac{p^2}{m^2c^2}$$

1.8 Mass and Energy

27. Dynamite liberates about 5.4×10^6 J/kg when it explodes. What fraction of its total energy content is this?

28. A certain quantity of ice at 0°C melts into water at 0°C and in so doing gains 1.00 kg of mass. What was its initial mass?

29. At what speed does the kinetic energy of a particle equal its rest energy?

30. How many joules of energy per kilogram of rest mass are needed to bring a spacecraft from rest to a speed of 0.90*c*?

31. An electron has a kinetic energy of 0.100 MeV. Find its speed according to classical and relativistic mechanics.

32. Verify that, for $E \gg E_0$,

$$\frac{v}{c} \approx 1 - \frac{1}{2}\left(\frac{E_0}{E}\right)^2$$

33. A particle has a kinetic energy 20 times its rest energy. Find the speed of the particle in terms of *c*.

34. (*a*) The speed of a proton is increased from 0.20*c* to 0.40*c*. By what factor does its kinetic energy increase? (*b*) The proton speed is again doubled, this time to 0.80*c*. By what factor does its kinetic energy increase now?

35. How much work (in MeV) must be done to increase the speed of an electron from 1.2×10^8 m/s to 2.4×10^8 m/s?

36. (*a*) Derive a formula for the minimum kinetic energy needed by a particle of rest mass *m* to emit Cerenkov radiation in a medium of index of refraction *n*. [*Hint*: Start from Eqs. (1.21) and (1.23).] (*b*) Use this formula to find KE_{min} for an electron in a medium of $n = 1.5$.

37. Prove that $\frac{1}{2}\gamma m v^2$, does *not* equal the kinetic energy of a particle moving at relativistic speeds.

38. A moving electron collides with a stationary electron and an electron-positron pair comes into being as a result (a positron is a positively charged electron). When all four particles have the same velocity after the collision, the kinetic energy required for this process is a minimum. Use a relativistic calculation to show that $\text{KE}_{\text{min}} = 6mc^2$, where *m* is the rest mass of the electron.

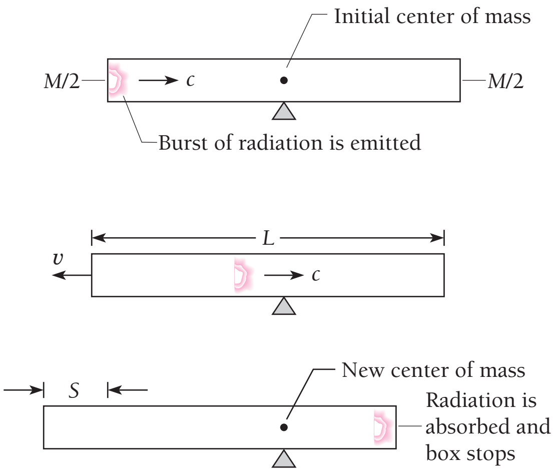

Initial center of mass
M/2
c
M/2
Burst of radiation is emitted
L
v
c
S
New center of mass
Radiation is absorbed and box stops

Figure 1.27 The box has moved the distance S to the left when it stops.

39. An alternative derivation of the mass-energy formula $E_0 = mc^2$, also given by Einstein, is based on the principle that the location of the center of mass (CM) of an isolated system cannot be changed by any process that occurs inside the system. Figure 1.27 shows a rigid box of length L that rests on a frictionless surface; the mass M of the box is equally divided between its two ends. A burst of electromagnetic radiation of energy E_0 is emitted by one end of the box. According to classical physics, the radiation has the momentum $p = E_0/c$, and when it is emitted, the box recoils with the speed $\boldsymbol{v} \approx E_0/Mc$ so that the total momentum of the system remains zero. After a time $t \approx L/c$ the radiation reaches the other end of the box and is absorbed there, which brings the box to a stop after having moved the distance S. If the CM of the box is to remain in its original place, the radiation must have transferred mass from one end to the other. Show that this amount of mass is $m = E_0/c^2$.

1.9 Energy and Momentum

40. Find the SI equivalents of the mass unit MeV/c^2 and the momentum unit MeV/c.

41. In its own frame of reference, a proton takes 5 min to cross the Milky Way galaxy, which is about 10^5 light-years in diameter. (*a*) What is the approximate energy of the proton in electronvolts? (*b*) About how long would the proton take to cross the galaxy as measured by an observer in the galaxy's reference frame?

42. What is the energy of a photon whose momentum is the same as that of a proton whose kinetic energy is 10.0 MeV?

43. Find the momentum (in MeV/c) of an electron whose speed is $0.600c$.

44. Find the total energy and kinetic energy (in GeV) and the momentum (in GeV/c) of a proton whose speed is $0.900c$. The mass of the proton is 0.938 GeV/c^2.

45. Find the momentum of an electron whose kinetic energy equals its rest energy of 511 keV.

46. Verify that $\boldsymbol{v}/c = pc/E$.

47. Find the speed and momentum (in GeV/c) of a proton whose total energy is 3.500 GeV.

48. Find the total energy of a neutron (m = 0.940 GeV/c^2) whose momentum is 1.200 GeV/c.

49. A particle has a kinetic energy of 62 MeV and a momentum of 335 MeV/c. Find its mass (in MeV/c^2) and speed (as a fraction of c).

50. (*a*) Find the mass (in GeV/c^2) of a particle whose total energy is 4.00 GeV and whose momentum is 1.45 GeV/c. (*b*) Find the total energy of this particle in a reference frame in which its momentum is 2.00 GeV/c.

Appendix I: The Lorentz Transformation

51. An observer detects two explosions, one that occurs near her at a certain time and another that occurs 2.00 ms later 100 km away. Another observer finds that the two explosions occur at the same place. What time interval separates the explosions to the second observer?

52. An observer detects two explosions that occur at the same time, one near her and the other 100 km away. Another observer finds that the two explosions occur 160 km apart. What time interval separates the explosions to the second observer?

53. A spacecraft moving in the $+x$ direction receives a light signal from a source in the xy plane. In the reference frame of the fixed stars, the speed of the spacecraft is $\boldsymbol{v}$ and the signal arrives at an angle θ to the axis of the spacecraft. (*a*) With the help of the Lorentz transformation find the angle θ' at which the signal arrives in the reference frame of the spacecraft. (*b*) What would you conclude from this result about the view of the stars from a porthole on the side of the spacecraft?

54. A body moving at $0.500c$ with respect to an observer disintegrates into two fragments that move in opposite directions relative to their center of mass along the same line of motion as the original body. One fragment has a velocity of $0.600c$ in the backward direction relative to the center of mass and the other has a velocity of $0.500c$ in the forward direction. What velocities will the observer find?

55. A man on the moon sees two spacecraft, A and B, coming toward him from opposite directions at the respective speeds of $0.800c$ and $0.900c$. (*a*) What does a man on A measure for the speed with which he is approaching the moon? For the speed with which he is approaching B? (*b*) What does a man on B measure for the speed with which he is approaching the moon? For the speed with which he is approaching A?

56. An electron whose speed relative to an observer in a laboratory is $0.800c$ is also being studied by an observer moving in the same direction as the electron at a speed of $0.500c$ relative to the laboratory. What is the kinetic energy (in MeV) of the electron to each observer?

Class: B. Tech (Unit IV)

I have taken all course materials for Unit IV from Book Concept of Modern Physics by Arthur Besier, Shobhit Mahajan & S. Rai Choudhury (McGraw Hill Education).

Students can download this book form given web address;

Web Address : **https://b-ok.cc/book/2700591/864ac0**

Some topics (mainly LASER) of unit IV (Laser and Fiber Optics) have been taken from **Chapter4** from above said book (**https://b-ok.cc/book/2700591/864ac0**). I am sending pdf file of Chapter 4 which have **LASER** notes.

UNIT-IV: LASER & FIBER OPTICS

Introduction; Absorption and Emission, Einstein's coefficients & equations; Metastable states, Population inversion, Pumping (three and four level laser schemes), Basic parts of a Laser, Characteristics of Laser Radiations; Classification of Lasers, Ruby Laser, He-Ne Laser, GaAs Laser; Applications of lasers in holography

Basics of optical fiber, Total Internal Reflection, Acceptance angle, Numerical Aperture; Modes of Propagation, Single Mode Step Index Optical Fiber, Multimode Step Index Optical Fiber, Graded Index Fiber, Losses, Dispersion in Optical Fiber, Intermodal and intramodal dispersion, Applications of optical fiber; Problems**.**

where m is the electron mass. From Eq. (4.23) the energy levels of a positronium "atom" are

$$E'_n = \left(\frac{m'}{m}\right)\frac{E_1}{n^2} = \frac{E_1}{2n^2}$$

This means that the Rydberg constant—the constant term in Eq. (4.18)—for positronium is half as large as it is for ordinary hydrogen. As a result the wavelengths in the positronium spectral lines are all twice those of the corresponding lines in the hydrogen spectrum.

Example 4.7

A **muon** is an unstable elementary particle whose mass is $207m_e$ and whose charge is either $+e$ or $-e$. A negative muon (μ^-) can be captured by a nucleus to form a muonic atom. (*a*) A proton captures a μ^-. Find the radius of the first Bohr orbit of this atom. (*b*) Find the ionization energy of the atom.

Solution

(*a*) Here $m = 207m_e$ and $M = 1836m_e$, so the reduced mass is

$$m' = \frac{mM}{m+M} = \frac{(207m_e)(1836m_e)}{207m_e + 1836m_e} = 186m_e$$

According to Eq. (4.13) the orbit radius corresponding to $n = 1$ is

$$r_1 = \frac{h^2\epsilon_0}{\pi m_e e^2}$$

where $r_1 = a_0 = 5.29 \times 10^{-11}$ m. Hence the radius r' that corresponds to the reduced mass m' is

$$r'_1 = \left(\frac{m}{m'}\right)r_1 = \left(\frac{m_e}{186m_e}\right)a_0 = 2.85 \times 10^{-13}\text{ m}$$

The muon is 186 times closer to the proton than an electron would be, so a muonic hydrogen atom is much smaller than an ordinary hydrogen atom.

(*b*) From Eq. (4.23) we have, with $n = 1$ and $E_1 = -13.6$ eV,

$$E'_1 = \left(\frac{m'}{m}\right)E_1 = 186E_1 = -2.53 \times 10^3\text{ eV} = -2.53\text{ keV}$$

The ionization energy is therefore 2.53 keV, 186 times that for an ordinary hydrogen atom.

4.8 ATOMIC EXCITATION

How atoms absorb and emit energy

There are two main ways in which an atom can be excited to an energy above its ground state and thereby become able to radiate. One of these ways is by a collision with another particle in which part of their joint kinetic energy is absorbed by the atom. Such an excited atom will return to its ground state in an average of 10^{-8} s by emitting one or more photons (Fig. 4.18).

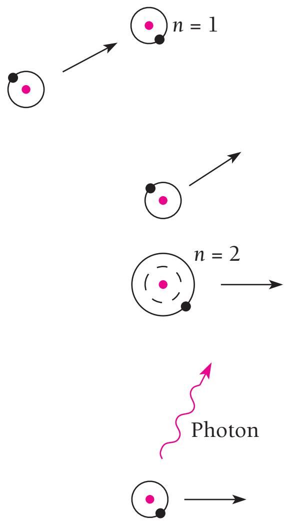


Figure 4.18 Excitation by collision. Some of the available energy is absorbed by one of the atoms, which goes into an excited energy state. The atom then emits a photon in returning to its ground (normal) state.

To produce a luminous discharge in a rarefied gas, an electric field is established that accelerates electrons and atomic ions until their kinetic energies are sufficient to

Auroras are caused by streams of fast protons and electrons from the sun that excite atoms in the upper atmosphere. The green hues of an auroral display come from oxygen, and the reds originate in both oxygen and nitrogen. This aurora occurred in Alaska.

excite atoms they collide with. Because energy transfer is a maximum when the colliding particles have the same mass (see Fig. 12.22), the electrons in such a discharge are more effective than the ions in providing energy to atomic electrons. Neon signs and mercury-vapor lamps are familiar examples of how a strong electric field applied between electrodes in a gas-filled tube leads to the emission of the characteristic spectral radiation of that gas, which happens to be reddish light in the case of neon and bluish light in the case of mercury vapor.

Another excitation mechanism is involved when an atom absorbs a photon of light whose energy is just the right amount to raise the atom to a higher energy level. For example, a photon of wavelength 121.7 nm is emitted when a hydrogen atom in the $n = 2$ state drops to the $n = 1$ state. Absorbing a photon of wavelength 121.7 nm by a hydrogen atom initially in the $n = 1$ state will therefore bring it up to the $n = 2$ state (Fig. 4.19). This process explains the origin of absorption spectra.

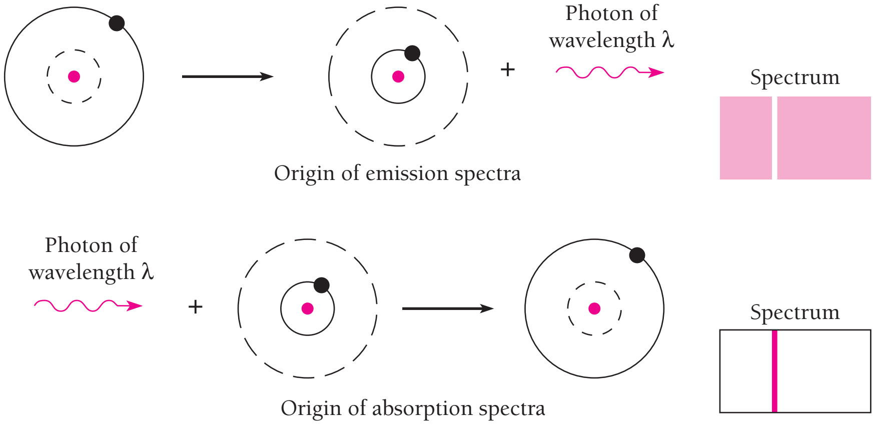


Figure 4.19 How emission and absorption spectral lines originate.

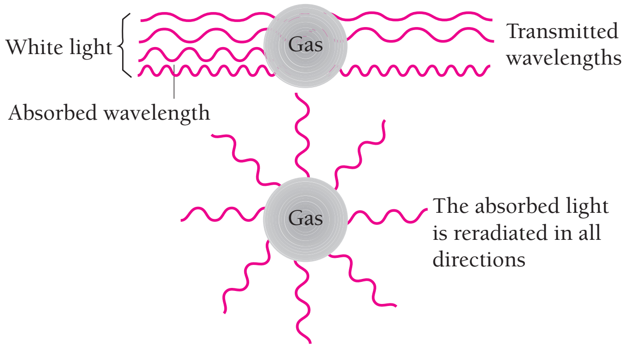


Figure 4.20 The dark lines in an absorption spectrum are never totally dark.

When white light, which contains all wavelengths, is passed through hydrogen gas, photons of those wavelengths that correspond to transitions between energy levels are absorbed. The resulting excited hydrogen atoms reradiate their excitation energy almost at once, but these photons come off in random directions with only a few in the same direction as the original beam of white light (Fig. 4.20). The dark lines in an absorption spectrum are therefore never completely black but only appear so by contrast with the bright background. We expect the lines in the absorption spectrum of any element to coincide with those in its emission spectrum that represent transitions to the ground state, which agrees with observation (see Fig. 4.9).

Franck-Hertz Experiment

Atomic spectra are not the only way to investigate energy levels inside atoms. A series of experiments based on excitation by collision was performed by James Franck and Gustav Hertz (a nephew of Heinrich Hertz) starting in 1914. These experiments demonstrated that atomic energy levels indeed exist and, furthermore, that the ones found in this way are the same as those suggested by line spectra.

Franck and Hertz bombarded the vapors of various elements with electrons of known energy, using an apparatus like that shown in Fig. 4.21. A small potential difference V_0 between the grid and collecting plate prevents electrons having energies less than a certain minimum from contributing to the current I through the ammeter. As the accelerating potential V is increased, more and more electrons arrive at the plate and I rises (Fig. 4.22).

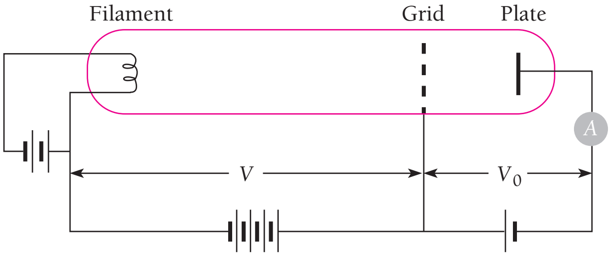


Figure 4.21 Apparatus for the Franck-Hertz experiment.

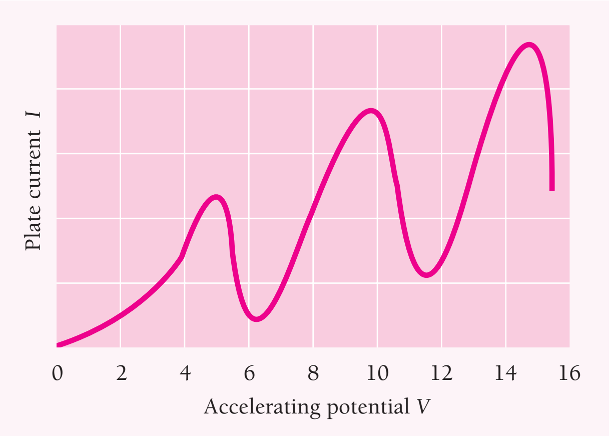


Figure 4.22 Results of the Franck-Hertz experiment, showing critical potentials in mercury vapor.

If KE is conserved when an electron collides with one of the atoms in the vapor, the electron merely bounces off in a new direction. Because an atom is much heavier than an electron, the electron loses almost no KE in the process. After a certain critical energy is reached, however, the plate current drops abruptly. This suggests that an electron colliding with one of the atoms gives up some or all of its KE to excite the atom to an energy level above its ground state. Such a collision is called inelastic, in contrast to an elastic collision in which KE is conserved. The critical electron energy equals the energy needed to raise the atom to its lowest excited state.

Then, as the accelerating potential V is raised further, the plate current again increases, since the electrons now have enough energy left to reach the plate after undergoing an inelastic collision on the way. Eventually another sharp drop in plate current occurs, which arises from the excitation of the same energy level in other atoms by the electrons. As Fig. 4.22 shows, a series of critical potentials for a given atomic vapor is obtained. Thus the higher potentials result from two or more inelastic collisions and are multiples of the lowest one.

To check that the critical potentials were due to atomic energy levels, Franck and Hertz observed the emission spectra of vapors during electron bombardment. In the case of mercury vapor, for example, they found that a minimum electron energy of 4.9 eV was required to excite the 253.6-nm spectral line of mercury—and a photon of 253.6-nm light has an energy of just 4.9 eV. The Franck-Hertz experiments were performed shortly after Bohr announced his theory of the hydrogen atom, and they independently confirmed his basic ideas.

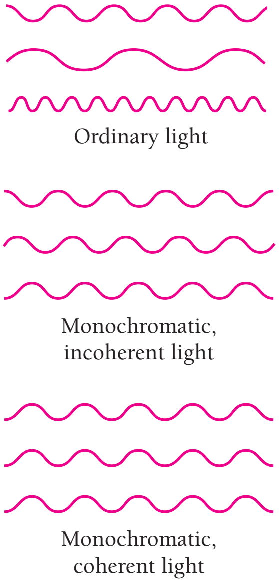

Ordinary light
Monochromatic, incoherent light
Monochromatic, coherent light

Figure 4.23 A laser produces a beam of light whose waves all have the same frequency (monochromatic) and are in phase with one another (coherent). The beam is also well collimated and so spreads out very little, even over long distances.

4.9 THE LASER

How to produce light waves all in step

The **laser** is a device that produces a light beam with some remarkable properties:

1 The light is very nearly monochromatic.
2 The light is coherent, with the waves all exactly in phase with one another (Fig.4.23).

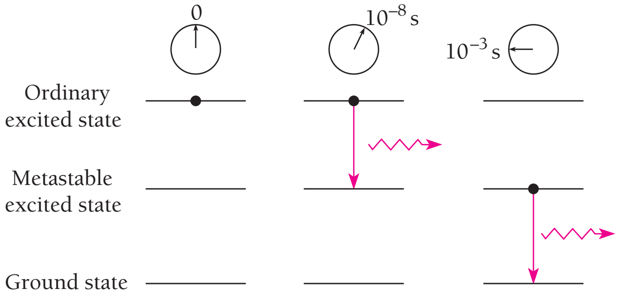


Figure 4.24 An atom can exist in a metastable energy level for a longer time before radiating than it can in an ordinary energy level.

3 A laser beam diverges hardly at all. Such a beam sent from the earth to a mirror left on the moon by the Apollo 11 expedition remained narrow enough to be detected on its return to the earth, a total distance of over three-quarters of a million kilometers. A light beam produced by any other means would have spread out too much for this to be done.

4 The beam is extremly intense, more intense by far than the light from any other source. To achieve an energy density equal to that in some laser beams, a hot object would have to be at a temperature of 10^{30} K.

The last two of these properties follow from the second of them.

The term *laser* stands for *l*ight *a*mplification by *s*timulated *e*mission of *r*adiation. The key to the laser is the presence in many atoms of one or more excited energy levels whose lifetimes may be 10^{-3} s or more instead of the usual 10^{-8} s. Such relatively long-lived states are called **metastable** (temporarily stable); see Fig. 4.24.

Three kinds of transition involving electromagnetic radiation are possible between two energy levels, E_0 and E_1, in an atom (Fig. 4.25). If the atom is initially in the lower state E_0, it can be raised to E_1 by absorbing a photon of energy $E_1 - E_0 = h\nu$. This process is called **stimulated absorption.** If the atom is initially in the upper state E_1, it can drop to E_0 by emitting a photon of energy $h\nu$. This is **spontaneous emission.**

Einstein, in 1917, was the first to point out a third possibility, **stimulated emission,** in which an incident photon of energy $h\nu$ causes a transition from E_1 to E_0. In stimulated emission, the radiated light waves are exactly in phase with the incident ones, so the result is an enhanced beam of coherent light. Einstein showed that stimulated emission has the same probability as stimulated absorption (see Sec. 9.7). That is, a photon of energy $h\nu$ incident on an atom in the upper

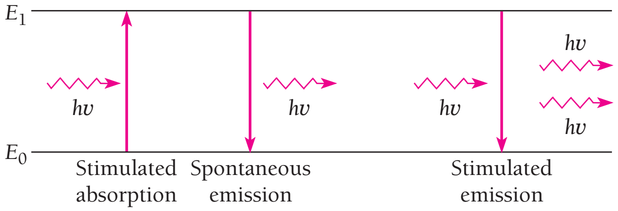


Figure 4.25 Transitions between two energy levels in an atom can occur by stimulated absorption, spontaneous emission, and stimulated emission.

Charles H. Townes (1915–) was born in Greenville, South Carolina, and attended Furman University there. After graduate study at Duke University and the California Institute of Technology, he spent 1939 to 1947 at the Bell Telephone Laboratories designing radar-controlled bombing systems. Townes then joined the physics department of Columbia University. In 1951, while sitting on a park bench, the idea for the **maser** (*m*icrowave *a*mplification by *s*timulated *e*mission of *r*adiation) occurred to him as a way to produce high-intensity microwaves, and in 1953 the first maser began operating. In this device ammonia (NH_3) molecules were raised to an excited vibrational state and then fed into a resonant cavity where, as in a laser, stimulated emission produced a cascade of photons of identical wavelength, here 1.25 cm in the microwave part of the spectrum. "Atomic clocks" of great accuracy are based on this concept, and solid-state maser amplifiers are used in such applications as radioastronomy.

In 1958 Townes and Arthur Schawlow attracted much attention with a paper showing that a similar scheme ought to be possible at optical wavelengths. Slightly earlier Gordon Gould, then a graduate student at Columbia, had come to the same conclusion, but did not publish his calculations at once since that would prevent securing a patent. Gould tried to develop the laser—his term—in private industry, but the Defense Department classified as secret the project (and his original notebooks) and denied him clearance to work on it. Finally, twenty years later, Gould succeeded in establishing his priority and received two patents on the laser, and still later, a third. The first working laser was built by Theodore Maiman at Hughes Research Laboratories in 1960. In 1964 Townes, along with two Russian laser pioneers, Aleksander Prokhorov and Nikolai Basov, was awarded a Nobel Prize. In 1981 Schawlow shared a Nobel Prize for precision spectroscopy using lasers.

Soon after its invention, the laser was spoken of as a "solution looking for a problem" because few applications were then known for it. Today, of course, lasers are widely employed for a variety of purposes.

state E_1 has the same likelihood of causing the emission of another photon of energy $h\nu$ as its likelihood of being absorbed if it is incident on an atom in the lower state E_0.

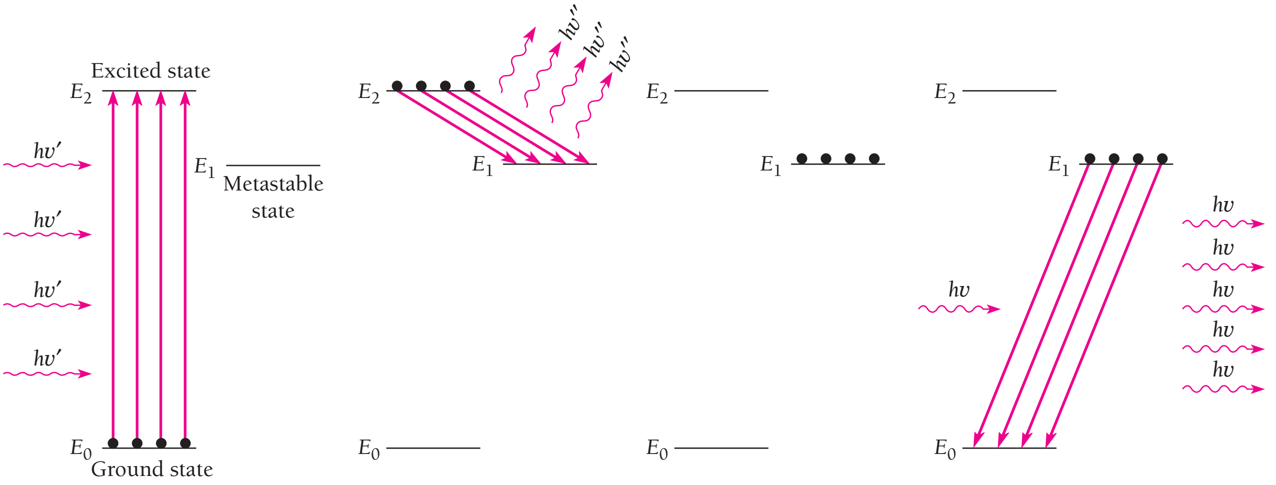
Stimulated emission involves no novel concepts. An analogy is a harmonic oscillator, for instance a pendulum, which has a sinusoidal force applied to it whose period is the same as its natural period of vibration. If the applied force is exactly in phase with the pendulum swings, the amplitude of the swings increases. This corresponds to stimulated absorption. However, if the applied force is 180° out of phase with the pendulum swings, the amplitude of the swings *decreases*. This corresponds to stimulated emission.

A **three-level laser,** the simplest kind, uses an assembly of atoms (or molecules) that have a metastable state $h\nu$ in energy above the ground state and a still higher excited state that decays to the metastable state (Fig. 4.26). What we want is more atoms in the metastable state than in the ground state. If we can arrange this and then shine light of frequency ν on the assembly, there will be more stimulated emissions from atoms in the metastable state than stimulated absorptions by atoms in the ground state. The result will be an amplification of the original light. This is the concept that underlies the operation of the laser.

The term **population inversion** describes an assembly of atoms in which the majority are in energy levels above the ground state; normally the ground state is occupied to the greatest extent.

A number of ways exist to produce a population inversion. One of them, called **optical pumping,** is illustrated in Fig. 4.27. Here an external light source is used some of whose photons have the right frequency to raise ground-state atoms to the excited state that decays spontaneously to the desired metastable state.

Why are three levels needed? Suppose there are only two levels, a metastable state $h\nu$ above the ground state. The more photons of frequency ν we pump into the assembly



Atoms in ground state are pumped to state E_2 by photons of $h\nu' = E_2 - E_0$ (or by collisions).

Rapid transition to metastable state E_1 by spontaneous emission of photons of $h\nu'' = E_2 - E_1$ (or in some other way).

Metastable states occupied in many atoms.

Stimulated emission occurs when photons of $h\nu = E_1 - E_0$ are incident, with the secondary photons themselves inducing further transitions to produce an avalanche of coherent photons.

Figure 4.26 The principle of the laser.

of atoms, the more upward transitions there will be from the ground state to the metastable state. However, at the same time the pumping will stimulate downward transitions from the metastable state to the ground state. When half the atoms are in each state, the rate of stimulated emissions will equal the rate of stimulated absorptions, so the assembly cannot ever have more than half its atoms in the metastable state. In this situation laser amplification cannot occur. A population inversion is only possible when the stimulated absorptions are to a higher energy level than the metastable one from which the stimulated emission takes place, which prevents the pumping from depopulating the metastable state.

In a three-level laser, more than half the atoms must be in the metastable state for stimulated induced emission to predominate. This is not the case for a **four-level laser.**

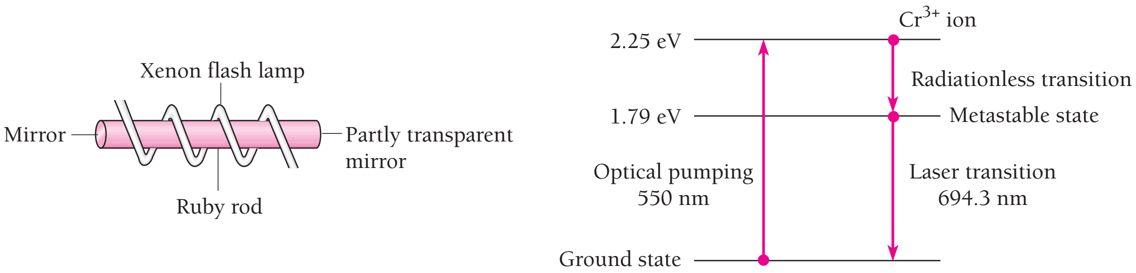


Figure 4.27 The ruby laser. In order for stimulated emission to exceed stimulated absorption, more than half the Cr^{3+} ions in the ruby rod must be in the metastable state. This laser produces a pulse of red light after each flash of the lamp.

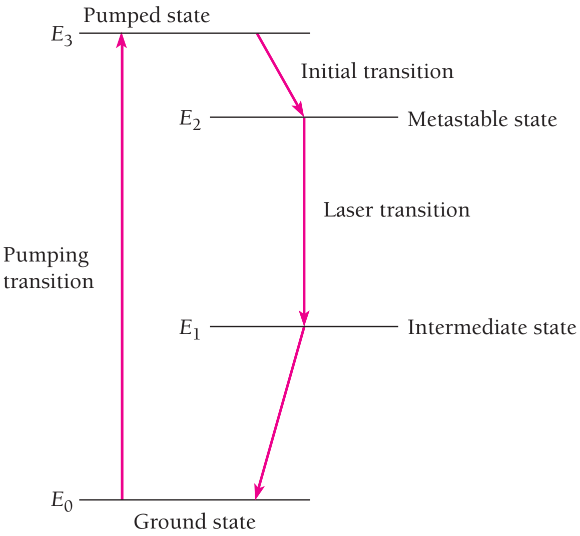
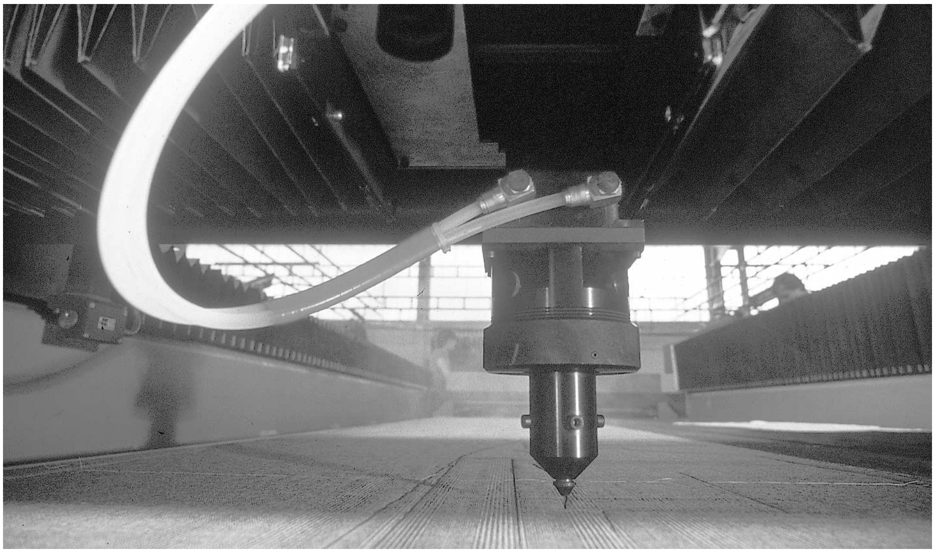


Figure 4.28 A four-level laser.

As in Fig. 4.28, the laser transition from the metastable state ends at an unstable intermediate state rather than at the ground state. Because the intermediate state decays rapidly to the ground state, very few atoms are in the intermediate state. Hence even a modest amount of pumping is enough to populate the metastable state to a greater extent than the intermediate state, as required for laser amplification.

Practical Lasers

The first successful laser, the **ruby laser,** is based on the three energy levels in the chromium ion Cr^{3+} shown in Fig. 4.27. A ruby is a crystal of aluminum oxide, Al_2O_3,



A robot arm carries a laser for cutting fabric in a clothing factory.

in which some of the Al^{3+} ions are replaced by Cr^{3+} ions, which are responsible for the red color. A Cr^{3+} ion has a metastable level whose lifetime is about 0.003 s. In the ruby laser, a xenon flash lamp excites the Cr^{3+} ions to a level of higher energy from which they fall to the metastable level by losing energy to other ions in the crystal. Photons from the spontaneous decay of some Cr^{3+} ions are reflected back and forth between the mirrored ends of the ruby rod, stimulating other excited Cr^{3+} ions to radiate. After a few microseconds the result is a large pulse of monochromatic, coherent red light from the partly transparent end of the rod.

The rod's length is made precisely an integral number of half-wavelengths long, so the radiation trapped in it forms an optical standing wave. Since the stimulated emissions are induced by the standing wave, their waves are all in step with it.

The common **helium-neon gas laser** achieves a population inversion in a different way. A mixture of about 10 parts of helium and 1 part of neon at a low pressure (~1 torr) is placed in a glass tube that has parallel mirrors, one of them partly transparent, at both ends. The spacing of the mirrors is again (as in all lasers) equal to an integral number of half-wavelengths of the laser light. An electric discharge is produced in the gas by means of electrodes outside the tube connected to a source of high-frequency alternating current, and collisions with electrons from the discharge excite He and Ne atoms to metastable states respectively 20.61 and 20.66 eV above their ground states (Fig. 4.29). Some of the excited He atoms transfer their energy to ground-state Ne atoms in collisions, with the 0.05 eV of additional energy being provided by the kinetic energy of the atoms. The purpose of the He atoms is thus to help achieve a population inversion in the Ne atoms.

The laser transition in Ne is from the metastable state at 20.66 eV to an excited state at 18.70 eV, with the emission of a 632.8-nm photon. Then another photon is spontaneously emitted in a transition to a lower metastable state; this transition yields only incoherent light. The remaining excitation energy is lost in collisions with the tube walls. Because the electron impacts that excite the He and Ne atoms occur all the time, unlike the pulsed excitation from the xenon flash lamp in a ruby laser, a He-Ne laser operates continuously. This is the laser whose narrow red beam is used in supermarkets to read bar codes. In a He-Ne laser, only a tiny

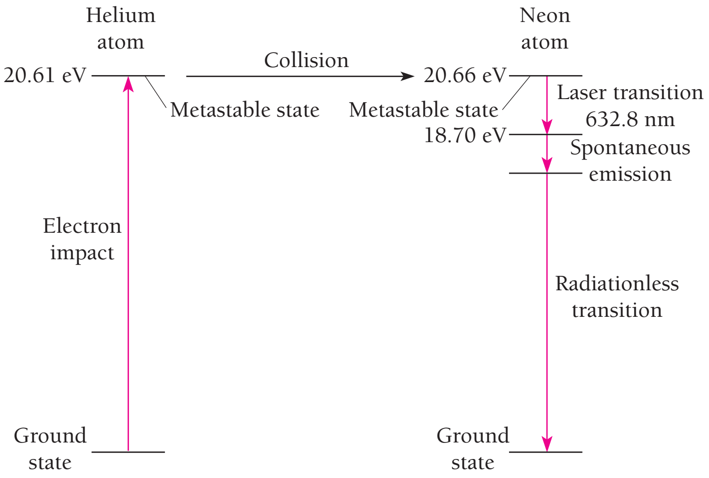


Figure 4.29 The helium-neon laser. In a four-level laser such as this, continuous operation is possible. Helium-neon lasers are commonly used to read bar codes.