Classical Electromagnetic Theory

MIKE GUIDRY

Contents

Preface

The material contained in these lecture notes represents an introduction to classical electromagnetism, written approximately at the level of classical texts such as Jackson [15]. It is suitable for a graduate course in classical electrodynamics and assumes students to have a basic familiarity with the material summarized in Appendix A, which would typically be covered in advanced undergraduate courses in electromagnetism (for example, the book by Griffiths [8]). These lectures are a work in progress, so please do not distribute them without notifying me.

Mike Guidry
Knoxville, Tennessee
April 25, 2024

1 Overview

The first inkling of the properties that we now understand of electricity and magnetism traces to the distant past when humans began to realize and remark upon the behavior of naturally occurring electricity in amber, and the properties of naturally occurring magnetism in lodestones. Although these properties were known qualitatively to the ancient Greeks, the modern quantitative understanding of electricity and magnetism emerged over a period of only about a century, beginning in the late 1700s.

1.1 The Synthesis of Classical Electromagnetism

The explosion in knowledge of electricity and magnetism, and the forging of that knowledge into a theoretically and mathematically coherent understanding, was highlighted by

1. Cavendish's pioneering experiments in electrostatics in the early 1770s,
2. the publication of Coulomb's work beginning in 1785,
3. Faraday's study of time-varying currents and magnetic fields in the mid-1800s,
4. Maxwell's famous 1865 paper (read to the Royal Society in 1864) synthesizing in equations a dynamical theory of the electromagnetic field [22],
5. Heaviside's reformulation of Maxwell's equations in their more modern vector calculus form in the late 1800s, and
6. Herz's publication in 1888 of data showing that *transverse electromagnetic waves propagated at the speed of light*, putting Maxwell's theory on a firm experimental footing.

Thus, by the late 19th century the basic understanding of classical electromagnetic theory was in place, and could be summarized concisely in the *Maxwell equations*, which may be written in free space using SI units (see Section 2.1) and modern notation[1] as

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)}, \tag{1.1a}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{1.1b}$$

$$\nabla \cdot \boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{1.1c}$$

$$\nabla \times \boldsymbol{B} - \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} = \mu_0 \boldsymbol{J} \qquad \text{(Ampère's law, as modified by Maxwell)}, \tag{1.1d}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, ρ is the charge density, $\boldsymbol{J}$ is the current vector, ε_0 is the *permittivity of free space* [defined for SI units in Eq. (2.3)], and μ_0 is the *permeability of free space*

$$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N\,A^{-2}}, \tag{1.2}$$

(with $\mu_0\varepsilon_0 = 1/c^2$, where c is the speed of light), with the charge density ρ and the current vector $\boldsymbol{J}$ satisfying the *continuity equation*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{J} = 0, \tag{1.3}$$

which ensures *conservation of charge* by requiring that electrical charge variation in some arbitrary volume is caused by flow of electrical current through the surface of that volume. An important property of the Maxwell equations is that they obey a set of symmetries that we expect physical laws to obey, as summarized in Box 1.2.[2]

The four Maxwell equations of Eqs. (1.1), supplemented by boundary conditions,[3] may be solved (in principle) for the fields $\boldsymbol{E}$ and $\boldsymbol{B}$, if the charge density ρ and current $\boldsymbol{J}$ are known. However, these equations make no reference to forces, which are often the connection to experimental results. This is remedied by introducing the *Lorentz force law*,

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \qquad \text{(Lorentz force)}, \tag{1.4}$$

which implies that the force on a particle with charge q and velocity $\boldsymbol{v}$ at the point $\boldsymbol{x}$ is determined completely by the instantaneous values of the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ at $\boldsymbol{x}$. Then Newton's

[1] Maxwell's original formulation was in terms of the vector and scalar potentials, required 20 equations, and did not use modern vector calculus notation, which was invented later by Oliver Heaviside and independently by J. Willard Gibbs. Heaviside reduced Maxwell's 20 equations to the four in Eqs. (1.1) by writing them entirely in terms of the electric and magnetic fields in his new vector calculus notation (see Ref. [14] for a concise history).

[2] The Maxwell equations written in the form (1.1) are *Lorentz covariant* (their validity is unchanged under Lorentz transformations) and thus are *consistent with special relativity*. However, in the form (1.1) this isn't obvious because in special relativity space and time must enter on an equal footing and the use in Eqs. (1.1) of 3-vectors instead of 4-vectors, and of separate derivatives for space and time, obscures whether this condition is satisfied. Later we will show that the Maxwell equations can be expressed in a covariant notation that makes their Lorentz covariance manifest.

[3] We will have more to say about this later but loosely physically acceptable boundary conditions in static problems require fields to vanish rapidly enough at infinity, and in dynamical problems that they represent outgoing solutions so that there is no flow of energy from infinity into regions of interest.

Box 1.1 Maxwell's Modification of Ampère's law

Ampère's original law was valid only for stationary charge densities because it lacked the $\partial \boldsymbol{E}/\partial t$ term of Eq. (1.1d). Maxwell realized that for time-dependent densities Ampère's law in its original form was incompatible with the continuity equation (1.3). As discussed in Section 7.1.1, Maxwell then added the $\partial \boldsymbol{E}/\partial t$ term (which is called the *displacement current*) to Ampère's law in Eq. (1.1d); this brought the full set of equations (1.1) into harmony with Eq. (1.3), and effectively brought together the previously separate subjects of electricity and magnetism. As a result, Eq. (1.1d) is also called the Ampère–Maxwell law. The addition of the displacement current to Eq. (1.1d) has a number of far-reaching implications. (1) We may now speak of the unified subject of *electromagnetism*. (2) The fundamental equations of electromagnetism are now consistent with charge conservation. (3) This modification will lead eventually to the interpretation of electromagnetic waves as light. (4) That Maxwell's equation obey the continuity equation and thus conserve charge will lead to the idea of classical *electromagnetic gauge invariance*, which will underlie a quantum field theory of electromagnetism (*quantum electrodynamics* or QED). (5) Electromagnetic gauge invariance will eventually be generalized to more complex gauge invariance in the weak and strong interactions, resulting in the quantum field theory that we term the *Standard Model* of elementary particle physics.

second law,

$$\frac{d\boldsymbol{p}}{dt} = \boldsymbol{F}, \tag{1.5}$$

allows calculating the complete motion of charges in the electromagnetic field.

Then the understanding of classical electromagnetic theory could be summarized in a succinct admonition:

> *Solve Maxwell's equations with appropriate boundary conditions for the electric and magnetic fields, and use those to compute observables for the problem at hand.*

While this admonition is not wrong, it is in danger of leaving two things at loose ends.

1. The solution of Maxwell's equations with appropriate boundary conditions can be highly non-trivial, often requiring considerable mathematical and computational prowess.
2. Although classical electromagnetism itself has changed little since the late 1800s, the context in which we view classical electromagnetism has changed dramatically because of modern advances in quantum field theory (often inspired and guided by the theoretical understanding of classical electrodynamics).

Box 1.2 Symmetries of the Maxwell Equations

Symmetry plays a fundamental role in physics [12]. The Maxwell equations (1.1) exhibit some symmetries that we expect to be valid for physical systems.

1. *Invariance under Space and Time Translations:* Invariance under translations in the spatial and time axes is associated with conservation of momentum and energy. That the Maxwell equations satisfy this invariance automatically implies that we can *assign energy and momentum to the electromagnetic field*.
2. *Invariance under Rotations:* The formulation of the Maxwell equations as vector equations ensures rotational invariance (isotropy of space), which in turn implies conservation of angular momentum by the electromagnetic field.

> The association of invariance under space translations, time translations, and rotations with conservation of linear momentum, energy, and angular momentum, respectively, is a general consequence of **Noether's theorem:** *For every continuous symmetry of a field theory Lagrangian there is a corresponding conserved quantity.* See Section 16.2 of Ref. [12].

3. *Symmetry under Space Inversion (Parity) P:* In the presence of rotational invariance, parity is equivalent to mirror reflection. Under inversion $\boldsymbol{E} \to -\boldsymbol{E}$, which is the transformation law for a *polar vector* (normal 3-vector), but under inversion $\boldsymbol{B} \to \boldsymbol{B}$, which is the transformation law for a *pseudovector* or *axial vector*.
4. *Symmetry under Time Reversal (T):* A motion picture of electromagnetic events would look the same if run backwards or forward.
5. *Lorentz Invariance:* Electromagnetism differs fundamentally from Newtonian particle mechanics in that it is not galilean invariant but is Lorentz invariant (consistent with special relativity), while Newtonian mechanics is galilean invariant but not Lorentz invariant. Einstein was strongly influenced by this property of electromagnetism in formulating the special theory of relativity.
6. *Gauge Invariance (Charge Conservation):* That the Maxwell equations are consistent with the continuity equation (1.3) implies that *charge is conserved locally*. This conservation law is associated with *local gauge invariance* for the electromagnetic field. We shall have a great deal to say about the local gauge symmetry of electromagnetism and its far-reaching implications in later chapters.
7. *Electric and Magnetic Field Asymmetry:* The Maxwell equations in CGS units [Eqs. (B.3) of Appendix B] exhibit a large asymmetry between the roles of electric and magnetic fields. If the zero on the right side of Eq. (B.3c) were replaced by a magnetic charge $4\pi\rho_{\mathrm{m}}$ and a magnetic current term $(4\pi/c)\boldsymbol{J}_{\mathrm{m}}$ were added to the right side of Eq. (B.3b), the Maxwell equations would become highly symmetric under a transformation $\boldsymbol{E} \to \boldsymbol{B}$ and $\boldsymbol{B} \to -\boldsymbol{E}$. However, no magnetic charges (no magnetic monopoles) have ever been observed.

The following chapters of these notes will address the first point at a practical level, by providing a variety of mathematical and computational tools to facilitate solution of Maxwell's equations in various physically important contexts. The remainder of the current chapter will address the second point at a more philosophical level, by setting electromagnetism in the context of modern theoretical physics.

1.2 Electromagnetism in Modern Physics

One of the remarkable findings of modern physics is that the already beautiful edifice of Maxwell's equations for electromagnetism that we shall elaborate on hides within it a deceptively powerful symmetry called *(local) gauge invariance*—in essence a symmetry under local phase transformations—that, with suitable exposition, explains the origin of both classical and quantum theories of electromagnetism. But that is not all! The local gauge symmetry describing electromagnetism can be generalized mathematically into a more powerful theory that can *partially unify* the electromagnetic and weak nuclear forces into a single *electroweak interaction*, and this can be generalized into the *Standard Model* (of elementary particle physics) that partially unifies the electromagnetic, weak, and strong interactions in a single non-abelian gauge theory.[4]

1.2.1 Quantization of Electrical Charge

The basic unit of electrical charge is given by the magnitude of the charge on an electron, which is measured to be

$$|q_\mathrm{e}| = 1.60217733 \times 10^{-19}\,\mathrm{C} \qquad \text{(in SI or MKS units)}. \tag{1.6}$$

The charges on protons, and all presently known particles or systems of particles, are found to be *integral multiples of this unit* (with positive or negative signs for non-zero charges). It is known experimentally that the ratios of charges between different particles are integers to one part in 10^{20}. Indeed, the stability of the atomic matter all around us would be compromised by even a tiny difference in the absolute values of the electron and proton charges. There is presently no convincing explanation for this quantization of charge.[5]

[4] In quantum filed theory local gauge symmetries are generated by a set of quantum operators. If the quantum operators commute among themselves, the gauge symmetry is said to be *abelian*. If they do not all commute among themselves, the gauge symmetry is said to be *non-abelian*. Because of the constraints arising from non-zero commutators among operators, non-abelian gauge symmetries have the potential to engender more complex behavior than abelian gauge symmetries. Electromagnetism corresponds to an abelian local gauge symmetry (with the quantum version known as *quantum electrodynamics or QED*). The Standard Electroweak Model and its generalization to the Standard Model incorporating the strong interactions (*quantum chromodynamics or QCD*) generally correspond to more complex non-abelian local gauge symmetries. Non-abelian models of elementary particles are also known as *Yang–Mills field theories*.

[5] Dirac proposed long ago that, if magnetic monopoles existed, there could be a topological reason for charge quantization. However, no reproducible observations of magnetic monopoles have ever been reported, so

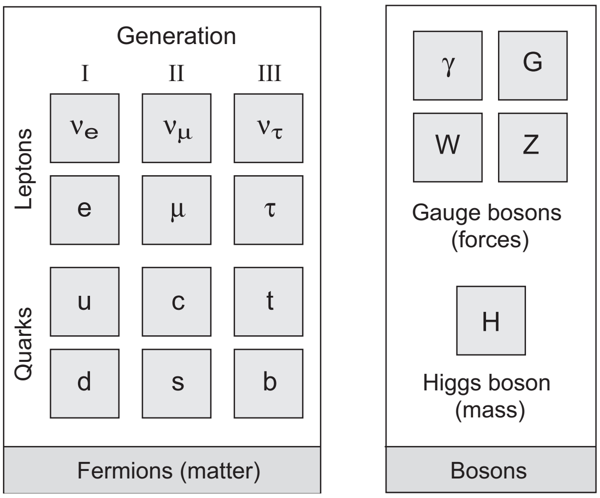


Fig. 1.1

Elementary particles of the Standard Model. Photons are labeled by γ and gluons by G. Elementary particles of half-integer spin that don't undergo strong interactions are called *leptons;* electrons and electron neutrinos are examples. Particles made from quarks, antiquarks, and gluons (and thus that undergo strong interactions) are called *hadrons*; pions (pi mesons) and protons are examples. A subset of hadrons corresponding to more massive particles containing three quarks are called *baryons*; protons and neutrons are examples. The different types of neutrinos $(\nu_e, \nu_\mu, \ldots)$ and the different types of quarks (u, d, s, ...) are called *flavors.* For simplicity the largely parallel classification of antiparticles has been omitted.

1.2.2 Particles in the Universe

Our modern view is that matter in the Universe consists of the fermions in the Standard Model (of elementary particle physics). The elementary particles of the Standard Model are summarized in Fig. 1.1. In the Standard Model matter is formed from fermions, while interactions between elementary particles are mediated by the exchange of gauge bosons, and masses for bare (that is, non-interacting) particles arise from interactions with the Higgs boson. For reasons that are not yet understood, the fermions (matter fields) of the Standard Model may be divided into three generations (or families), I, II, and III, with the fermions of each generation being successively more massive than in the preceding generation, and with many properties repeating themselves in successive generations. (For example, the muon μ in generation II acts in many respects as if it were a heavier version of the electron e in generation I.) Table 1.1 gives Standard Model quark quantum number assignments for the six known quark flavors displayed in Fig. 1.1. As indicated in Table 1.1,

Dirac's idea remains only conjecture. As we have noted in Box 1.2, if magnetic monopoles did exist, the Maxwell equations (1.1) would become much more symmetric with respect to electric and magnetic fields.

Table 1.1 Quantum number assignments for quarks [9]

	Up	Down	Strange	Charm	Bottom	Top
Symbol	u	d	s	c	b	t
Baryon number (B)	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
Spin	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
Charge (Q)	$\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$
Isospin (T)	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0
Projection of isospin (T_3)	$\frac{1}{2}$	$-\frac{1}{2}$	0	0	0	0
Strangeness number (S)	0	0	-1	0	0	0
Charm number (c)	0	0	0	1	0	0
Bottom number (b)	0	0	0	0	-1	0
Top number (t)	0	0	0	0	0	1

The additive quantum numbers Q, T_3, S, c, B, b, and t of the corresponding antiquarks are the negative of those for quarks. The charge is given by $Q = T_3 + \frac{1}{2}(B+S+c+b+t)$.

the (electrical) charge Q of each quark is given by

$$Q = T_3 + \frac{1}{2}(B+S+c+b+t), \tag{1.7}$$

which leads to third-integer charges for all the quarks (row 4 of Table 1.1).

Example 1.1 From Eq. (1.7) and Table 1.1, the charge of the down quark d is

$$\begin{aligned} Q_d &= T_3 + \frac{1}{2}(B+S+c+b+t) \\ &= -\frac{1}{2} + \frac{1}{2}\left(\frac{1}{3}+0+0+0+0\right) \\ &= -\frac{1}{3} \end{aligned}$$

This quantization of charge in integer multiples of $\pm\frac{1}{3}$ of the charge of the electron is characteristic of quarks, as may be seen from Table 1.1.

However, the non-abelian gauge field theory of the strong interactions (quantum chromodynamics) predicts that *quarks are confined and can never appear as free particles,* and indeed no free particles with 3rd-integer charges have ever been observed in experiments. For example, a proton has a *udu* quark structure (two up quarks and one down quark) and from Table 1.1 the charge for a proton is

$$Q_p = 2Q_u + Q_d = 2\left(\frac{2}{3}\right) + \frac{1}{3} = +1,$$

in terms of the fundamental charge unit $|q_\mathrm{e}|$ given by Eq. (1.6), while a neutron has a *udd* quark structure and

$$Q_n = Q_u + 2Q_d = \frac{2}{3} + 2\left(-\frac{1}{3}\right) = 0,$$

and neutrons have zero net electrical charge.

Thus, even though modern high energy physics has found a variety of elementary particles, some of which have electrical charges that are fractions of the fundamental unit (1.6) set by the charge on the electron, the *physically observable particles* entering into classical electromagnetism all appear to have electrical charges that are integer multiples of the fundamental charge unit given by Eq. (1.6).

Background and Further Reading

The history of classical electromagnetism is summarized in various parts of Jackson [15]. A view of classical electromagnetism with emphasis on the relationship of modern quantum field theory to classical electromagnetism may be found in Wald [25]. Introductions to the Standard Model of elementary particle physics and the origin of the particle and quantum number assignments in Fig. 1.1 and Table 1.1 may be found in many places; for example, in Refs. [9, 12].

Problems

1.1 Prove that Maxwell's equations (1.1) are consistent with the continuity equation (1.3).

2 Electrostatics in Vacuum

The fundamental problem to be solved in electrodynamics is illustrated schematically in Fig. 2.1(a). If we have a distribution of n distinct *source charges* q_1, q_2, q_3, $\ldots q_n$, what net force do they exert on a *test charge* Q? In the most general case both the source charges and test charges may be in motion but we shall begin with the simpler case of *electrostatics,* where the source charges are assumed to be fixed in spatial position (but the test charge may move). We shall also assume initially that the source and test charges are embedded in vacuum or a medium of negligible susceptibility, so that we do not initially need to worry about the effects of a medium altering the interactions between charges (that will be the subject of later chapters). Let us now consider a quantitative description of this idealized problem, utilizing experimental data, mathematics, and physical intuition as our guides.

2.1 Coulomb's Law

Consider first the force acting between two isolated charges at rest with respect to each other, as illustrated in Fig. 2.1(b). The force exerted on a single charge q_1 located at position $\boldsymbol{x}_1$ by a single charge q_2 located at position $\boldsymbol{x}_2$ may be measured experimentally and is found to be given by *Coulomb's law*,

$$\boldsymbol{F} = kq_1q_2\,\frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{|\boldsymbol{x}_1 - \boldsymbol{x}_2|^3}, \tag{2.1}$$

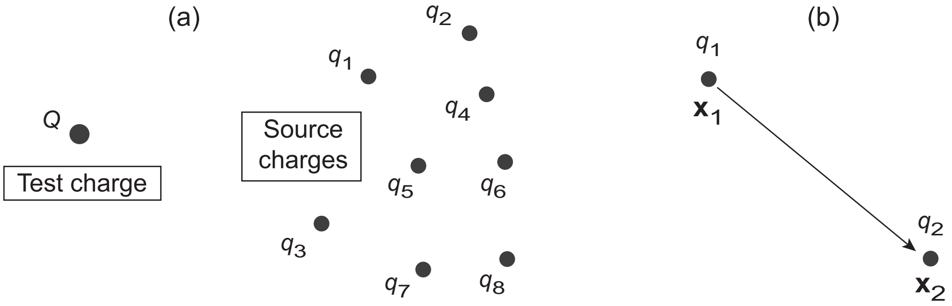


Fig. 2.1 (a) The prototype electrostatics problem: interaction in vacuum of a test charge Q with a set of stationary source charges q_n. (b) Coulomb interaction between two isolated and stationary test charges.

where the charges q_n are algebraic quantities that may be positive or negative, the force points along the line from q_1 to q_2 and is attractive if the signs of the changes are opposite and repulsive if they are the same. The constant k depends on the system of units that is in use. Two common ones are *electrostatic units* and *the SI system*.

1. In *electrostatic units* (*esu*; also termed *Gaussian* or *CGS*), $k = 1$ and unit charge is chosen such that it exerts a force of one dyne on an equivalent point charge located one centimeter away. In the esu system the unit charge is called a *statcoulomb*.
2. In the *SI system* of units,

$$k = \frac{1}{4\pi\varepsilon_0}, \tag{2.2}$$

where the constant ε_0 is called the *permittivity of free space.* In the SI system the unit of force is the Newton (N), the unit of distance is the meter (m), the unit of charge is the coulomb (C), and

$$\varepsilon_0 \simeq 8.85 \times 10^{-12} \, \frac{\mathrm{C}^2}{\mathrm{N\,m}^2} = 8.85 \times 10^{-12} \, \frac{\mathrm{F}}{\mathrm{m}}, \tag{2.3}$$

where the farad (F) is the derived SI unit of electrical capacitance.[1] The constants μ_0, $\varepsilon_)$, and the speed of light c that may appear in SI units are related by

$$\mu_0\varepsilon_0 = \frac{1}{c^2}. \tag{2.4}$$

Thus Coulomb's law (2.1) expressed in SI units is

$$\boldsymbol{F} = \frac{q_1 q_2}{4\pi\varepsilon_0} \, \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{|\boldsymbol{x}_1 - \boldsymbol{x}_2|^3}. \tag{2.5}$$

Other systems of units such as *Heaviside–Lorentz* (where $k = 1/4\pi$) are favored in certain areas of physics, but we shall primarily use the SI system of units.

The correctness of Eq. (2.5) has been explored in some depth by a variety of techniques and over a range of distance scales. As discussed in Box 2.1, experimental limits suggest that for classical electromagnetism we can assume the photon to be massless and the deviation from Coulomb's law to be negligible on all length scales investigated thus far.

2.2 The Electric Field

In experiments determining the interaction of charges one typically measures a force, as implied by Eq. (2.1). However, much of the power of theoretical physics derives from the ability to abstract broader implications from direct measurements. Perhaps no abstract concept has been more powerful in the development of physics than that of a *field*, which is

[1] The three constants that play a role in SI units, the permeability of free space μ_0 defined in Eq. (1.2), the permittivity of free space ε_0 defined in Eq. (2.3), and the speed of light c, are related by $\mu_0\varepsilon_0 c^2 = 1$. This has the unfortunate implication that the form of equations written in SI units can be changed by using $\mu_0\varepsilon_0 c^2 = 1$ to interchange constants appearing in them.

Box 2.1 Deviations from the Inverse-Square Force Law

As established in the original experiments of Cavendish and Coulomb, the electrostatic force between two charges in vacuum obeys an inverse square law (2.1). Since the time of Cavendish and Coulomb, the precision of testing for possible deviations from the inverse square law has improved substantially. It is common to report possible experimental deviation from the inverse square law in one of two ways.

1. Assume that the electrostatic force has the dependence $F \sim 1/r^{2+\varepsilon}$ and report an upper experimental limit on ε.
2. Assume that the electrostatic potential has the Yukawa form

$$V \sim r^{-1}e^{-\mu r} = r^{-1}e^{-(m_\gamma c/\hbar)r} \qquad \mu \equiv \frac{m_\gamma c}{\hbar},$$

where m_γ is the mass of the photon (which should be identically massless for an inverse square force law). Then possible deviations from the inverse square law are often reported as an upper limit on μ or an upper limit on m_γ.

The original experiments of Cavendish using concentric spheres in 1772 established an upper limit $|\varepsilon| \leq 0.02$.

1. Greatly improved modern determinations based on Gauss's law have pushed this limit to $\varepsilon = (2.7 \pm 3.1) \times 10^{-16}$ [26].
2. The best limits on the mass of the photon come from measuring planetary magnetic fields. Such measurements place a limit on the photon mass of $m_\gamma < 4 \times 10^{-51}$ kg [7, 15, 18].

Hence the photon mass can be assumed to be zero for the entire regime of classical electrodynamics, implying no significant deviation from the inverse square law.

simply an instance of something that is defined at every point of spacetime. Let us introduce the concept of an *electric field* $\boldsymbol{E}$ acting on a test charge q by defining

$$\boldsymbol{F} = q\boldsymbol{E}. \tag{2.6}$$

Thus the electric field may be interpreted as the force per unit test charge at a particular point in spacetime. Since force is a vector and charge is a scalar, the electric field generated by a charge is a *vector field* $\boldsymbol{E}(t,\boldsymbol{x})$ defined at each point of spacetime $(t,\boldsymbol{x})$.[2] Comparing Eqs. (2.6) and (2.1), the electric field at a point $P(\boldsymbol{x})$ produced by a point charge q_1 at $\boldsymbol{x}_1$ is

$$\boldsymbol{E}(\boldsymbol{x}) = kq_1 \frac{\boldsymbol{x}-\boldsymbol{x}_1}{|\boldsymbol{x}-\boldsymbol{x}_1|^3}, \tag{2.7}$$

as illustrated in Fig. 2.2. The electric field is a function of position, but is independent of

[2] In the present discussion of electrostatics we ignore time dependence so the electric field is assumed to be defined at every point $\boldsymbol{x}$ of the spatial manifold, with no dependence on time.

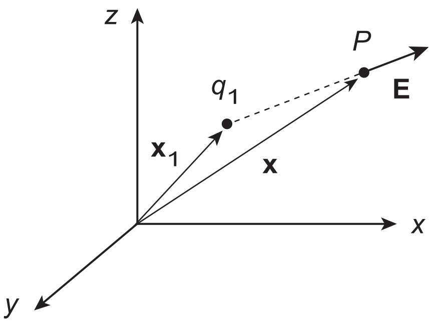


Fig. 2.2 The electric field vector $\boldsymbol{E}$ at a point $P(\boldsymbol{x})$, generated by a charge q_1 located at position $\boldsymbol{x}_1$.

whether there is a test charge located at P. (The field is generated by the source charges.) The SI unit of charge is the coulomb (C) and the electric field $\boldsymbol{E}$ has units of volts per meter in the SI system. The importance of the field concept in the development of modern physics is elaborated in Box 2.2.

2.3 Principle of Superposition

Now let us address the more general case in Fig. 2.1(a) of the interaction of a test charge with a set of n source charges. In principle this could be a quite complicated problem, but it is an experimentally observed fact that as long as the interactions between charged particles are not too large, and as long as we can ignore quantum effects at the microscopic level, the interaction between any two charges is unaffected by the presence of all other charges.[3]

> Thus, the total force acting on the test charge Q in Fig. 2.1(a) can be obtained to excellent approximation by summing the interactions of the charges pairwise, while holding all other charges constant. This is termed the *principle of linear superposition (of forces).* The ultimate source of this linearity may be viewed as the linearity of the Maxwell equations (1.1) in the electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{B}$.

Since the principle of linear superposition applies to the forces computed from Eq. (2.1)

[3] In the presence of strong electric fields such as those generated by powerful lasers, and in various *vacuum polarization phenomena* observed in modern quantum field theory experiments, this linear superposition principle may fail. However, we shall restrict ourselves explicitly in these notes to the *classical linear regime*, where such effects may be neglected by hypothesis. It is fortunate for the development of the theory that the formative experiments that laid the foundations of the classical theory of electricity and magnetism were all performed under conditions where linear superposition holds very precisely, which greatly expedited their original physical interpretation.

Box 2.2 Significance of Fields in Electromagnetism

As suggested by Eq. (2.6), we generally measure forces and fields may be inferred from that, suggesting a derivative role for fields. But modern physics places strong emphasis directly on the fields. Indeed, Maxwell's equations (1.1) are formulated in terms of electric and magnetic fields, and the electromagnetic field is the central concept of the theory of electromagnetism. The importance of fields is most obvious in relativistic quantum field theory, which is not our subject here. However, the field concept traces historically to the introduction of electric and magnetic fields in the description of classical electromagnetism, which is our subject.

Action at a Distance

Originally it was believed that forces associated with charges, currents, and magnets constituted *action at a distance,* meaning that the forces acted *instantaneously over any distance.*[a] The modern view—shaped by experimental measurement and the development of quantum field theory—is that fields are every bit as fundamental as particles (arguably even more so). In this picture, forces are mediated by fields, and the lightspeed limit set by special relativity means that no signal can transmit a force faster than the speed of light, relegating action at a distance to the dustbin.

Actions Mediated by Fields

If, for example, a charge moves, the fields created by the charge change, but that change isn't felt immediately at every point $(\boldsymbol{x},t)$ of spacetime. Maxwell's equations (1.1) describing electromagnetism require the change to be propagated through changes in two vector fields, the *electric field* $\boldsymbol{E}(\boldsymbol{x},t)$ and the *magnetic field* $\boldsymbol{B}(\boldsymbol{x},t)$, defined at each point of spacetime, and that those changes can propagate only at a finite speed (less than or equal to the speed of light).

This abstract view was initially resisted by many, particularly because the fields could exist in vacuum, and did not describe tangible matter. Modern physics is of quite a different mind. Yes, the introduction of electric and magnetic fields allows a simple and clean mathematical description of electromagnetic phenomena, but the fields are *not just mathematical abstractions;* they are *real physical entities* ("as real as a rinoceros" [6]). They carry concrete physical properties such as energy, momentum, and angular momentum. This is most clear in quantum field theory, where these physical attributes become properties of *quanta of the field* (photons) and one views electromagnetic interactions as being mediated by exchange of virtual photons. These photons associated with the electromagnetic field have observable consequences: at sufficiently high energy the collision of two photons can produce matter in the form of an electron and positron pair; real as a rinoceros, indeed! Quantum field theory implies that all non-gravitational fundamental interactions (electromagnetic, strong, and weak) are mediated by the gauge bosons of Fig. 1.1 that are the quanta of gauge fields generalizing the gauge symmetry of photons. Such is the rich legacy of introducing the concept of fields in classical electromagnetism.

[a] Similar statements apply to gravity, since Newton's gravitational law implicitly assumes the force to be applied instantly to a distant object. Replacement of Newtonian gravity with general relativity (a field theory consistent with special relativity and the limiting speed of light) eliminated action at a distance in gravitational physics.

it applies also to the electric fields calculated from Eq. (2.7), by virtue of the linear relationship (2.6). Therefore, the electric field acting at position $\boldsymbol{x}$ in Fig. 2.1(a) is given by a vector sum of contributions from all the source charges q_i,

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{n} q_i \frac{\boldsymbol{x}-\boldsymbol{x}_i}{|\boldsymbol{x}-\boldsymbol{x}_i|^3}, \tag{2.8}$$

where now to be definite we shall work in the SI system of units. In most practical problems we will be able to assume that the charges are sufficiently small and numerous that they can be approximated by a continuous charge density $\rho(\boldsymbol{x}')$, such that the charge contained in a small 3D volume element $d^3x' \equiv dx'dy'dz'$ centered at $\boldsymbol{x}'$ is $\Delta q = \rho(\boldsymbol{x}')\Delta x\Delta y\Delta z$. Then the sum over discrete charges in Eq. (2.8) may be replaced by an integral over a continuous charge distribution,

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \rho(\boldsymbol{x}') \frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\, d^3x'. \tag{2.9}$$

A discrete set of charges also can be described by a continuous charge distribution $\rho(\boldsymbol{x})$ if the *Dirac delta function* described in Appendix A.9 is used to pick out the locations of each of the discrete charges,

$$\rho(\boldsymbol{x}) = \sum_{i=1}^{n} q_1 \delta(\boldsymbol{x}-\boldsymbol{X}), \tag{2.10}$$

since substitution of the charge density (2.10) in Eq. (2.9) and integrating using the properties of the delta function described in Appendix A.9 gives the sum over contributions from discrete charges to the electric field in Eq. (2.8).

2.4 Gauss's Law

The electric field for a continuous charge distribution may be calculated from Eq. (2.9). However, evaluating the integral in this equation isn't always the easiest solution for the electric field. If a problem has some level of symmetry, often another integral result called *Gauss's law* can lead to an easier solution. In Fig. 2.3 we indicate a charge q enclosed by a surface S. The normal component of $\boldsymbol{E}$ times a surface area element da is

$$\boldsymbol{E}\cdot\boldsymbol{n}\,da = \frac{q}{4\pi\varepsilon_0}\frac{\cos\theta}{r^2}\,da = \frac{q}{4\pi\varepsilon_0}\,d\Omega, \tag{2.11}$$

where in the last step we have used that $\cos\theta\, da = r^2\, d\Omega$, with $d\Omega$ being the solid angle subtended by da at the position of the charge. If the normal component of $\boldsymbol{E}$ is now integrated over the entire surface,

$$\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da = \begin{cases} q/\varepsilon_0 & \text{for } q \text{ inside } S, \\ 0 & \text{for } q \text{ outside } S. \end{cases} \tag{2.12}$$

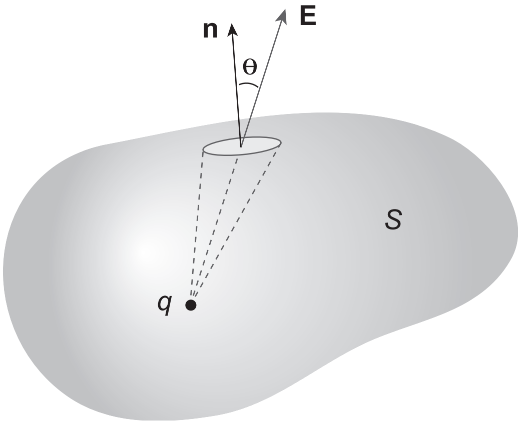


Fig. 2.3 Closed surface S illustrating Gauss's law. The electric field generated by the charge q is $\boldsymbol{E}$, the normal vector to the surface is $\boldsymbol{n}$, and da is an element of surface area.

where $\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da$ is called the *flux* through the surface S. Equation (2.12) is *Gauss's law in integral form* for a single charge. For a set of discrete charges Gauss's law takes the form,

$$\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da = \frac{1}{\varepsilon_0}\sum_i q_i, \tag{2.13}$$

where the summation over i is restricted to charges q_i that are *inside the surface S*. If the charge distribution is continuous, Gauss's law takes the form

$$\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da = \frac{1}{\varepsilon_0}\int_V \rho(\boldsymbol{x})\,d^3x, \tag{2.14}$$

where the integration is over the volume V contained within the surface S.[4]

Gauss's law in Eqs. (2.12)-(2.14) may be viewed as integral formulations of the law of electrostatics. Corresponding differential forms of Gauss's law may be obtained using the *divergence theorem* of Eq. (A.33).

Divergence theorem: For a vector field defined within a volume V that is enclosed by a surface S,

$$\oint_S \boldsymbol{A}\cdot\boldsymbol{n}\,da = \int_V \boldsymbol{\nabla}\cdot\boldsymbol{A}\,d^3x, \tag{2.15}$$

where the left side is the surface integral of the outwardly directed normal component of the vector $\boldsymbol{A}$ and the right side is the volume integral of the divergence of $\boldsymbol{A}$.

[4] Gauss's law (2.14) is a consequence of (1) the forces between charges being of inverse square form, (2) the central nature of the force, and (3) linear superposition of charges. Newtonian gravity obeys similar conditions, so one can construct a Gauss's law for Newtonian gravity, where the gravitational "charge" is mass and mass density replaces charge density. Of course there is only one sign for the gravitational charge in Newtonian gravity, since the gravitational force is always attractive. (In the absence of dark energy, which can effectively turn gravity repulsive on cosmological scales.)

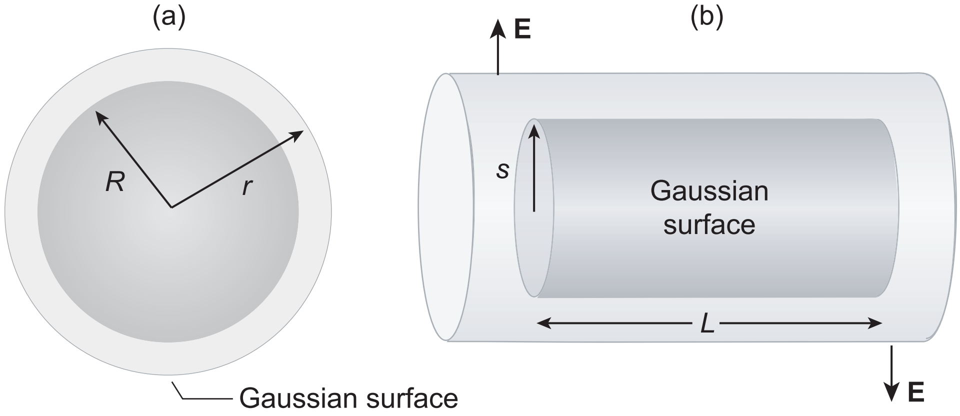


Fig. 2.4 Examples of applying Gauss's law to find the electric field. (a) A ball of uniformly distributed charge with radius R, used in Example 2.1. (b) A cylinder carrying a charge density proportional to the distance s from the cylindrical axis, used in Example 2.2.

The divergence theorem (2.15) allows Eq. (2.14) to be written in the form

$$\int_V \left(\boldsymbol{\nabla} \cdot \boldsymbol{E} - \frac{\rho}{\varepsilon_0} \right) d^3x = 0.$$

But this can be true generally only if the integrand vanishes, giving

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)}, \tag{2.16}$$

which is the *differential form of Gauss's law* (2.14) for continuous charge distributions. Thus, we have now obtained the first of Maxwell's equations (1.1). Before proceeding, let's look at two examples of Gauss's law in action [8].

Example 2.1 Figure 2.4(a) shows a charged solid 2-sphere (a ball)[5] of radius R, for which the total charge Q is assumed to be evenly distributed. What is the electric field outside the ball? We may solve this easily using Gauss's law in integral form. Imagine surrounding the charged ball with a 2-sphere of radius $r > R$ (this is called a *Gaussian surface*). Applying Gauss's law (2.13)

$$\oint \boldsymbol{E} \cdot \boldsymbol{n}\, da = \frac{1}{\varepsilon_0} Q.$$

Because of the spherical symmetry, $\boldsymbol{E}$ and $\boldsymbol{n}\, da$ point radially outward so that the scalar product is trivial to evaluate:

$$\oint \boldsymbol{E} \cdot \boldsymbol{n}\, da = \oint |\boldsymbol{E}| da,$$

[5] A 2-sphere is hollow when displayed in 3D space, consisting only of the 2D spherical surface. Mathematically, a "solid 2-sphere", which includes the interior of a 2-sphere, is called a *ball* (an *open ball* if the surface points are not included and a *closed ball* if they are).

and the magnitude $|\boldsymbol{E}|$ is constant over the surface by symmetry, so it can be pulled out of the integral,

$$\oint |\boldsymbol{E}| da = |\boldsymbol{E}| \oint da = 4\pi r^2 |\boldsymbol{E}|.$$

Combining the preceding results,

$$4\pi r^2 |\boldsymbol{E}| = \frac{1}{\varepsilon_0} Q,$$

or finally,

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0} \frac{Q}{r^2} \hat{\boldsymbol{r}},$$

where $\hat{\boldsymbol{r}}$ is a unit vector in the radial direction. Notice the well-known result that the field external to the charge distribution is the same that would have been obtained by putting all charge at the center of the ball.[6]

Example 2.2 Consider Fig. 2.4(b), where a long cylinder carries a charge density proportional to the distance s' from the axis of the cylinder, $\rho = ks'$, where k is a constant. What is the electric field inside the cylinder? Let's draw a Gaussian cylinder of radius s and length L, as illustrated in Fig. 2.4(b). From Eq. (2.14), Gauss's law for this surface is

$$\oint \boldsymbol{E} \cdot \boldsymbol{n} da = \frac{1}{\varepsilon_0} Q,$$

where Q is the total charge enclosed by the Gaussian surface, which is given by integrating the charge over the volume within the Gaussian surface using cylindrical coordinates (Appendix A.7.2) with a cylindrical volume element $d\tau = s ds d\phi dz$

$$\begin{aligned} Q = \int \rho d\tau &= k \int_0^s s'^2 ds' \int_0^{2\pi} d\phi \int_0^L dz \\ &= 2\pi kL \int_0^s s'^2 ds' = \frac{2}{3}\pi k L s^3. \end{aligned}$$

By symmetry $\boldsymbol{E}$ must point radially outward from the cylinder's central axis, as indicated in Fig. 2.4(b), so for the curved portion of the Gaussian cylinder

$$\int \boldsymbol{E} \cdot \boldsymbol{n} da = \int |\boldsymbol{E}| da = |\boldsymbol{E}| \int da = 2\pi s L |\boldsymbol{E}|,$$

while the two ends contribute zero because $\boldsymbol{E}$ is perpendicular to $\boldsymbol{n} da$. Thus, from Gauss's law,

$$2\pi s L |\boldsymbol{E}| = \frac{1}{\varepsilon_0} \frac{2}{3} \pi k L s^3$$

and the electric field is given by

$$\boldsymbol{E} = \frac{1}{3\varepsilon_0} k s^2 \hat{\boldsymbol{s}},$$

[6] By similar arguments applied to Newtonian gravity, one finds that for a sphere containing gravitating matter the gravitational field external to the sphere is the same as if all mass were concentrated at the center of the sphere.

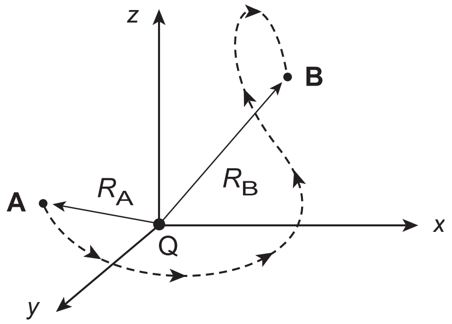


Fig. 2.5 Path for a line integral in an electric field generated by a charge Q at the origin.

where $\hat{\boldsymbol{s}}$ is a unit vector pointing radially from the central axis of the cylinder.

One sees generally from such examples that using Gauss's theorem to determine the electric field is most useful when there is a high degree of symmetry that can be exploited to evaluate integrals. Generally our goal with Gauss's law is to simplify the integral on the left side of Eqs. (2.13) or (2.14). Often this can be done if we can choose a gaussian surface such that one or more of the following conditions holds.

1. The electric field is zero over the surface.
2. The electric field is constant over the surface, by symmetry arguments.
3. The integrand $\boldsymbol{E} \cdot \boldsymbol{n}\, da$ reduces to the algebraic product $|\boldsymbol{E}|\, da$ because the vectors $\boldsymbol{E}$ and $\boldsymbol{n}$ are parallel.
4. The integrand $\boldsymbol{E} \cdot \boldsymbol{n}\, da$ is zero because the vectors $\boldsymbol{E}$ and $\boldsymbol{n}$ are orthogonal.

For example, we used the third condition in Example 2.1 and the third and fourth conditions in Example 2.2 above. If none of these possibilities are fulfilled, Gauss's law is still valid but it may not be the easiest way to determine the electric field.

2.5 The Scalar Potential

Consider a line integral between two points A and B in a field generated by a single charge at the origin, as illustrated in Fig. 2.5. In spherical coordinates (Appendix A.7.1) the electric field $\boldsymbol{E}$ and line element $d\boldsymbol{l}$ are

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\boldsymbol{r}} \qquad d\boldsymbol{l} = dr\,\hat{\boldsymbol{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\phi\,\hat{\boldsymbol{\phi}}, \tag{2.17}$$

where $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$ are unit vectors. Therefore, the line integral is

$$\begin{aligned}\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l} &= \frac{1}{4\pi\varepsilon_0}\int_{\boldsymbol{A}}^{\boldsymbol{B}} \frac{Q}{r^2}\,dr \\ &= \left.\frac{-Q}{4\pi\varepsilon_0 r}\right|_{R_A}^{R_B} \\ &= \frac{1}{4\pi\varepsilon_0}\left(\frac{Q}{R_A}-\frac{Q}{R_B}\right).\end{aligned} \tag{2.18}$$

The integral around a *closed path* is then zero,

$$\oint \boldsymbol{E}\cdot d\boldsymbol{l} = 0, \tag{2.19}$$

since $R_a = R_b$ in that case. Now we may invoke *Stokes' theorem* [Eq. (A.29)]:

Stokes' theorem: If $\boldsymbol{A}$ is a vector field and S is an arbitrary open surface bounded by a closed curve C, then

$$\int_S (\boldsymbol{\nabla}\times\boldsymbol{A})\cdot\boldsymbol{n}\,da = \oint_C \boldsymbol{A}\cdot d\boldsymbol{l}, \tag{2.20}$$

where $\boldsymbol{n}$ is the normal to S, the line element on the curve C is $d\boldsymbol{l}$, and the path in the line integration is traversed in a right-hand screw sense relative to $\boldsymbol{n}$. A geometrical interpretation of Stokes' theorem is given in Box 2.3.

This implies that

$$\int_S (\boldsymbol{\nabla}\times\boldsymbol{E})\cdot\boldsymbol{n}\,da = \oint_P \boldsymbol{E}\cdot d\boldsymbol{l} = 0, \tag{2.21}$$

and since this must be valid for any closed path, the integrand on the left side must vanish,

$$\boldsymbol{\nabla}\times\boldsymbol{E} = \boldsymbol{0}. \tag{2.22}$$

Thus, we have shown that *the curl of an electric field* $\boldsymbol{E}$ *is zero*.

Because of Eq. (2.22) and Stokes' theorem, the line integral of the electric field around any closed loop is zero, which implies that the line integral between points $\boldsymbol{A}$ and $\boldsymbol{B}$ in Fig. 2.5 has *the same value for all possible paths*. Thus we can define a function $\Phi(\boldsymbol{x})$ by

$$\Phi(\boldsymbol{x}) = -\int_{\mathscr{O}}^{\boldsymbol{x}} \boldsymbol{E}(\boldsymbol{x})\cdot d\boldsymbol{l}, \tag{2.23}$$

where $\mathscr{O}$ is a chosen standard reference point. The function $\Phi(\boldsymbol{x})$ is called *scalar potential* or the *electric potential*. The scalar potential associated with a point charge q at the origin is given by

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{r}\right) \tag{2.24}$$

Box 2.3 Geometrical Interpretation of Stokes' Theorem [12]

For a 3D vector field $\boldsymbol{A}$, *Stokes' theorem* relates a surface integral of the curl vector field $\boldsymbol{\nabla} \times \boldsymbol{A}$ to a line integral around the (smooth) boundary of that surface:

$$\oint_C \boldsymbol{A} \cdot d\boldsymbol{r} = \int_S (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot \boldsymbol{n}\, ds \equiv \int_S (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot d\boldsymbol{s} \qquad \text{(Stokes' theorem)},$$

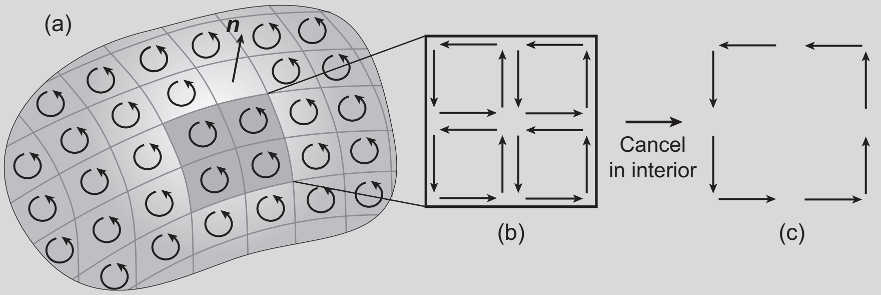
where S is the 2D surface enclosed by the 1D boundary C, the outward normal to the surface is $\boldsymbol{n}$, and the curl $\boldsymbol{\nabla} \times \boldsymbol{A}$ is given in cartesian coordinates by

$$\boldsymbol{\nabla} \times \boldsymbol{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \right) \hat{\boldsymbol{x}} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \right) \hat{\boldsymbol{y}} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) \hat{\boldsymbol{z}}.$$

The orientation of the surface $\boldsymbol{n}$ and the direction of the integration path around the boundary are related by a *right-hand rule:*

> Point the thumb of your right hand in the direction of a unit normal vector near the edge of the surface and curl your fingers; the direction that your fingers point indicates the integration direction around the boundary.

The physical content of Stokes' theorem may be understood geometrically:



For the darker gray plaquettes in (b) of (a) the circulating currents cancel in the interior, leaving only the contribution from the boundary (c). Generalizing to the whole surface, as plaquette size tends to zero all interior contributions to the surface integral may be expected to cancel, leaving only the boundary contributions.

where r is the separation between charge and point, and invoking superposition the potential generated by a collection of n charges is

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{n} \frac{q_i}{r_i}. \tag{2.25}$$

For a continuous charge distribution the potential evaluates to

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x' \qquad \text{(3D volume charge)}, \tag{2.26a}$$

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da' \qquad \text{(2D surface charge)}, \tag{2.26b}$$

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\lambda(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,dl' \qquad \text{(1D line charge)}, \tag{2.26c}$$

where ρ is a volume charge density, σ is a surface charge density, and λ is a line charge density.

2.6 The Electric Field and the Scalar Potential

The electric field is special because it has vanishing curl. We shall now use this to reduce finding the electric field (a vector problem) to a simpler scalar problem. The *potential difference* between points $\boldsymbol{A}$ and $\boldsymbol{B}$ (see Fig. 2.5) is given by

$$\begin{aligned}\Phi(\boldsymbol{B}) - \Phi(\boldsymbol{A}) &= -\int_{\mathscr{O}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l} + \int_{\mathscr{O}}^{\boldsymbol{A}} \boldsymbol{E}\cdot d\boldsymbol{l} \\ &= -\int_{\mathscr{O}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l} - \int_{\boldsymbol{A}}^{\mathscr{O}} \boldsymbol{E}\cdot d\boldsymbol{l} \\ &= -\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l}.\end{aligned} \tag{2.27}$$

Applying the fundamental theorem for gradients (A.27),

$$F(\boldsymbol{B}) - F(\boldsymbol{A}) = \int_{\boldsymbol{A}}^{\boldsymbol{B}} (\boldsymbol{\nabla} F)\cdot d\boldsymbol{l}, \tag{2.28}$$

to Eq. (2.27) then gives

$$\int_{\boldsymbol{A}}^{\boldsymbol{B}} (\boldsymbol{\nabla}\Phi)\cdot d\boldsymbol{l} = -\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l}. \tag{2.29}$$

But since Eq. (2.29) must be true for any points $\boldsymbol{A}$ and $\boldsymbol{B}$, the integrands of the integrals on the two sides of Eq. (2.29) must be equal and we obtain

$$\boldsymbol{E} = -\boldsymbol{\nabla}\Phi, \tag{2.30}$$

which is a differential version of Eq. (2.23).

> The electric field vector $\boldsymbol{E}$ may be obtained by taking the (negative) gradient of the scalar potential Φ.

In SI units force is measured in newtons and charge in coulombs, so electric fields $\boldsymbol{E}$ have units of newtons per coulomb. The potential Φ has units then of newton-meters per coulomb, or joules per coulomb, where a joule per coulomb is defined to be a volt.

A big advantage of the potential formulation is that if you know Φ you can obtain the electric field simply by taking the gradient, as in Eq. (2.30). This is perhaps surprising because Φ is a scalar with only *one component*, while $\boldsymbol{E}$ is a vector with *three components*. The source of this seeming miracle is the restrictions placed on the components $\boldsymbol{E}$ by Eq. (2.22), since expansion of $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$ leads to the constraint equations (see Problem 2.5)

$$\frac{\partial E_x}{\partial y} = \frac{\partial E_y}{\partial x} \qquad \frac{\partial E_z}{\partial y} = \frac{\partial E_y}{\partial z} \qquad \frac{\partial E_x}{\partial z} = \frac{\partial E_z}{\partial x}. \tag{2.31}$$

There is an essential ambiguity in defining the potential Φ because of the arbitrary reference point $\mathscr{O}$ appearing in Eq. (2.23), since changing it shifts the potential by a constant amount. Thus the potential itself has no clear physical meaning, but as long as we choose the same reference $\mathscr{O}$ for all potentials,

1. *differences between potentials* $(\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_j)$ are independent of the reference point and
2. *gradients of the potential* $\boldsymbol{\nabla}\Phi$ are unaffected by shifting the reference point a constant amount, since the derivative of a constant is zero,

so these have physical meaning. The selection of a reference point is in principle arbitrary, but the most common choice in electrostatics is to take the reference point for potentials to be an infinite distance away from the charges, where the potential drops to zero.[7]

2.7 Superposition of Scalar Potentials

An important property of the potential is that it obeys the superposition principle. We introduced the principle of linear superposition in Section 2.3 with the hypothesis that the forces acting on a test charge Q are the vector sum of contributions from each source charge q_i,

$$\boldsymbol{F} = \boldsymbol{F}_1 + \boldsymbol{F}_2 + \boldsymbol{F}_3 + \dots \tag{2.32}$$

and since $\boldsymbol{F} = Q\boldsymbol{E}$, dividing the terms in Eq. (2.32) by Q implies linearity for the electric fields $\boldsymbol{E}$ also,

$$\boldsymbol{E} = \boldsymbol{E}_1 + \boldsymbol{E}_2 + \boldsymbol{E}_3 + \dots \tag{2.33}$$

Likewise, from the preceding definitions the scalar potential Φ is expected to obey linear superposition,

$$\Phi = \Phi_1 + \Phi_2 + \Phi_3 + \dots \tag{2.34}$$

meaning that the potential at a point $\boldsymbol{x}$ is the sum of the potentials due to all source charges separately. However, there is a fundamental difference between linear superposition for potentials and linear superposition for forces and electric fields.

[7] This prescription requires more thought if a problem hypothesizes a charge distribution that extends to infinity. In that case a different reference point $\mathscr{O}$ must be used. We won't worry about that here, since real-world charge distributions are typically bounded spatially.

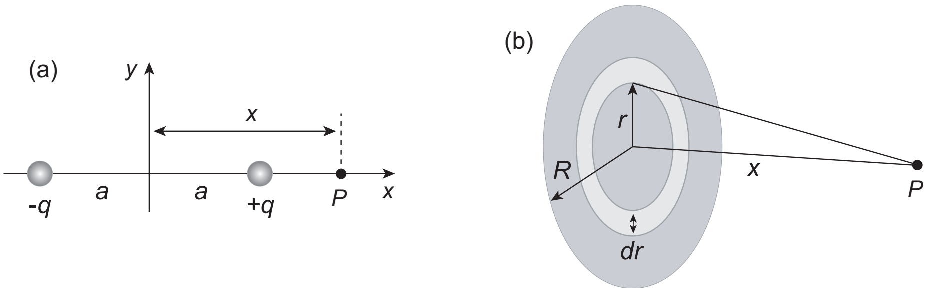


Fig. 2.6 (a) An electric dipole charge configuration used in Example 2.3. (b) Uniformly charged disk of radius R and surface charge density σ used in Example 2.4.

> Linear superposition of electrostatic forces and electric fields in Eqs. (2.32) and (2.33) corresponds to *vector sums*, since $\boldsymbol{F}$ and $\boldsymbol{E}$ are vectors, but linear superposition of potentials in Eq. (2.34) entails an ordinary *arithmetic sum* over scalar quantities Φ_i, which is typically easier to deal with than a sum over vectors.

Examples 2.3 and 2.4 illustrate calculating the potential Φ of a charge distribution and then taking its gradient to give the electric field.

Example 2.3 Let's calculate the electric field for the electric dipole charge configuration displayed in Fig. 2.6(a) at the point P, and also at a very large distance $x \gg a$ along the x axis from the charges. At the point P the potential Φ is

$$\Phi = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{x-a}+\frac{-q}{x+a}\right) = \frac{1}{2\pi\varepsilon_0}\left(\frac{qa}{x^2-a^2}\right).$$

The electric field is then oriented along the x axis and given by minus the gradient of the potential

$$E_x = -\boldsymbol{\nabla}\Phi = -\frac{d\Phi}{dx} = \frac{1}{4\pi\varepsilon_0}\left(\frac{aqx}{(x^2-a^2)^2}\right).$$

The scalar potential and electric field may be approximated as

$$\Phi \simeq \frac{1}{2\pi\varepsilon_0}\left(\frac{aq}{x^2}\right) \qquad E_x \simeq -\frac{d\Phi}{dx} = \frac{1}{\pi\varepsilon_0}\left(\frac{aq}{x^3}\right),$$

if $x \gg a$, so that $x^2 - a^2 \sim x^2$.

Example 2.4 Let's calculate the electric field at the point P along the x axis for the charged disk of radius R and uniform surface charge density σ illustrated in Fig. 2.6(b). For a ring of radius r and width dr the charge element at a distance $(r^2+x^2)^{1/2}$ from the point P is $dq = (2\pi r dr)\sigma$. Then

$$d\Phi = \frac{1}{2\varepsilon_0}\left(\frac{\sigma r dr}{\sqrt{x^2+r^2}}\right)$$

and the total Φ at the point P is obtained by integration over the disk,

$$\begin{aligned}\Phi = \int d\Phi &= \int_0^R \frac{1}{2\varepsilon_0}\left(\frac{\sigma r dr}{\sqrt{x^2+r^2}}\right)\\ &= \frac{\sigma}{2\varepsilon_0}\int_0^R \frac{r\,dr}{\sqrt{x^2+r^2}}\\ &= \frac{\sigma}{4\varepsilon_0}\int_0^R \frac{d(x^2+r^2)}{\sqrt{x^2+r^2}}\\ &= \frac{\sigma}{2\varepsilon_0}\left[\sqrt{x^2+r^2}\right]_0^R\\ &= \frac{\sigma}{2\varepsilon_0}\left(\sqrt{x^2+R^2}-x\right),\end{aligned}$$

where we changed variables in the third line. Finally, the electric field at P is minus the gradient of the potential,

$$E_x = -\frac{d\Phi}{dx} = \frac{\sigma}{2\varepsilon_0}\left(1-\frac{x}{\sqrt{x^2+R^2}}\right),$$

where by symmetry $\boldsymbol{E}$ has only x components.

For relatively simple charge distributions like those in Examples 2.3 and 2.4, we can work out the potentials easily and then determine the electric field by taking the gradient of the potential. However, for more complex situations we may need more powerful and systematic ways to determine potentials. In the next section we show that finding the potentials can be cast in the form of solving a second-order partial differential equation. This approach may be preferred over the ones we have examined so far, particularly for problems with complicated boundary conditions.

2.8 The Poisson and Laplace Equations

We already know from Eq. (2.22) that the curl of $\boldsymbol{E}$ vanishes. What is the divergence of the electric field equal to? From Eq. (2.30) the divergence of $\boldsymbol{E}$ is

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = \boldsymbol{\nabla}\cdot(-\boldsymbol{\nabla}\Phi) = -\nabla^2\Phi,$$

and comparing with Gauss's law (2.16) gives *Poisson's equation,*

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon_0} \qquad \text{(Poisson's equation),} \tag{2.35}$$

where in cartesian coordinates the *Laplacian operator* ∇^2 is given by [see Eq. (A.19)][8]

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \tag{2.36}$$

which operates on a scalar to return a scalar. For regions where there is no charge density $\rho = 0$ and Poisson's equation reduces to *Laplace's equation,*

$$\nabla^2 \Phi = 0 \qquad \text{(Laplace's equation).} \tag{2.37}$$

By construction, solving Poisson's equation (or Laplace's equation if $\rho = 0$) is equivalent physically to solving for the potential Φ using Eq. (2.23), and once Φ has been determined by either means the electric field can be calculated from Eq. (2.30).

Laplace's equation (2.37) is *linear:* if each of a set of potentials $\{\Phi_1, \Phi_2, \Phi_3, \dots, \Phi_n\}$ satisfy it, then a linear combination of those solutions,

$$\Phi \equiv a_1\Phi_1 + a_2\Phi_2 + a_3\Phi_3, + \dots + \Phi_n,$$

where the a_i are arbitrary constants, also satisfies it.

Example 2.5 Earlier it was asserted that Eq. (2.26a) gives the scalar potential $\Phi(\boldsymbol{x})$ for a 3D continuous charge distribution. If correct, this $\Phi(\boldsymbol{x})$ should be a solution of the 3D Poisson equation (2.35). Let's check that it is. From Eq. (2.26a),

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x'$$

and applying the Laplacian operator to both sides of this equation gives

$$\nabla^2 \Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \int \rho(\boldsymbol{x}') \nabla^2 \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) d^3x'.$$

Evaluating the right side is potentially tricky in that the integrand is singular as $\boldsymbol{x}' \to \boldsymbol{x}$, but this may be handled elegantly using that from Eq. (A.67) the Laplacian of $|\boldsymbol{x} - \boldsymbol{x}'|^{-1}$ is proportional to a Dirac delta function.

$$\nabla^2 \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) = -4\pi\delta(\boldsymbol{x} - \boldsymbol{x}'),$$

giving immediately

$$\nabla^2 \Phi(\boldsymbol{x}) = -\frac{1}{\varepsilon_0} \int \rho(\boldsymbol{x}')\, \delta(\boldsymbol{x} - \boldsymbol{x}')\, d^3x' = -\frac{\rho(\boldsymbol{x})}{\varepsilon_0},$$

which is Poisson's equation (2.35). A longer solution that doesn't invoke the Dirac δ-function may be found in Ch. 1 of Jackson [15].

[8] The Laplacian operator in spherical coordinates is given in Eq. (A.50) and in cylindrical coordinates in Eq. (A.57). The Poisson and Laplace equations are partial differential equations, which are typically more difficult to deal with than ordinary differential equations. We shall address method of solution for these equations shortly.

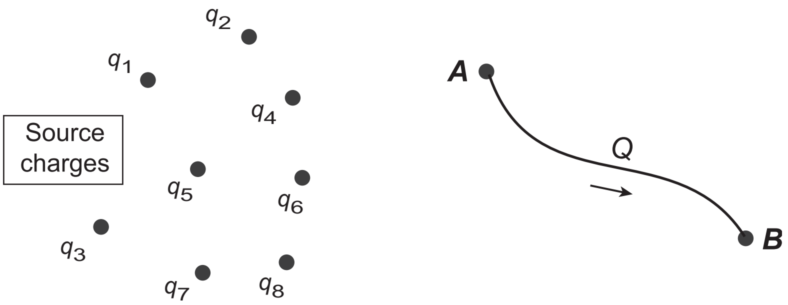


Fig. 2.7 Moving an electrical charge Q on a path between endpoints $\boldsymbol{A}$ and $\boldsymbol{B}$ in an electric field generated by a set of source charges.

2.9 Work and Energy in Electric Fields

A question of fundamental importance in electrostatics is illustrated in Fig. 2.7: if we have a stationary configuration of source charges and a test charge Q is moved along some path between two points $\boldsymbol{A}$ and $\boldsymbol{B}$, *how much work will be done?*

2.9.1 Work to Move a Test Charge

The electrical force exerted on the charge Q is the product of Q and the electric field $\boldsymbol{E}$ generated by the source charges, $\boldsymbol{F} = Q\boldsymbol{E}$. Thus a minimal force $-Q\boldsymbol{E}$ must be exerted at each point to move the charge along the path and the total work W done is given by the line integral

$$W = \int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{F} \cdot d\boldsymbol{l} = -Q \int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E} \cdot d\boldsymbol{l} = Q\,[\Phi(\boldsymbol{B}) - \Phi(\boldsymbol{A})], \tag{2.38}$$

where Eq. (2.27) has been used. Because the work was done against an electric field, it is *independent of path.* Therefore the electrostatic force is *conservative,* meaning that it depends only on the *difference in potentials between the endpoints of the path* and not on the nature of the path followed. If we wish to move the charge from an infinite distance away to a point $\boldsymbol{B}$,

$$W = Q\,[\Phi(\boldsymbol{B}) - \Phi(\infty)], \tag{2.39}$$

so if we make the standard choice that the reference point for the potential is $\Phi(\infty) = 0$, the work done in moving a test charge from infinity to a point $\boldsymbol{x} \equiv \boldsymbol{B}$ is

$$W = Q\Phi(\boldsymbol{x}). \tag{2.40}$$

Thus $\Phi(\boldsymbol{x}) = W/Q$ and the scalar potential $\Phi(\boldsymbol{x})$ may be interpreted as the *work per unit charge* required to move a test charge from infinity to the position $\boldsymbol{x}$ in an electric field. It may also be interpreted as a potential energy stored in the fields of the assembled charge configuration that could be released by moving all charges back to infinity. Example 2.6 illustrates using Eq. (2.40) to calculate the total work done in assembling a set of charges.

Example 2.6 Let's use Eq. (2.40) to calculate the total work done in assembling a set of n charges q_i, by bringing them from infinity to their final positions in a local assembly of charges. The first charge costs nothing to move, since there are no assembled charges and no electric field to fight against. Adding each additional charge will require the work specified in Eq. (2.40) summed pairwise over contributions from all charges. Therefore the total work to assemble the charge distribution is

$$\begin{aligned} W &= \frac{1}{2}\frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{q_i q_j}{r_{ij}} \\ &= \frac{1}{2}\sum_{i=1}^{n} q_i\left(\sum_{j\neq i}^{n}\frac{1}{4\pi\varepsilon_0}\frac{q_j}{r_{ij}}\right) \\ &= \frac{1}{2}\sum_{i=1}^{n} q_i\,\Phi(\boldsymbol{r}_i), \end{aligned} \tag{2.41}$$

where r_{ij} is the distance between charges i and j, the initial factor of $\frac{1}{2}$ is to correct for the double counting of pairs in the double summation, and the factor in parentheses on the second line represents the potential $\Phi(\boldsymbol{r}_i)$ at the position $\boldsymbol{r}_i$ of charge q_i generated by all the other charges [see Eq. (2.25)]. Equation (2.41) represents the work required to assemble the n charges into some local configuration. Therefore, Eq. (2.41) also represents a total (potential) energy stored in the electric fields of the final assembly of charges that could be released by moving all charges back to infinity.

2.9.2 Energy of a Continuous Charge Distribution

We have seen in previous sections how to calculate the energy of a discrete set of charges. Let's now address the energy of a continuous distribution of charge. From Eq. (2.41) we infer that assembling a continuous volume charge requires an amount of work

$$W = \frac{1}{2}\int \rho\,\Phi\,d\tau, \tag{2.42}$$

where $d\tau$ is a volume element. This can be rewritten to eliminate the charge density ρ and the scalar potential Φ in favor of the electric field $\boldsymbol{E}$ in the following way. Use Gauss's law $\rho = \varepsilon_0\boldsymbol{\nabla}\cdot\boldsymbol{E}$ from Eq. (2.16) to express ρ in terms of $\boldsymbol{E}$,

$$W = \frac{\varepsilon_0}{2}\int(\boldsymbol{\nabla}\cdot\boldsymbol{E})\,\Phi\,d\tau, \tag{2.43}$$

and integrate this by parts to give

$$W = \frac{\varepsilon_0}{2}\left(-\int_V \boldsymbol{E}\cdot(\boldsymbol{\nabla}\Phi)\,d\tau + \oint_S \Phi\boldsymbol{E}\cdot d\boldsymbol{a},\right) \tag{2.44}$$

where $d\boldsymbol{a} \equiv \boldsymbol{n}\,da$. But from Eq. (2.30), $\boldsymbol{\nabla}\Phi = -\boldsymbol{E}$, so

$$W = \frac{\varepsilon_0}{2}\left(\int_V E^2\,d\tau + \oint_S \Phi\boldsymbol{E}\cdot d\boldsymbol{a},\right) \tag{2.45}$$

where $E \equiv |\boldsymbol{E}|$, and the integration volume V in the first term is large enough to enclose all charge. If we let $V \to \infty$ then the second (surface) term tends to zero relative to the first term and we obtain

$$W = \frac{\varepsilon_0}{2} \int E^2 d\tau, \tag{2.46}$$

where it is understood that the integration is over all space. The use of Eq. (2.46) is illustrated in Example 2.7.

Example 2.7 Let's calculate the energy of a uniformly charged spherical shell of total charge Q and radius R. From the solution of Problem 2.4, inside the sphere $\boldsymbol{E} = 0$ and outside

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0} \frac{Q}{r^2} \hat{\boldsymbol{r}} \quad \longrightarrow \quad E^2 = \frac{Q^2}{(4\pi\varepsilon_0)^2 r^4},$$

where we work in spherical coordinates. Therefore, from Eq. (2.46),

$$W = \frac{Q^2}{8\pi\varepsilon_0} \int_R^\infty \frac{dr}{r^2} = \frac{1}{8\pi\varepsilon_0} \frac{Q^2}{R},$$

where all contributions to the energy have come from fields outside the sphere.

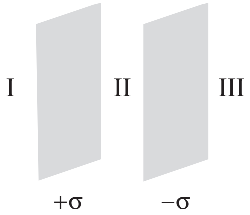
Background and Further Reading

Introductions to the material of this chapter may be found in Corson and Lorrain [4], Griffiths [8], Kamberaj [17], Lorrain and Corson [21], and Purcell and Morin [23]. More advanced treatments may be found in Jackson [15], Garg [6], Chaichian et al [3], Zangwill [27], and Wald [25].

Problems

2.1 Find the electric field produced by an infinite plane that carries a uniform surface charge σ.

2.2 Two infinite parallel planes have equal but opposite charge densities $\pm\sigma$.



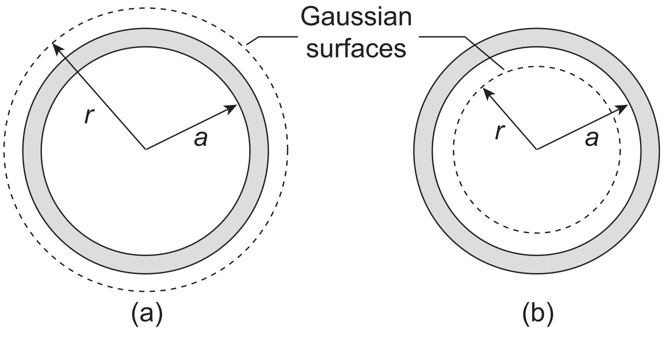
Find the electric fields in regions I, II, and III.

2.3 An isolated good conductor in electrostatic equilibrium (no net motion of electrons) has the following properties.

1. The interior electric field is zero.
2. Any excess charge resides on the surface.
3. The electric field just outside a charged conductor is perpendicular to the surface with a magnitude σ/ε_0, where σ is the surface charge density at that point.
4. If the conductor is of irregular shape, the surface charge density σ is greatest where the surface has the largest local curvature (smallest radius of curvature).

Use Gauss's law to prove each of these properties.

2.4 A thin spherical shell of radius a has a total charge of Q distributed uniformly over its surface, as illustrated in the following figure.



Find the electric field inside and outside the spherical shell using the Gaussian surfaces indicated by dashed circles.

2.5 P rove that the components of the electric field $\boldsymbol{E}$ satisfy the constraints of Eq. (2.31).

3 Electrostatic Boundary Value Problems

In the preceding chapter we introduced the Poisson equation in Eq. (2.35), with solutions corresponding to scalar potentials $\Phi(\boldsymbol{x})$ in the presence of a charge density, and the Laplace equation in Eq. (2.37) with solutions corresponding to scalar potentials $\Phi(\boldsymbol{x})$ in the absence of a charge density. Those solutions result from solving the corresponding partial differential equations subject to boundary conditions, which can be a highly nontrivial matter. In this chapter we address the nature of the solutions of the Poisson and Laplace equations, and some actual means of obtaining solutions.

3.1 Properties of Conductors

There are a number of categories for the classification of matter in materials science. One of the most fundamental distinctions is between[1]

1. *insulators* (often termed *dielectrics* in electromagnetism) which correspond to matter with charge carriers tightly bound to atoms or molecules that do not transport electrical charge well, and
2. *conductors*, which have many delocalized charge carriers that are free to transport electrical charge. Conductors are also often called *metals*.

The charge carriers are typically electrons, electron holes, or ions, but for purposes of discussion we shall normally assume electrons to be the charge carriers.

The free mobility of charge carriers in good conductors leads to many of their basic properties. A good conductor in electrostatic equilibrium is expected to exhibit the following properties (see Problem 2.3)

1. The electric field inside a conductor vanishes, $\boldsymbol{E} = 0$. Qualitatively this is because if there were an electric field inside the conductor electrons would be accelerated, violating the assumption of electrostatic equilibrium. More precisely, consider Fig. 3.1(a) where a conducting slab is immersed in an external electric field $\boldsymbol{E}$. Initially the external field will attract negative charges to the left side of the slab, leaving a net positive charge on the right side of the slab. This polarization of charge will create an internal electric field $\boldsymbol{E}'$ that opposes the external field (remember the convention: the electric field vector points from positive to negative charge). This piling up of negative charges

[1] There are also things in between such as *semimetals* or *semiconductors* that transport charge poorly, or only under certain conditions. We will leave finer classification to materials science for now and concentrate on the behavior of good (ideal) conductors and insulators

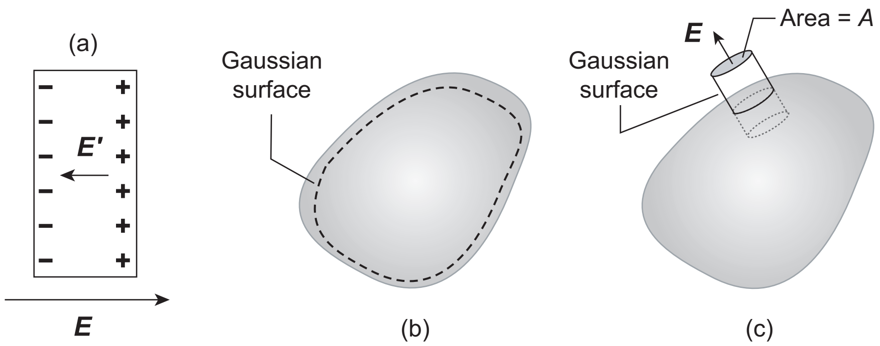


Fig. 3.1 (a) Conducting slab in a uniform electric field $\boldsymbol{E}$. (b) Gaussian surface (dashed curve) inside a conductor. (c) Cylindrical gaussian surface perpendicular to the surface of a conductor.

on one side and positive charge on the right in response to an external field is called an *induced charge*. Charge will continue to flow until the induced internal electric field $\boldsymbol{E}'$ exactly cancels the external field, leaving a net zero electric field inside the conductor. Because this is a conductor the charge carriers are extremely mobile and the generation of internal fields that cancel the external field typically occurs on a timescale so short that it can be assumed to be instantaneous for most considerations.

2. Consider the gaussian surface shown in Fig. 3.1(b). From Gauss's law with zero internal electric field (point 1 above),

$$\oint_S \boldsymbol{E} \cdot \boldsymbol{n}\, da = \oint_S \mathbf{0} \cdot \boldsymbol{n}\, da = 0 = \frac{Q}{\varepsilon_0},$$

where Q is the total enclosed charge. Hence the absence of an electric field means necessarily that the charge density $\rho = 0$ in the interior of the conductor.

3. As a consequence of Gauss's law, any excess charge in a conductor must *reside at the surface of the conductor* (see Problem 2.3). This follows immediately from point 2 above. Since the gaussian surface (dashed curve) in Fig. 3.1(b) can be place arbitrarily close to the surface, any excess charge can only exist at the surface.

4. Any electric field *outside the conductor* is generated by the surface charge and is perpendicular to the surface. Consider Fig. 3.1(c), where we construct a cylindrical gaussian surface with the end faces parallel to the surface of the conductor that is partially inside and partially outside the conductor. If at a point on the surface $\boldsymbol{E}$ had a component tangent to the surface of the conductor, this would cause electrons to move along the surface and disturb electrostatic equilibrium. Thus $\boldsymbol{E}$ is perpendicular the the surface of the conductor and there is no flux through the curved part of the gaussian cylinder. There is also no flux through the flat face of the cylinder *inside the conductor*, because $\boldsymbol{E}$ is zero there (point 1 above). Hence the net flux is only through the flat face of the gaussian cylinder outside the conductor. Applying Gauss's law to this surface,

$$\oint_S \boldsymbol{E} \cdot \boldsymbol{n}\, da = \int_{\text{base}} |\boldsymbol{E}| da = EA = \frac{Q}{\varepsilon_0} = \frac{\sigma}{\varepsilon_0},$$

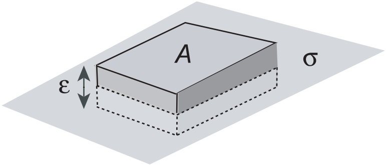


Fig. 3.2 Discontinuous normal of $\boldsymbol{E}$ at surface of a conductor. The surface charge density is σ and the top area of the rectangular gaussian surface is A.

where σ is the surface charge density and A is the area of the endplate of the cylinder. Thus any external electric field is proportional to the surface charge density and perpendicular to the surface.

5. The potential $\Phi(\boldsymbol{x})$ has the same value at each point in the conductor (a conductor is an *equipotential*), since for any two points in the conductor or on its surface,

$$\Phi(\boldsymbol{A}) = \Phi(\boldsymbol{B}) = -\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l} = 0,$$

because $\boldsymbol{E} = 0$.

6. If the conductor is of irregular shape, the surface charge density σ is greatest where the surface has the largest local curvature (smallest radius of curvature). *Proof:* Consider an irregularly shaped conductor, as in Fig. 3.1(b), and partition the surface into small elements of area da_i subtending equal angles measured from the center of the conductor. Points 1 and 3 above then require that $\sigma_i\, da_i$ be constant, and since da_i depends on the radius of curvature, the smaller the radius of curvature (the larger the curvature) the larger the local surface charge density σ.

If conductors are present in a problem, these properties of conductors will often play a significant role in determining boundary conditions and corresponding solutions.

3.2 Capacitance

By Coulomb's law the electric field for an isolated conductor is proportional to the source charge Q and therefore the potential Φ is also, so Q and Φ are proportional to each other. The constant of proportionality is called the *capacitance* C. For an isolated conductor carrying a charge Q with potential Φ, the capacitance is

$$C = \frac{Q}{\Phi}. \tag{3.1}$$

In the SI system the charge is measured in coulombs, the potential in volts, and the capacitance in *farads* (F),[2] with 1 F ≡ 1 coulomb/volt.

[2] The farad is a very large unit for typical phenomena, so it is common to use microfarads ($1\,\mu\mathrm{F} = 10^{-6}\,\mathrm{F}$) and picofarads ($1\,\mathrm{pF} = 10^{-9}\,\mathrm{F}$) as units of capacitance in practical calculations.

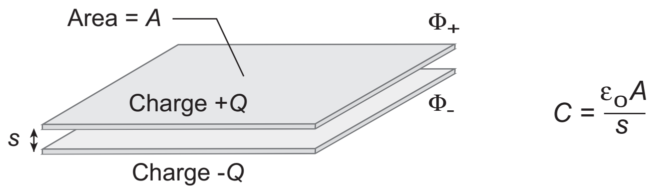


Fig. 3.3 A parallel-plate capacitor. Adapted from Ref. [23].

3.2.1 Parallel-Plate Capacitors

Capacitance is also a very useful concept in dealing with the change and potential associated with two or more conductors. The generic example is the *parallel-plate capacitor* illustrated in Fig. 3.3. For two conductors the capacitance is defined to be the charge on the positive plate divided by the difference in potential between the two plates. If we define the difference in potential for two conductors to be V,

$$V \equiv \Phi_+ - \Phi_-, \tag{3.2}$$

the capacitance is given by

$$C = \frac{Q}{V} \qquad \text{(two conductors).} \tag{3.3}$$

For the parallel-plate capacitor in Fig. 3.3 the capacitance depends only on the geometry,

$$C = \frac{\varepsilon_0 A}{s} \qquad \text{(parallel-plate capacitor),} \tag{3.4}$$

where A is the surface area of a plate and s is the separation between the (assumed parallel) plates.

As will be discussed in Section 4.1, the capacitance of a capacitor like that in Fig. 3.3 can be increased significantly by replacing the air gap between the electrodes with a layer of insulating (dielectric) material such as teflon. As will be illustrated in Box 4.1, this is a consequence of the electric field polarizing the charge distribution in the dielectric layer between the plates.

3.2.2 Energy Stored in a Capacitor

Charging a capacitor (for example, by connecting the two plates in Fig. 3.3 to the poles of a battery) stores energy in the electric field created between the oppositely-charged electrodes of the device. We may determine how much energy by computing the work done to charge the capacitor to a particular level. Consider the parallel-plate capacitor of Fig. 3.3. If at some point in the charging process the charge on the positive plate is q, from Eqs. (2.38) and (3.3) the work increment dW required to add the next charge increment dq is given by

$$dW = \left(\frac{q}{C}\right) dq, \tag{3.5}$$

implying that the total work that must be done to charge the plate from $q = 0$ to $q = Q$ is

$$W = \int dW = \int_0^Q \left(\frac{q}{C}\right) dq = \frac{Q^2}{2C}. \tag{3.6}$$

Therefore, using $Q = CV$ from Eq. (3.3), the work required to charge to a potential difference V between the electrodes is

$$W = \frac{1}{2} C V^2, \tag{3.7}$$

for a capacitor having capacitance C. For a parallel-plate capacitor this takes the specific form

$$W = \frac{1}{2}\frac{A\varepsilon_0}{d} V^2, \tag{3.8}$$

where A is the area of a plate and d is the separation of the plates.

3.3 Electric Fields and Surface Charge Layers

The electric field exhibits a discontinuity if a surface charge layer is crossed. Consider Fig. 3.2, which shows a small piece of a surface having a surface charge density σ. We place a very thin rectangular gaussian box of height ε that extends vertically just below and just above the surface. From Gauss's law,

$$\oint_S \boldsymbol{E} \cdot d\boldsymbol{a} = \frac{Q}{\varepsilon_0} = \frac{\sigma A}{\varepsilon_0},$$

where $Q = \sigma A$ is the enclosed charge and $d\boldsymbol{a} = \boldsymbol{n} a$. In the limit that $\varepsilon \to 0$ the sides of the box contribute no flux. The electric fields due to the surface charge are perpendicular to the local plane and parallel to $\boldsymbol{n}$, so

$$\oint_S \boldsymbol{E} \cdot d\boldsymbol{a} = E \int da = EA$$

and

$$\boldsymbol{E} = \frac{\sigma}{2\varepsilon_0} \hat{\boldsymbol{n}}, \tag{3.9}$$

where $\hat{\boldsymbol{n}}$ is a unit vector perpendicular to the surface and pointing away from it in Fig. 3.2. Then the difference between electric fields above and below the plane is

$$\boldsymbol{E}_{\text{above}} - \boldsymbol{E}_{\text{below}} = \frac{\sigma}{\varepsilon_0} \hat{\boldsymbol{n}}, \tag{3.10}$$

Thus, the electric field is discontinuous at the boundary, changing by a magnitude σ/ε_0.

The scalar potential, on the other hand, is continuous across the boundary since if $\boldsymbol{A}$ is a point just below the surface and $\boldsymbol{B}$ is a point just above it,

$$\Phi_{\text{above}} - \Phi_{\text{below}} = -\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E} \cdot d\boldsymbol{l}, \tag{3.11}$$

which tends to zero as the distance between $\boldsymbol{A}$ and $\boldsymbol{B}$ is decreased. The gradient of Φ is related to $\boldsymbol{E}$ by $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$, so it inherits the discontinuity in $\boldsymbol{E}$,

$$\boldsymbol{\nabla}\Phi_{\text{above}} - \boldsymbol{\nabla}\Phi_{\text{below}} = -\frac{\sigma}{\varepsilon_0}\hat{\boldsymbol{n}}. \tag{3.12}$$

This can also be rewritten as

$$\frac{\partial\Phi_{\text{above}}}{\partial n} - \frac{\partial\Phi_{\text{below}}}{\partial n} = -\frac{\sigma}{\varepsilon_0} \qquad \frac{\partial\Phi}{\partial n} = \boldsymbol{\nabla}\Phi\cdot\hat{\boldsymbol{n}}, \tag{3.13}$$

where $\partial\Phi/\partial n$ is the *normal derivative* (directional derivative evaluated in a direction perpendicular to the surface). These boundary conditions are valid only just above and just below the surface, so the above equations are valid only in the limit that we approach the surface in Fig. 3.2 very closely from the top or bottom.

3.4 Properties of Poisson and Laplace Solutions

Since the Poisson or Laplace equations can be difficult to solve with appropriate boundary conditions for many real-world problems, it is useful to catalog those features that are generic. Let us consider one-dimensional equations first before tackling 2D and 3D versions.

3.4.1 One-Dimensional Laplace Equation

Laplace's equation in one dimension is a function of a single variable, which we choose to be x,

$$\nabla^2\Phi = 0 \quad \longrightarrow \quad \frac{\partial^2\Phi}{\partial x^2} = 0 \quad \longrightarrow \quad \frac{d^2\Phi}{dx^2} = 0, \tag{3.14}$$

where we indicate explicitly in the last step that the usual partial differential equation (PDE) reduces to an ordinary differential equation (ODE), if there is only one variable. This ordinary differential equation (3.14) has a general solution

$$\Phi(x) = ax + b, \tag{3.15}$$

which graphs as a straight line parameterized by the constants a and b (there are two parameters because it is a solution to a second-order ordinary differential equation). The values of the parameters are determined by imposing boundary conditions. In this case two boundary conditions are required, since there are two undetermined parameters. In this example, the boundary conditions could consist of specifying $\Phi(x)$ at two different values of x. The boundary conditions in actual applications reflect the detailed physics of the system being modeled by Eq. (3.14).

This solution of the 1D Laplace equation has two unique features (which will carry over in suitable form to 2D and 3D).

1. A solution $\Phi(x)$ is an average of $\Phi(x-c)$ and $\Phi(x+c)$,

$$\Phi(x)=\frac{1}{2}\left[\Phi(x+c)+\Phi(x-c)\right], \tag{3.16}$$

for any c.

2. There can be no local maxima or minima for the solution as a function of the parameter x, which actually follows from the first point. If there were a local maximum or minimum of Φ at some value of x it could not be the average of points on either side of it, contradicting feature 1.

Thus, maxima or minima can occur only at the endpoints of the plot.

3.4.2 2D and 3D Laplace Equations

Let us now move to more realistic 2D and 3D versions of Laplace's equation, where we must now solve equations of the form

$$\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}=0 \qquad \frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}+\frac{\partial^2\Phi}{\partial z^2}=0 \tag{3.17}$$

In general this introduces a higher level of difficulty than for the 1D case because solution of PDE's with boundary conditions are typically more complex than solution of ODEs. We will illustrate for 3D.

The solutions of the Laplace and Poisson equations have two unique features that generalize those found in 1D.

1. The value of the solution $\Phi(\boldsymbol{x})$ at a point $\boldsymbol{x}$ is an average of values over a sphere centered at $\boldsymbol{x}$.

$$\Phi(\boldsymbol{x})=\frac{1}{4\pi R^2}\oint \Phi\, da, \tag{3.18}$$

where the integral is over the surface of a sphere of radius R centered at $\boldsymbol{x}$.

2. Because of point 1, the solution cannot have local minima or maxima and extrema must occur on the boundaries.

Example 3.1 Let's check point 1 for a single point charge q outside a sphere. For Fig. 3.4 the law of cosines gives for the distance d from the charge q to the patch da,

$$d^2=r^2+R^2-2rR\cos\theta.$$

At a single point on the surface of the sphere the potential is (see Fig. 3.4),

$$\Phi=\frac{1}{4\pi\varepsilon_0}\frac{q}{\sqrt{r^2+R^2-2rR\cos\theta}},$$

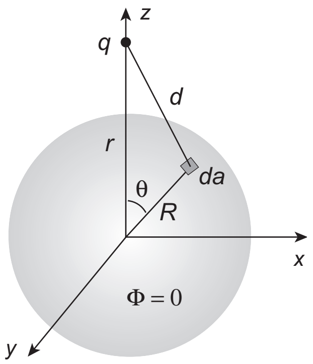


Fig. 3.4 Averaging Φ over a sphere of radius R with a single point charge q external to the sphere.

and the average over the sphere is from Eq. (3.18)

$$\begin{aligned}\Phi_{\text{avg}} &= \frac{1}{4\pi R^2}\frac{q}{4\pi\varepsilon_0}\int (r^2+R^2-2rR\cos\theta)^{-1/2}R^2\sin\theta\,d\theta\,d\phi \\ &= \frac{q}{4\pi\varepsilon_0}\frac{1}{2rR}\left.(r^2+R^2-2rR\cos\theta)^{1/2}\right|_0^\pi \\ &= \frac{q}{4\pi\varepsilon_0}\frac{1}{2rR}[(r+R)-(r-R)] \\ &= \frac{1}{4\pi\varepsilon_0}\frac{q}{r},\end{aligned}$$

which is the potential due to q at the center of the sphere.

3.5 Uniqueness Theorems

Use of the Poisson or Laplace equations to determine the potential Φ requires the solution of partial differential equations with boundary conditions. A simple example of such a problem is illustrated in Fig. 3.5. For partial differential equations it is often not immediately obvious what constitutes appropriate boundary conditions (meaning that they allow a solution and that it is physically well-behaved). The proof that a given set of boundary conditions fits the bill is called a *uniqueness theorem*. Two categories of boundary conditions are common in electrostatics problems.

1. *Dirichlet boundary conditions* correspond to specification of the potential on a closed surface.
2. *Neumann boundary conditions* correspond to specification of the electric field at every point on the surface (which is equivalent to specifying the normal derivative of the potential, or the surface charge density, everywhere the surface).

This leads to two uniqueness theorems.

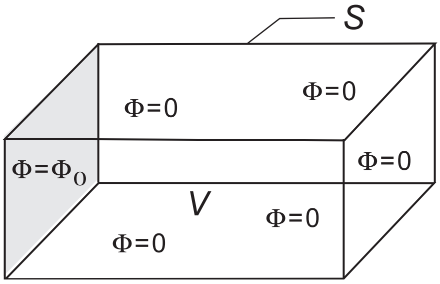


Fig. 3.5 Generic example of an electrostatics problem formulated in a 3D volume V that is bounded by a 2D surface S. In this case the (finite) volume V is surrounded by a rectangular conducting box S with boundary conditions corresponding to five of the six conducting box faces held at zero potential, $\Phi = 0$, and one face at a finite potential, $\Phi = \Phi_0$. Since the potential is being specified on the bounding surface S, this is an example of Dirichlet boundary conditions. The problem would typically be to solve for the electrostatic potential $\Phi(x,y,z)$ in the volume V, subject to the boundary conditions on S. If the charge density $\rho(x,y,z)$ is zero in the volume V, this would correspond to solving the Laplace equation, subject to the boundary conditions.

Uniqueness Theorem I: The solution of Poisson's or Laplace's equations in some volume V is uniquely determined if Φ is specified everywhere on the boundary surface S of the volume V.

The simplest way to set boundary conditions is to specify the value of the scalar field Φ on all surfaces surrounding the region (Dirichlet boundary conditions). Then the first uniqueness theorem applies. However, in some situations we may not know the potential at the boundaries but we do know the total charge on conducting surfaces. This leads to a second uniqueness theorem.

Uniqueness Theorem II:
If a volume V is surrounded by conductors of specified charge density ρ, the electric field is uniquely determined if the total charge on each conductor is specified.

Thus the second uniqueness theorem is appropriate for Neumann boundary conditions. Let us now turn to proof and a deeper look at these uniqueness theorems.

3.6 Uniqueness Theorems by Green's Methods

From a formal point of view, the solution of the Laplace or Poisson equation in a volume V bounded by a surface S with either Dirichlet or Neumann boundary conditions on S (for example, Fig. 3.5) can be obtained using *Green's function methods* [15]. These may be derived beginning with the divergence theorem,

$$\oint_S \boldsymbol{A}\cdot\boldsymbol{n}\,da = \int_V \boldsymbol{\nabla}\cdot\boldsymbol{A}\,d^3x, \tag{3.19}$$

which is valid for any well-behaved vector field $\boldsymbol{A}$ in a volume V bounded by the closed surface S. Let $\boldsymbol{A} = \phi\boldsymbol{\nabla}\psi$ where ϕ and ψ are arbitrary scalar fields. Then

$$\boldsymbol{\nabla}\cdot(\phi\boldsymbol{\nabla}\psi) = \phi\nabla^2\psi + \boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\psi \tag{3.20}$$

and

$$\phi\boldsymbol{\nabla}\psi\cdot\boldsymbol{n} = \phi\frac{\partial\psi}{\partial n}, \tag{3.21}$$

where $\partial/\partial n$ is the normal derivative at the surface, directed from inside to outside the volume V. Substitution of Eqs. (3.20) and (3.21) into the divergence theorem (3.19) then gives *Green's first identity*

$$\int_V (\phi\nabla^2\psi + \boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\psi)\,d^3x = \oint_S \frac{\partial\psi}{\partial n}\,da \qquad \text{(Green's first identity)}. \tag{3.22}$$

If we then subtract from Eq. (3.22) the same expression but with ϕ and ψ interchanged, the $\boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\psi$ terms cancel and we obtain *Green's theorem* (also termed *Green's second identity*)

$$\int_V (\phi\nabla^2\psi - \psi\nabla^2\phi)\,d^3x = \oint_S \left[\phi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\phi}{\partial n}\right]da \qquad \text{(Green's theorem)}. \tag{3.23}$$

Let us now choose a particular function for ψ,

$$\psi \equiv \frac{1}{R} = \frac{1}{\boldsymbol{x}-\boldsymbol{x}'},$$

where $\boldsymbol{x}$ is the observation point and $\boldsymbol{x}'$ is the integration variable, set ϕ equal to the scalar potential $\phi = \Phi$, use Poisson's equation

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon_0}$$

and use that from Eq. (A.67)

$$\nabla^2\left(\frac{1}{R}\right) = 4\pi\delta(\boldsymbol{x}-\boldsymbol{x}')$$

so that Eq. (3.23) becomes

$$\int_V \left[-4\pi\Phi(\boldsymbol{x}')\delta(\boldsymbol{x}-\boldsymbol{x}') + \frac{1}{\varepsilon_0 R}\rho(\boldsymbol{x}')\right]d^3x' = \oint_S \left[\Phi\frac{\partial}{\partial n'}\left(\frac{1}{R}\right) - \frac{1}{R}\frac{\partial\Phi}{\partial n'}da'\right]. \tag{3.24}$$

Then, if $\boldsymbol{x}$ lies within the volume V, this becomes

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \frac{\rho(\boldsymbol{x}')}{R}\,d^3x' + \frac{1}{4\pi}\oint_S \left[\frac{1}{R}\frac{\partial\Phi}{\partial n'} - \Phi\frac{\partial}{\partial n'}\left(\frac{1}{R}\right)\right] da'. \tag{3.25}$$

Notice two important implications of this result.

1. If the surface S goes to infinity and the electric field on S falls off faster than R^{-1}, the surface term in Eq. (3.25) vanishes and we recover the usual result of Eq. (2.26a),

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'.$$

2. In the second (surface) term *both* the the scalar potential Φ and the normal derivative of the scalar potential $\partial\Phi/\partial n'$ appear. The first is associated with Dirichlet boundary conditions and the second with Neumann boundary conditions. Thus Eq. (3.25) is *overdetermined*, since it isn't permitted to impose both types of boundary conditions on the same closed surface. Since it is overdetermined, Eq. (3.25) is not a valid solution. In Section 3.7 we shall address how to correct this deficiency.

Let us now use Green's first identity to show that the use of either (but not both) Dirichlet or Neumann boundary conditions defines a unique potential problem. Assume that the solution is not unique and that there are two potentials, Φ_1 and Φ_2 that satisfy the same Poisson equation. Let

$$U = \Phi_2 - \Phi_1 \tag{3.26}$$

Then $\nabla^2 U = 0$ inside the volume V and on the boundary S either $U = 0$ (Dirichlet boundary conditions) or $\partial U/\partial n = 0$ (Neumann boundary conditions). From Eq. (3.22) with $\phi = \psi = U$,

$$\int_V (U\nabla^2 U + \boldsymbol{\nabla} U\cdot\boldsymbol{\nabla} U)\,d^3x = \oint_S U\frac{\partial U}{\partial n}\,da. \tag{3.27}$$

with the specified properties of U and either type of boundary condition this reduces to

$$\int_V |\boldsymbol{\nabla} U|^2\,d^3x \tag{3.28}$$

which implies that $\boldsymbol{\nabla} U = 0$. Thus, U is constant inside V. For Dirichlet boundary conditions $U = 0$ on S, so $U = \Phi_2 - \Phi_1 = 0$ inside V and the solution is unique. Likewise, for Neumann boundary conditions the solution is unique, apart from an arbitrary additive constant.

Thus we have shown that for either Dirichlet or Neumann boundary conditions the solution of the Poisson equation is *unique*. By similar proofs the solution is unique if the closed surface S has mixed boundary conditions (part with Dirichlet boundary conditions and part with Neumann boundary conditions). However, since Dirichlet and Neumann boundary conditions each define a unique solution, imposing both Dirichlet and Neumann conditions on the same surface (Cauchy boundary conditions) will overdetermine the system and no reliable solution will exist.

Let us summarize some general statements about Dirichlet and Neumann boundary conditions in electrostatics problems gleaned from the preceding discussion.

Box 3.1 The Helmholtz Theorem

Let $\boldsymbol{F}(\boldsymbol{r})$ be any continuous vector field with continuous first partial derivatives. Then $\boldsymbol{F}(\boldsymbol{r})$ can be uniquely expressed in terms of the negative gradient of a scalar potential $\Phi(\boldsymbol{r})$ and the curl of a vector potential $\boldsymbol{A}(\boldsymbol{r})$,

$$\boldsymbol{F}(\boldsymbol{r}) = -\boldsymbol{\nabla}\Phi(\boldsymbol{r}) + \boldsymbol{\nabla} \times \boldsymbol{A}(\boldsymbol{r}).$$

This can also be written as the *Helmholtz decomposition*,

$$\boldsymbol{F}(\boldsymbol{r}) = \boldsymbol{F}_{\mathrm{L}}(\boldsymbol{r}) + \boldsymbol{F}_{\mathrm{T}}(\boldsymbol{r}),$$

where L denotes a *longitudinal component* and T a *transverse component* of a vector field. The theorem is sometimes paraphrased as a uniqueness statement that any vector field whose curl and divergence are known is uniquely determined. To be precise:

> A vector field whose curl and divergence are known everywhere is uniquely determined, provided that the sources vanish at infinity, and that the field vanishes at infinity at least as fast as r^{-2}.

We will have more to say about this in later discussion of gauge invariance and transverse and longitudinal components of the electromagnetic field.

1. One can specify either Dirichlet or Neumann constraints at each point on a boundary, *but not both*. Since either Dirichlet or Neumann conditions are adequate, if both are specified the system is *overdetermined*.[3]

2. It is legitimate to specify parts of a boundary using Dirichlet conditions and other parts using Neumann conditions.

3. The uniqueness property means that a solution obtained by any method that we wish is the correct solution if it satisfies the equations and implements the correct boundary conditions.

Much of the uniqueness of solutions for the Poisson and Laplace equations follows from a more general theorem called the *Helmholtz Theorem* that is described in Box 3.1.

[3] If both Dirichlet and Neumann boundary conditions are specified for a given part of a bounding surface the boundary condition is said to be *Cauchy*. Mathematically, an overdetermined system effectively has more equation constraints than unknowns. Such a system is typically *inconsistent* (no set of values for parameters satisfies all equations), and its equations can be manipulated to obtain contradictory results.

3.7 Boundary-Value Problems by Green Functions

In obtaining the result of Eq. (3.25) (recall: it is not a valid solution, because the boundary conditions are overdetermined) we chose the function ψ to be $1/|\boldsymbol{x}-\boldsymbol{x}'|$, which satisfies[4]

$$\boldsymbol{\nabla}'^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -4\pi\delta^3(\boldsymbol{x}-\boldsymbol{x}'). \tag{3.29}$$

The function $1/|\boldsymbol{x}-\boldsymbol{x}'|$ is one of a class of functions called *Green functions* that satisfy Eq. (3.29). Generally a Green function $G(\boldsymbol{x},\boldsymbol{x}')$ satisfies

$$\boldsymbol{\nabla}'^2 G(\boldsymbol{x},\boldsymbol{x}') = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}') \tag{3.30}$$

where we define

$$G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} + F(\boldsymbol{x},\boldsymbol{x}'), \tag{3.31}$$

with the function $F(\boldsymbol{x},\boldsymbol{x}')$ satisfying the Laplace equation inside the volume V,

$$\boldsymbol{\nabla}'^2 F(\boldsymbol{x},\boldsymbol{x}') = 0. \tag{3.32}$$

The first term on the right side of Eq. (3.31) is the simplest Green function and is called the Green function of free space (physically it gives the response at a point $\boldsymbol{x}$ to a unit charge placed at $\boldsymbol{x}'$),

$$G_0(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}. \tag{3.33}$$

Recall that we derived Eq. (3.25) by substituting $G_0(\boldsymbol{x},\boldsymbol{x}')$ into Green's theorem (3.23), but Eq. (3.25) is not a valid solution because it mixes Dirichlet and Neumann boundary terms in the surface integral. The additional term $F(\boldsymbol{x},\boldsymbol{x}')$ in the generalized Green function (3.31) raises the possibility that if we substitute Eq. (3.31) into Green's theorem, the function $F(\boldsymbol{x},\boldsymbol{x}')$ can be chosen to eliminate from the resulting surface integral either the Dirichlet or the Neumann terms, thus leaving a result with consistent boundary conditions.

Indeed, if we substitute $\phi = \Phi$, and $\psi = G(\boldsymbol{x},\boldsymbol{x}')$ into Green's theorem (3.23) and use the properties (3.30), we obtain a generalization of Eq. (3.25),

$$\begin{aligned}\Phi(\boldsymbol{x}) &= \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' \\ &\quad + \frac{1}{4\pi}\oint_S \left[G(\boldsymbol{x},\boldsymbol{x}')\frac{\partial\Phi}{\partial n'} - \Phi(\boldsymbol{x}')\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'}\right] da'.\end{aligned} \tag{3.34}$$

Now the freedom to choose $F(\boldsymbol{x},\boldsymbol{x}')$ in the definition (3.31) of the Green function means that we can make the surface integral depend on a chosen type of boundary condition.

1. If we desire *Dirichlet boundary conditions* we require that $G(\boldsymbol{x},\boldsymbol{x}') = 0$ for $\boldsymbol{x}'$ on S.

[4] The operator $\boldsymbol{\nabla}'^2$ is the Laplace operator ∇^2 acting on $\boldsymbol{x}'$ rather than $\boldsymbol{x}$.

Then the first term in the surface integral of Eq. (3.34) vanishes and the solution with Dirichlet boundary conditions is

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' - \frac{1}{4\pi}\oint_S \left[\Phi(\boldsymbol{x}')\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'}\right]da'. \tag{3.35}$$

2. If instead we desire *Neumann boundary conditions*, we must be a little more careful. The obvious choice $\partial G(\boldsymbol{x},\boldsymbol{x}')/\partial n' = 0$ for $\boldsymbol{x}'$ on S indeed makes the second term in the surface integral of Eq. (3.34) vanish. But application of Gauss's theorem to Eq. (3.30) indicates that

$$\oint_S \frac{\partial G}{\partial n'}da' = -4\pi, \tag{3.36}$$

implying that the simplest allowable boundary condition is

$$\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'} = -\frac{4\pi}{\Sigma} \qquad \text{for } \boldsymbol{x}' \text{ on } S, \tag{3.37}$$

where Σ is the total area of the bounding surface S. Then the solution with Neumann boundary conditions is

$$\Phi(\boldsymbol{x}) = \langle\Phi\rangle_S + \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' + \frac{1}{4\pi}\oint_S \frac{\partial\Phi}{\partial n'}G\,da', \tag{3.38}$$

where $\langle\Phi\rangle_S$ is the average of the potential over the entire surface.

Thus, we have shown formally how to impose consistent boundary conditions in an electrostatic boundary-value problem.

3.8 Method of Images

In the method of images one replaces the actual Poisson or Laplace problem with a different one (analog problem) that is easier to solve, but that is tailored to have boundary conditions equivalent to those of the actual problem. Then, since the analog problem has the same boundary conditions and is a solution of the Laplace or Poisson equation just as the actual problem, Uniqueness Theorem I supports the validity of the solution.

In Fig. 3.6(a) we consider a charge q placed at a distance d above an infinite conducting plane that is grounded (held at the potential $\Phi = 0$). What is the potential in the region $z > 0$? The presence of the charge will polarize the infinite conducting plane, so the potential will have a part associated with the charge q and a part generated by the polarized infinite conducting plane.. How can we determine the field in the region above the conducting plane when we don't know beforehand the distribution of the polarized charge? One way is to solve Poisson's equation for $z > 0$ with the boundary conditions,

1. $\Phi = 0$ at $z = 0$, since the conducting plane is grounded.
2. $\Phi \to 0$ at very large distances from the charge.

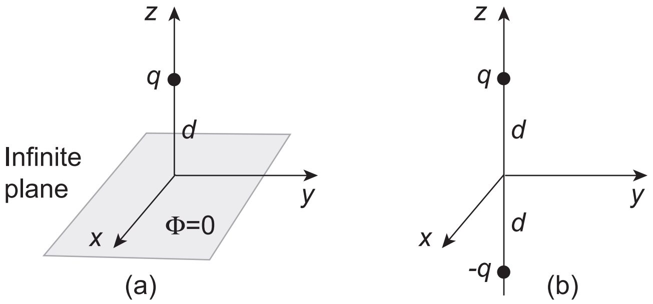


Fig. 3.6 (a) Charge a distance d above an infinite conducting plane held at zero potential. (b) An analogous image-charge configuration with the charge $-q$ on the negative z axis, and no infinite plane.

The *first uniqueness theorem* indicates that there can be *only one independent solution of the Poisson equation*. Thus, if we can find by any means such a solution (including by guessing), it must be the correct one.

Now consider the completely different problem illustrated in Fig. 3.6(b). This problem has the original charge q at a distance d above the origin, but also has an additional charge $-q$ at a distance d below the origin, and there is no conducting plane. The problem in Fig. 3.6(b) is simple and can be solved easily using Coulomb's law,

$$\Phi(x,y,z) = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{\sqrt{x^2+y^2+(z-d)^2}} - \frac{q}{\sqrt{x^2+y^2+(z+d)^2}}\right), \tag{3.39}$$

where the quantities in the denominators are distances between the charge and the point $P(x,y,z)$. But how does this help us? We want the solution for the problem of Fig. 3.6(a), not Fig. 3.6(b). Well, notice that the *boundary conditions are the same* for the two problems: There is a single charge q in the region $z > 0$ (which is where we seek a solution), and from Eq. (3.39),

1. the potential Φ is equal to zero in the $x-y$ plane, and
2. at large distances from the charges, $\Phi \to 0$.

Thus, we conclude that the solution of our original problem in Fig. 3.6(a) is given by Eq. (3.39), which is the solution of the problem Fig. 3.6(b). This approach is called the *method of images.* It can be powerful, but is dependent on thinking up an analog problem that has the same boundary conditions and is easy to solve. Notice that the image charges must in general be put in a region that is not in the domain of the original problem. In the present case we put the image charge in the region $z < 0$ but required a solution for $z > 0$.

Now that the potential has been found in Eq. (3.39) for the upper half plane we can determine the induced charge in the conducting plane caused by the positive charge at $z = d$. From Eq. (3.13) keeping the above contribution,

$$\sigma = -\varepsilon_0 \frac{\partial\Phi}{\partial n}. \tag{3.40}$$

The normal derivative of Φ at the surface is in the z direction, so

$$\sigma = -\varepsilon_0 \frac{\partial \Phi}{\partial z}\bigg|_{z=0}, \tag{3.41}$$

and the derivative may be evaluated from Eq. (3.39) as,

$$\begin{aligned} \frac{\partial \Phi}{\partial z}\bigg|_{z=0} &= \frac{1}{4\pi\varepsilon_0}\left(\frac{-q(z-d)}{[x^2+y^2+(z-d)^2]^{3/2}} + \frac{q(z+d)}{[x^2+y^2+(z+d)^2]^{3/2}}\right)\bigg|_{z=0}, \\ &= \frac{1}{4\pi\varepsilon_0}\left(\frac{qd}{[x^2+y^2+d^2]^{3/2}} + \frac{qd}{[x^2+y^2+d^2]^{3/2}}\right) \\ &= \frac{1}{2\pi\varepsilon_0}\frac{qd}{[x^2+y^2+d^2]^{3/2}} \end{aligned}$$

and from Eq. (3.41)

$$\sigma = -\,\varepsilon_0 \frac{\partial \Phi}{\partial n}\bigg|_{z=0} = -\frac{qd}{2\pi[x^2+y^2+d^2]^{3/2}}. \tag{3.42}$$

We can then obtain the total induced charge Q by integration. Using polar coordinates (r,ϕ) with $r^2 = x^2+y^2$ and $da = r\,dr\,d\phi$,

$$Q = \int_0^{2\pi}\int_0^{\infty} \frac{-qd}{2\pi[x^2+y^2+d^2]^{3/2}}\, r\,dr\,d\phi = \frac{qd}{\sqrt{r^2+d^2}}\bigg|_0^{\infty} = -q. \tag{3.43}$$

So the total induced charge in the infinite sheet has the same magnitude as the polarizing charge q, but with the opposite sign.

Let us also calculate the force $\boldsymbol{F}$ on q produced by the induced charge. From the analog problem displayed in Fig. 3.6(b), the force is, from Coulomb's law,

$$\boldsymbol{F} = -\frac{1}{4\pi\varepsilon_0}\frac{q^2}{(2d)^2}\,\hat{\boldsymbol{z}}, \tag{3.44}$$

where $\hat{\boldsymbol{z}}$ is a unit vector in the z direction. Since the potential, electric field, and force acting on q are expected to be the same in the analog and actual problem, we deduce that the force given by Eq. (3.44) for the analog problem is also the force felt in the actual problem in Fig. 3.6(a).

3.9 Green Function for the Conducting Sphere

As discussed in Section 3.7, solution of the Laplace or Poisson equations in a finite volume V with either Dirichlet or Neumann boundary conditions on the bounding surface S of V can be obtained using Green functions. For example, Eq. (3.35) solves for the potential Φ in terms of a Green function with Dirichelet boundary conditions and Eq. (3.38) solves for the potential in terms of a Green function with Neumann boundary conditions. However, choosing an appropriate Green function for a given problem can be difficult. As discussed by Jackson,[5] for image problems like those described in Section 3.8, the potential due to a

[5] See Section 1.10 of Jackson [15].

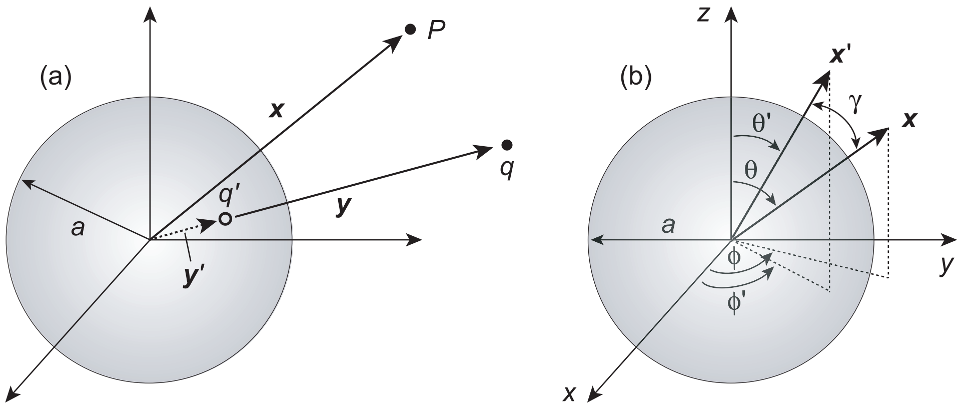


Fig. 3.7 (a) Geometry of image charge method for a charge q outside a conducting sphere, with image charge q' inside the sphere. (b) Coordinates associated with the Green function $G(\boldsymbol{x},\boldsymbol{x}')$ for a conducting sphere.

unit source and its image(s), chosen to satisfy homogeneous boundary conditions is just the Green function of Eqs. (3.35) (for Dirichlet boundary conditions) or (3.38) (for Neumann boundary conditions).

For example, consider the problem of a charge outside a conducting sphere solved by the image method in Problem 3.1 and illustrated in Fig. 3.7(a). In the Green function $G(\boldsymbol{x},\boldsymbol{x}')$, the variable $\boldsymbol{x}'$ refers to the location of the unit source and the variable $\boldsymbol{x}$ is the point P at which the potential is being evaluated. These coordinates and a sphere of radius a are illustrated in Fig. 3.7(b). For Dirichlet boundary conditions on the sphere of radius a, the Green function defined by Eq. (3.30),

$$\boldsymbol{\nabla}'^2 G(\boldsymbol{x},\boldsymbol{x}') = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}'),$$

for a unit source and its image is given by [see Fig. 3.7(a)]

$$\Phi(\boldsymbol{x}) = \frac{q/4\pi\varepsilon_0}{|\boldsymbol{x}-\boldsymbol{y}|} + \frac{q'/4\pi\varepsilon_0}{|\boldsymbol{x}-\boldsymbol{y}'|}, \tag{3.45}$$

with q' and $\boldsymbol{y}'$ chosen such that the potential vanishes on the surface of the sphere in Fig. 3.7(a) ($|\boldsymbol{x}| = a$), which requires that

$$q' = -\frac{a}{y}q \qquad y' = \frac{a^2}{y}, \tag{3.46}$$

and that q be replaced by $4\pi\varepsilon_0$ (to give unit source charge). With these changes the right side of Eq. (3.45) is converted into the Green function

$$G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} - \frac{a}{x'\,|\boldsymbol{x}-(a^2/x'^2)\boldsymbol{x}'|}, \tag{3.47}$$

which can be expressed in spherical coordinates as

$$G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{(x^2+x'^2-2xx'\cos\gamma)^{1/2}} - \frac{1}{(x^2x'^2/a^2+a^2-2xx'\cos\gamma)^{1/2}}, \tag{3.48}$$

where γ is the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$ that is illustrated in Fig. 3.7(b). For the solution (3.35),

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' - \frac{1}{4\pi}\oint_S \left[\Phi(\boldsymbol{x}')\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'}\right] da',$$

of the Poisson equation with Dirichlet boundary conditions, the second term requires the normal derivative $\partial G/\partial n'$. Recall that $\boldsymbol{n}'$ is the unit normal vector *outward from the volume of interest*. Thus, for the solution *outside the sphere* in Fig. 3.7(a), it is inward along $\boldsymbol{x}'$, toward the origin and

$$\left.\frac{\partial G}{\partial n'}\right|_{x'=a} = -\frac{x^2-a^2}{a(x^2+a^2-2ax\cos\gamma)^{3/2}}. \tag{3.49}$$

Therefore, from Eq. (3.35) quoted above, if there is no charge distribution $\rho(\boldsymbol{x}')$ in the problem, the solution *outside the conducting sphere* of the Laplace equation with the potential specified on its surface (Dirichlet boundary conditions) is

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi}\int \Phi(a,\theta',\phi')\frac{a(x^2-a^2)}{(x^2+a^2-2ax\cos\gamma)^{3/2}}d\Omega' \qquad \text{(outside)}, \tag{3.50}$$

where $d\Omega'$ is the element of solid angle at the point (a,θ',ϕ'), and

$$\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi-\phi'). \tag{3.51}$$

If there is a charge distribution $\rho(\boldsymbol{x}')$, then one must add to Eq. (3.51) the first term of Eq. (3.35) with the associated Green function (3.48). For the solution *interior to the sphere* the only thing that changes is that the normal derivative is radially outward, so the sign on the right side of Eq. (3.49) changes. Thus the interior solution is

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi}\int \Phi(a,\theta',\phi')\frac{a(a^2-x^2)}{(x^2+a^2-2ax\cos\gamma)^{3/2}}d\Omega' \qquad \text{(inside)}. \tag{3.52}$$

Example 3.2 Let's illustrate use of the exterior solution (3.50) by considering a conducting sphere of radius a, divided into two hemispherical shells separated by an insulating ring, as illustrated in Fig. 3.8. From Eq. (3.50) the solution for $\Phi(x,\theta,\phi)$ is

$$\Phi(x,\theta,\phi) = \frac{V}{4\pi}\int_0^{2\pi}d\phi'\left(\int_0^1 d(\cos\theta') - \int_{-1}^0 d(\cos\theta')\right)\frac{a(x^2-a^2)}{(x^2+a^2-2ax\cos\gamma)^{3/2}}.$$

By a change of variables in the second integral

$$\theta' \to \pi-\theta' \qquad \phi' \to \phi'+\pi,$$

this can be put in the form

$$\begin{aligned}\Phi(x,\theta,\phi) &= \frac{Va(x^2-a^2)}{4\pi}\int_0^{2\pi}d\phi' \\ &\times \int_0^1 d(\cos\theta')\left[(a^2+x^2-2ax\cos\gamma)^{-3/2} - (a^2+x^2+2ax\cos\gamma)^{-3/2}\right]. \end{aligned}\tag{3.53}$$

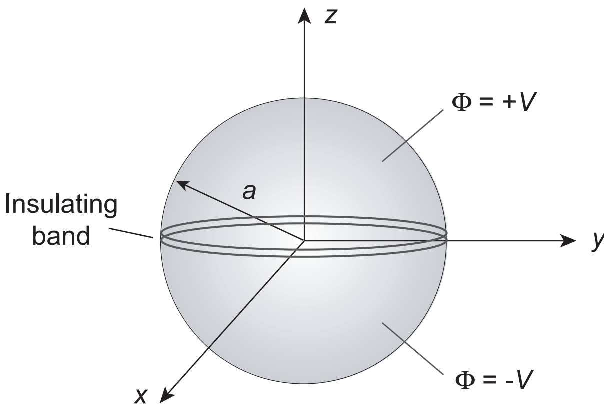


Fig. 3.8 Example 3.2: Two hemispheres of radius a separated by an insulating band lying in the $z = 0$ plane, with the upper hemisphere kept at potential $+V$ and the lower hemisphere kept at potential $-V$.

Because of the complicated dependence among the angles in Eq. (3.51) this cannot be integrated easily in closed form. If one restricts to the positive z axis the integrals can be done, with the result [15]

$$\Phi(z) = V\left[1 - \frac{z^2 - a^2}{z\sqrt{z^2 + a^2}}\right] \qquad \text{(valid on positive } z \text{ axis)}, \tag{3.54}$$

which correctly reduces to $\Phi = V$ at $z = a$. Of potentially more use is to expand the denominator of Eq. (3.53) in a power series and integrate term by term, which yields

$$\Phi(x, \theta, \phi) = \frac{3Va^2}{2x^2}\left[\cos\theta - \frac{7a^2}{12x^2}\left(\frac{5}{2}\cos^2\theta - \frac{3}{2}\cos\theta\right) + \cdots\right]. \tag{3.55}$$

This expansion has been shown to converge rapidly for large x/a, and agrees with the special solution (3.54) for $\cos\theta = 1$ [15].

3.10 Solving the Poisson and Laplace Equations

For relatively simple electrostatics problems solutions may be found using Coulomb's law directly, Gauss's theorem, or image methods, as we have shown in Ch. 2 and the initial sections of this chapter. For more complicated problems these methods may be difficult to apply and a more straightforward way to proceed may be to solve the Poisson or Laplace differential equations directly. One method to do so is by *separation of variables*.

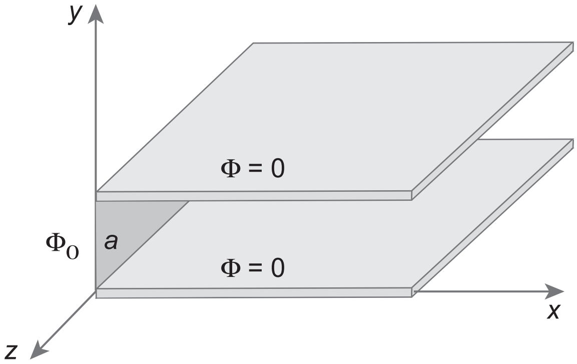


Fig. 3.9 Grounded ($\Phi = 0$) semi-infinite plane-parallel electrodes terminated by a planar electrode at potential Φ_0. The electrodes are assumed to extend infinitely to the right and to be infinite in the direction perpendicular to the page. (Maintaining the difference in potential between the plates at $\Phi = 0$ and $\Phi = \Phi_0$ requires a strip of insulator at the joints.) Adapted from Ref. [4].

3.10.1 Separation of Variables in Cartesian Coordinates

A typical case to be solved is where the potential or the charge density is specified on the boundaries of a region, and we wish to find the potential at arbitrary points in the interior. A standard approach to such a problem is to solve Laplace's equation by the *separation of variables method*, which we consider initially in cartesian coordinates. The basic idea is to assume that the solution can be written in the product form

$$\Phi(x,y,z) = X(x)Y(y)Z(z), \tag{3.56}$$

where $X(x)$, $Y(y)$, and $Z(z)$ are functions only of x, y, and z, respectively. Let us illustrate the method by considering the problem corresponding to Fig. 3.9 [4]. The problem is independent of the z direction, and we assume that there is no charge density between the plates. Therefore, we wish to solve the 2D Laplace equation in cartesian coordinates,

$$\frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} = 0, \tag{3.57}$$

subject to the boundary conditions

$$\Phi = 0 \qquad (y = 0, y = a), \tag{3.58a}$$

$$\Phi = \Phi_0 \qquad (x = 0), \tag{3.58b}$$

$$\Phi \to 0 \qquad (x \to \infty). \tag{3.58c}$$

Inserting the 2D product function

$$\Phi(x,y) = X(x)Y(y) \tag{3.59}$$

into Eq. (3.57) and dividing through by $\Phi = X(x)Y(y)$ gives

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} = 0, \tag{3.60}$$

where we write the derivatives as ordinary derivatives since $X(x)$ and $Y(y)$ are functions of a single variable. Now the second term of Eq. (3.60) is independent of x and the first term is independent of y, and the two terms must always sum to zero. Thus each term must be equal to a constant C_n,

$$\frac{1}{X}\frac{d^2X}{dx^2} = C_1 \qquad \frac{1}{Y}\frac{d^2Y}{dy^2} = C_2. \tag{3.61}$$

Equation (3.60) then implies that $C_1 + C_2 = 0$, so for later convenience we introduce a new constant k and set $C_1 = k^2$ and $C_2 = -k^2$. Therefore, the problem has been reduced to solving two *ordinary differential equations*

$$\frac{d^2X}{dx^2} - k^2X = 0 \tag{3.62a}$$

$$\frac{d^2Y}{dy^2} + k^2Y = 0. \tag{3.62b}$$

As can be verified by substitution, the second equation (3.62b) has a solution

$$Y = A\sin(ky) + B\cos(ky), \tag{3.63}$$

where A and B are arbitrary constants. We may determine constants by imposing the boundary conditions (3.58). From Eq. (3.58a), $\Phi = 0$ at $y = 0$ requires that $B = 0$, and $\Phi = 0$ at $y = a$ requires that

$$k = \frac{n\pi}{a} \qquad (n = 1, 2, 3, \cdots). \tag{3.64}$$

Therefore,

$$Y = A\sin\left(\frac{n\pi y}{a}\right) \qquad (n = 1, 2, 3, \cdots). \tag{3.65}$$

The equation for X in Eq. (3.62a) now takes the form

$$\frac{d^2X}{dx^2} - \left(\frac{n\pi}{a}\right)^2 X = 0, \tag{3.66}$$

which has a solution

$$X = Ge^{n\pi x/a} + He^{-n\pi x/a}, \tag{3.67}$$

as may be verified by substitution. However, the boundary condition $\Phi \to 0$ as $x \to \infty$ of Eq. (3.58c) can be satisfied only if $G = 0$, so we discard the first term of Eq. (3.67) and insert Eqs. (3.65) and (3.67) into Eq. (3.59) to give

$$\Phi(x, y) = C\sin\left(\frac{n\pi y}{b}\right) e^{-n\pi x/a}, \tag{3.68}$$

with C another arbitrary constant.

The solution (3.68) satisfies the boundary conditions (3.58a) and (3.58c), but does not satisfy the boundary condition (3.58b). However, the Laplace equation is linear, meaning that if $\{\Phi_1, \Phi_2, \Phi_3, \cdots, \Phi_n\}$ satisfy it, then the linear combination

$$\Phi = a_1\Phi_1 + a_2\Phi_2 + a_3\Phi_3 + \cdots + a_n\Phi_n$$

where a_n are arbitrary constants, satisfies it also. Therefore, we take as a better approximation a linear combination of solutions (3.68) in the form

$$\Phi(x,y) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right) e^{-n\pi x/a}, \tag{3.69}$$

The boundary condition (3.58b) at $x = 0$ implies that

$$\Phi(0,y) = \Phi_0 = \sum_{n=0}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right), \tag{3.70}$$

which is a Fourier sine series, so we can exploit the corresponding orthogonality properties to evaluate the coefficients c_n. Specifically, multiply both sides of Eq. (3.70) by $\sin[(pxy)/a]$, where p is an integer, and integrate from $y = 0$ to $y = a$,

$$\int_0^a \Phi_0 \sin\left(\frac{p\pi y}{a}\right) dy = \int_0^a \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right) \sin\left(\frac{p\pi y}{a}\right) dy. \tag{3.71}$$

For the integral on the left side of this equation

$$\int_0^a \Phi_0 \sin\left(\frac{p\pi y}{a}\right) dy = \begin{cases} \dfrac{2a\Phi_0}{p\pi} & (\text{ if } p \text{ is odd}), \\ 0 & (\text{if } p \text{ is even}), \end{cases} \tag{3.72}$$

while for the integral on the right side,

$$\int_0^a c_n \sin\left(\frac{n\pi y}{a}\right) \sin\left(\frac{p\pi y}{a}\right) dy = \begin{cases} 0 & (\text{ if } p \neq n), \\ \dfrac{a}{2}c_n & (\text{if } p = n), \end{cases} \tag{3.73}$$

Comparing Eqs. (3.72) and (3.73), we conclude that the expansion coefficients are given by

$$c_n = \begin{cases} \dfrac{4\Phi_0}{n\pi} & (\text{ if } n \text{ is odd}), \\ 0 & (\text{if } n \text{ is even}), \end{cases} \tag{3.74}$$

and the potential as a function of x and y is

$$\Phi(x,y) = \frac{4\Phi_0}{n\pi} \sum_{n=1,3,5,\cdots}^{\infty} \frac{1}{n} \sin\left(\frac{nxy}{a}\right) e^{-n\pi x/a}. \tag{3.75}$$

This series should converge fairly rapidly because of the rapid decrease of the factor $e^{-n\pi x/a}/n$ with n. The solution is plotted in Fig. 3.10. A convergent power series is a perfectly adequate way to define a solution, but it happens that the series (3.75) can be summed exactly, with the result [8],

$$\Phi(x,y) = \frac{2\Phi_0}{\pi} \tan^{-1}\left(\frac{\sin(\pi y/a)}{\sinh(\pi x/a)}\right), \tag{3.76}$$

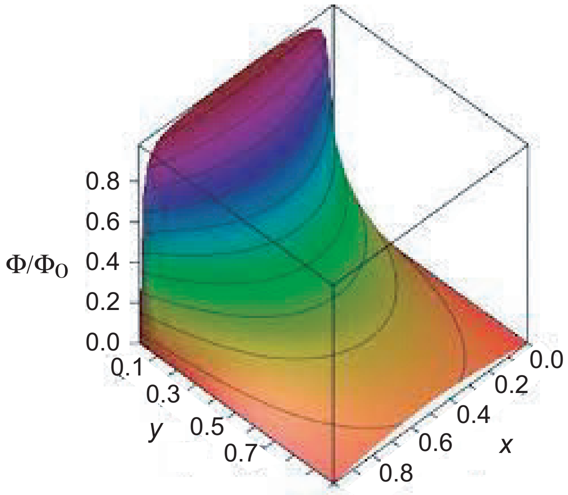


Fig. 3.10 Solution for 2D Laplace equation by separation of variables.(Excuse quality; temporary placeholder).

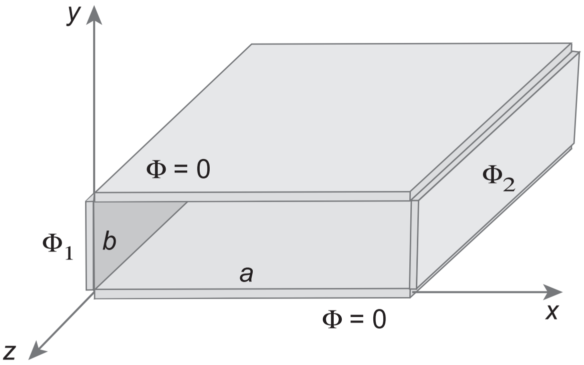


Fig. 3.11 Grounded ($\Phi = 0$) parallel-plane electrodes of width a and separated by a distance b are terminated on two sides with plane electrodes at potentials Φ_1 and Φ_2. The electrodes are assumed to be infinite in the direction perpendicular to the page. Adapted from Ref. [4].

in a more convenient closed form.

Example 3.3 Consider Fig. 3.11, where we wish to calculate the electrostatic potential in the region between the parallel-plane electrodes [4]. There is no z dependence so again let's solve the 2D Laplace equation by the method of separation of variables. This problem has much in common with the example from Fig. 3.9 just worked out, but the boundary conditions are different.

$$\Phi = 0 \qquad (y=0, y=b), \tag{3.77a}$$

$$\Phi = \Phi_1 \qquad (x=0), \tag{3.77b}$$

$$\Phi = \Phi_2 \qquad (x=a), \tag{3.77c}$$

The most general solution is of the form

$$\Phi(x,y) = \sum_{n=1}^{\infty} \left(A_n e^{-n\pi x/b} + B_n e^{n\pi x/b}\right) \sin\left(\frac{n\pi y}{b}\right), \tag{3.78}$$

where the constants A_n and B_n may be determined using the boundary conditions. From Eq. (3.77b), at $x = 0$ we have

$$\Phi_1 = \sum_{n=1}^{\infty} (A_n + B_n) \sin\left(\frac{n\pi y}{b}\right). \tag{3.79}$$

Upon multiplying this by $\sin(p\pi y/b)$ and integrating from $y = 0$ to $y = b$, only the single $p = n$ term survives,

$$\Phi_1 \int_0^b \sin\left(\frac{n\pi y}{b}\right) dy = (A_n + B_n)\frac{b}{2}, \tag{3.80}$$

implying that

$$A_n + B_n = \begin{cases} \dfrac{4\Phi_1}{n\pi} & (\text{ if } n \text{ is odd}), \\ 0 & (\text{if } n \text{ is even}). \end{cases} \tag{3.81}$$

The boundary condition (3.77c) yields another relationhip between A_n and B_n:

$$\Phi_2 = \sum_{n=1}^{\infty} (A_n e^{-n\pi a/b} + B_n e^{n\pi a/b}) \sin\left(\frac{n\pi y}{b}\right). \tag{3.82}$$

Multiplying this expression by $\sin(p\pi y/b)$ and integrating from $y = 0$ to $y = a$ yields

$$A_n e^{-n\pi a/b} + B_n e^{n\pi a/b} = \begin{cases} \dfrac{4\Phi_2}{n\pi} & (\text{ if } n \text{ is odd}), \\ 0 & (\text{if } n \text{ is even}). \end{cases} \tag{3.83}$$

Then from Eqs. (3.81) and (3.83),

$$A_n = \frac{4}{n\pi}\left(\frac{\Phi_1 - \Phi_2 e^{-n\pi a/b}}{1 - e^{-2n\pi a/b}}\right) \qquad B_n = \frac{4e^{-n\pi a/b}}{n\pi}\left(\frac{\Phi_2 - \Phi_1 e^{-n\pi a/b}}{1 - e^{-2n\pi a/b}}\right), \tag{3.84}$$

where $n = 1, 3, 5, \cdots$. This gives $\Phi(x,y)$ when inserted in Eq. (3.78). This solution is plotted in Fig. 3.12.

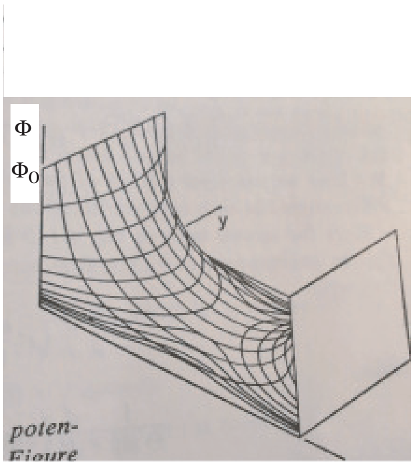


Fig. 3.12 Solution for 2D Laplace equation by separation of variables.(Excuse quality; temporary placeholder).

3.10.2 Separation of Variables in Spherical Coordinates

In the preceding examples of solving the Laplace equation by separation of variables the geometry was rectangular and cartesian coordinates were appropriate. However, for some problems other coordinate systems such as spherical or cylindrical may be more natural. In this section we consider solution of Laplace's equation in spherical coordinates. Laplaces equation in spherical coordinates (r,θ,ϕ) is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2\Phi}{\partial\phi^2}=0, \tag{3.85}$$

To illustrate we will consider problems having axial symmetry (no dependence on ϕ), in which case the Laplacian equation reduces to

$$\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right)+\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right)=0. \tag{3.86}$$

Just as in the cartesian coordinate examples we seek product solutions in which the variables (r,θ) are separated,

$$\Phi(r,\theta)=R(r)\,\Theta(\theta), \tag{3.87}$$

where $R(r)$ is a function only of r and $\Theta(\theta)$ is a function only of θ. Substituting Eq. (3.87) into Eq. (3.86) and dividing through by $R\Theta$ gives,

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)=0, \tag{3.88}$$

where now we use total derivatives rather than partial derivatives since the functions being differentiated are functions of a single variable. By similar logic as in the cartesian case, the second term in Eq. (3.88) is independent of r so the first term must also be independent of r. Thus we write the separated equations as two ordinary differential equations

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)=k \tag{3.89a}$$

$$\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)=-k \tag{3.89b}$$

where the constants are written as k and $-k$, since the sum of the two equations must be zero.

Consider first the R equation in Eq. (3.89a). Multiplying both sides by R and carrying out the leftmost d/dr operation gives

$$r^2\frac{d^2R}{dr^2}+2r\frac{dR}{dr}-kR=0, \tag{3.90}$$

which has a solution

$$R=Ar^n+\frac{B}{r^{n+1}}, \tag{3.91}$$

that when inserted in Eq. (3.89a) requires n and k to be related by,

$$n(n+1)=k. \tag{3.92}$$

Now consider the Θ equation (3.89b). Multiplying by $\Theta\sin\theta$ and using Eq. (3.92), it may be written

$$\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)+n(n+1)\sin\theta\,\Theta=0 \tag{3.93}$$

It is convenient to change variables by letting $\mu=\cos\theta$. By the chain rule, for any function $f(\mu)$ of μ,

$$\frac{df}{d\theta}=\frac{df}{d\mu}\frac{d\mu}{d\theta}=-\sin\theta\frac{df}{d\mu}=-\sqrt{1-\mu^2}\frac{df}{d\mu} \tag{3.94}$$

where we have used $d\mu/d\theta=-\sin\theta$ and $1-\mu^2=\sin^2\theta$. Then, Eq. (3.93) becomes

$$\frac{d}{d\mu}\left[(1-\mu^2)\frac{d\Theta}{d\mu}\right]+n(n+1)\Theta=0 \qquad \text{(Legendre's equation).} \tag{3.95}$$

This is called *Legendre's equation* and its solutions are polynomials in $\cos\theta$ called *Legendre polynomials*, designated by $P_n(\cos\theta)$, where n is termed the order of the polynomial. Normalized Legendre polynomials[6] are typically defined by

$$P_n(\cos\theta)=\frac{1}{2^n n!}\frac{\partial^n}{\partial(\cos\theta)^n}(\cos^2\theta-1)^n. \tag{3.96}$$

Some polynomials are tabulated in Table 3.1. A general solution of Laplace's equation in spherical coordinates assuming axial symmetry is then given by

$$\Phi(r,\theta)=\sum_{n=0}^{\infty}A_n r^n P_n(\cos\theta)+\sum_{n=0}^{\infty}B_n r^{-(n+1)}P_n(\cos\theta) \tag{3.97}$$

[6] The normalization is chosen to make all Legendre polynomials equal to one at $\cos\theta=1$.

Table 3.1 Legendre polynomials $P_n(\cos\theta)$

n	$P_n(\cos\theta)$
0	1
1	$\cos\theta$
2	$\frac{1}{2}(3\cos^2\theta - 1)$
3	$\frac{1}{2}(5\cos^3\theta - 3\cos\theta)$
4	$\frac{1}{8}(35\cos^4\theta - 30\cos^2\theta + 3)$
5	$\frac{1}{8}(63\cos^5\theta - 70\cos^3\theta + 15\cos\theta)$

These functions are a complete set, so arbitrary boundary conditions with axial symmetry can be satisfied. Furthermore, the Legendre polynomials satisfy the orthogonality condition

$$\int_{-1}^{+1} P_m(x)P_n(x)dx = \frac{2}{2n+1}\delta_{mn}, \tag{3.98}$$

or explicitly in spherical coordinates,

$$\begin{aligned}\int_{-1}^{+1} P_m(\cos\theta)P_n(\cos\theta)d(\cos\theta) &= \int_0^{\pi} P_m(\cos\theta)P_n(\cos\theta)\sin\theta d\theta \\ &= \begin{cases} \dfrac{2}{2n+1} & (\text{ if } m = n), \\ 0 & (\text{if } m \neq n), \end{cases}\end{aligned} \tag{3.99}$$

which is important in evaluating the coefficients in Eq. (3.97).

Example 3.4 Consider the conducting ball of radius a in an electric field illustrated in Fig. 3.13. Inside the ball the field will be zero and far outside it will be the undisturbed field $\boldsymbol{E}_0$. Near the ball the field will be distorted by polarization. Let's solve the Laplace equation for the outside field in 2D by separation of variables in spherical coordinates, assuming axial symmetry. We take as boundary conditions

$$\Phi = 0 \qquad (r = a) \tag{3.100}$$

$$\Phi = -E_0 r\cos\theta \qquad (r = \infty) \tag{3.101}$$

At $r = a$ (radius of ball), from Eqs. (3.97) and (3.100)

$$\Phi(r,\theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\theta) + \sum_{n=0}^{\infty} B_n r^{-(n+1)} P_n(\cos\theta) = 0. \tag{3.102}$$

The coefficients A_n and B_n in Eq. (3.102) may be evaluated using the orthogonality of the Legendre polynomials. Multiply this expression by $P_m(\cos\theta)$ and integrate over $d(\cos\theta)$.

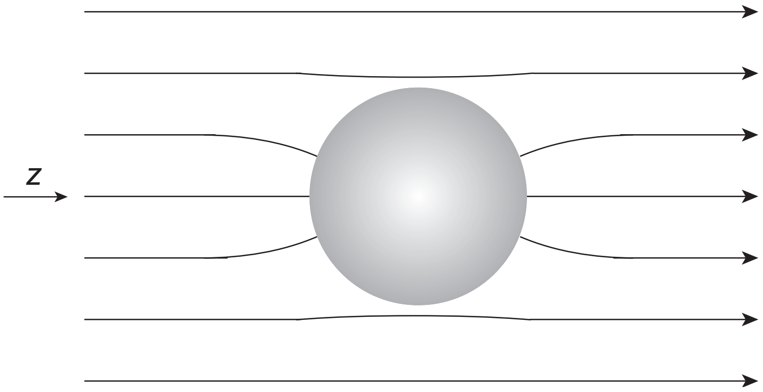


Fig. 3.13 Conducting ball of radius a in an external electric field $\boldsymbol{E}_0$ directed along the z axis. Assume axial symmetry around the left-right axis.

From Eq. (3.99), the only non-vanishing terms are those for which $n = m$. Thus,

$$\begin{aligned}0 &= A_n a^n \int_{-1}^{+1} P_n^2(\cos\theta)d(\cos\theta) + B_n a^{-(n+1)} \int_{-1}^{+1} P_n^2(\cos\theta)d(\cos\theta)\\ &= A_n a^n \left(\frac{2}{2n+1}\right) + B_n a^{-(n+1)} \left(\frac{2}{2n+1}\right),\end{aligned}$$

from which the coefficients are related by

$$B_n = -A_n a^{2n+1}. \tag{3.103}$$

Therefore, from Eqs. (3.102) and (3.103),

$$\Phi(r,\theta) = \sum_{n=0}^{\infty} A_n \left(r^n - \frac{a^{2n+1}}{r^{n+1}}\right) P_n(\cos\theta). \tag{3.104}$$

Now as $r \to \infty$ the second term in parentheses in Eq. (3.104) becomes negligible compared with the first and the boundary condition (3.101) requires that

$$-E_0 r\cos\theta = -E_0 r P_1(\cos\theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\theta). \tag{3.105}$$

Thus, the only non-zero term on the right side is $n = 1$, implying that

$$A_1 = -E_0,$$

with all other $A_n = 0$. Then from Eq. (3.103) all the B_n are zero except for

$$B_1 = -A_1 a^3 = E_0 a^3.$$

Thus the potential at any point (r,θ) is given by

$$\begin{aligned}\Phi(r,\theta) &= -E_0 r\cos\theta + E_0 \frac{a^3\cos\theta}{r^2}\\ &= -E_0\left(1 - \frac{a^3}{r^3}\right) r\cos\theta,\end{aligned} \tag{3.106}$$

where the first term is the potential corresponding to the applied field E_0 and the second term is the induced polarization potential. The electric field components follow from Eq. (3.106) by taking the gradient [see Eq. (A.47) for spherical coordinates],

$$E_r = -\frac{\partial \Phi}{\partial r} = E_0\left(1+\frac{2a^3}{r^3}\right)\cos\theta, \tag{3.107}$$

$$E_\theta = -\frac{1}{r}\frac{\partial \Phi}{\partial \theta} = -E_0\left(1-\frac{a^3}{r^3}\right)\sin\theta. \tag{3.108}$$

At the surface of the conductor we know that

$$E_r|_{r=a} = \frac{\sigma}{\varepsilon_0},$$

so solving for σ and using Eq. (3.107) with $r=a$ gives

$$\sigma = 3\varepsilon_0 E_0 \cos\theta, \tag{3.109}$$

for the induced charge density.

The 3D volume charge density $\rho(\boldsymbol{x})$ appearing in Eq. (2.26a),

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x', \tag{3.110}$$

describes the distribution of charge localized in some finite volume V. In many situations we are interested in the potential $\Phi(\boldsymbol{x})$ resulting from this charge at large distances from the charge distribution such that $|\boldsymbol{x}| \gg |\boldsymbol{x}'|$. Then a series expansion of $1/|\boldsymbol{x}-\boldsymbol{x}'|$ is suggested. Two types of expansions are common in the literature:

1. A Taylor series in the cartesian coordinates $(x, y.z)$.
2. An expansion in terms of spherical harmonics, associated Legendre polynomials, or Legendre polynomials depending on spherical coordinates (r, θ, ϕ).

We will discuss here the multipole expansion in terms of Legendre polynomials or spherical harmonics.

3.10.3 Spherical Harmonic Expansions

A potential due to a unit point charge at $\boldsymbol{x}'$ can be expanded as

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \sum_{l=0}^{\infty} \frac{r_<^l}{r_>^{l+1}} P_l(\cos\theta), \tag{3.111}$$

where $P_l(\cos\theta)$ is a Legendre polynomial [solution (3.96) of the Legendre equation (3.95)], $r_<$ is the smaller and $r_>$ the larger of $|\boldsymbol{x}|$ and $|\boldsymbol{x}'|$,[7] and θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$ (see Fig. 3.14). Equation (3.111) is termed a *multipole expansion*. In this expression

[7] The factor $r_<^l/r_>^{l+1}$ in Eq. (3.111) allows the expansion to be made in terms of either r/r' or r'/r, whichever is smaller.

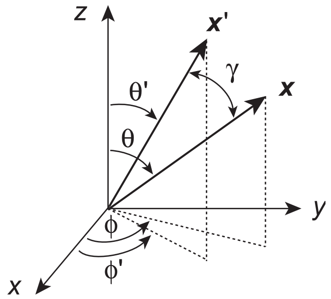


Fig. 3.14 Geometry for the Legendre polynomials and the spherical harmonics in Eqs. (3.111)-(3.114).

- the $l = 0$ term is called the *monopole term*,
- the $l = 1$ term is called the *dipole term*,
- the $l = 2$ term is called the *quadrupole term*, and so on.

The reason for this terminology is suggested by the point-charge distributions discussed in Box 3.2. Comparison of these expressions with the terms in Eq. (3.111) using $r_> = r$ and $r_< = d$ (since $r \gg d$) suggests why Eq. (3.111) is called a multipole expansion: the $l = 1$ term is of the dipole form $P_1(\cos\theta)/r^2$ and the $l = 2$ term is of the quadrupole form $P_2(\cos\theta)/r^3$. (And the $l = 0$ term is $1/r$, which is termed the monopole term.)

Equation (3.111) can be expressed in terms of spherical harmonics using the *spherical harmonic addition theorem*

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_{lm}^*(\theta', \phi') Y_{lm}(\theta, \phi), \tag{3.112}$$

where $P_l(\cos\theta)$ is a Legendre polynomial, $Y_{lm}(\theta, \phi)$ is a spherical harmonic, $\boldsymbol{x}$ has the spherical coordinates (r, θ, ϕ), $\boldsymbol{x}'$ has the spherical coordinates (r', θ', ϕ'), and the angle γ between the vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ is given by

$$\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi'). \tag{3.113}$$

The spherical harmonic addition theorem (3.112) may then be used to express the expansion (3.111) of the potential at $\boldsymbol{x}$ due to a unit charge at $\boldsymbol{x}'$ as

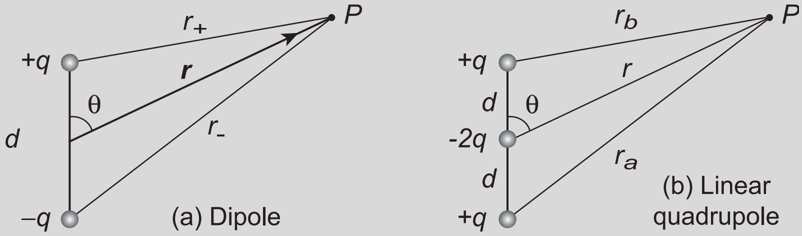
$$\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{1}{2l+1} \frac{r_<^l}{r_>^{l+1}} Y_{lm}^*(\theta', \phi') Y_{lm}(\theta, \phi), \tag{3.114}$$

where angles are defined in Fig. 3.14. This expression gives the potential completely factorized in the coordinates $\boldsymbol{x}$ and $\boldsymbol{x}'$.

If the localized distribution of charge $\rho(\boldsymbol{x}')$ in Eq. (3.110) is assumed to vanish outside of a small sphere of radius R centered on the origin, the potential generated by that charge

Box 3.2 **Multipole moments**

What is the potential generated at P by the following charge distributions?



(a) Dipole

(b) Linear quadrupole

(1) Dipole Potential

For the dipole at P in Fig. (a), the potential is given by

$$\Phi = \frac{q}{4\pi\varepsilon_0}\left(\frac{1}{r_+} - \frac{1}{r_-}\right).$$

From the law of cosines (A.58) the distances are

$$r_\pm^2 = r^2 + \left(\frac{d}{2}\right)^2 \mp rd\cos\theta = r^2\left(1 \mp \frac{d}{r}\cos\theta + \frac{d^2}{4r^2}\right) \simeq r^2\left(1 \mp \frac{d}{r}\cos\theta\right),$$

where $r = |\boldsymbol{r}|$ and the term $d^2/4r^2$ was dropped since we assume that $r \gg d$. Thus,

$$\frac{1}{r_\pm} \simeq \frac{1}{r}\left(1 \mp \frac{d}{r}\cos\theta\right)^{-1/2} \simeq \frac{1}{r}\left(1 \pm \frac{d}{2r}\cos\theta\right),$$

where a binomial expansion was used in the last step. Therefore,

$$\Phi_{\text{dipole}} = \frac{q}{4\pi\varepsilon_0}\left(\frac{1}{r_+} - \frac{1}{r_-}\right) = \frac{q}{4\pi\varepsilon_0}\frac{d\cos\theta}{r^2} = \left(\frac{qd}{4\pi\varepsilon_0}\right)\frac{P_1(\cos\theta)}{r^2} \qquad (r \gg d),$$

and the electric dipole potential is proportional to the Legendre polynomial $P_1(\cos\theta)$ and falls off at large distance as $1/r^2$.

(b) Quadrupole Potential

Carrying out a similar analysis for the linear quadrupole in Fig. (b) above (see Problem 3.3), the linear quadrupole potential is given by

$$\Phi_{\text{quadrupole}} = \left(\frac{2Qd^2}{4\pi\varepsilon_0}\right)\frac{P_2(\cos\theta)}{r^3} \qquad (r \gg d),$$

which is proportional to $P_2(\cos\theta)$ and varies inversely as the cube of the distance.

outside the radius R can be expanded in spherical harmonics as

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{4\pi}{2l+1}\, q_{lm}\frac{Y_{lm}(\theta,\phi)}{r^{l+1}} \tag{3.115}$$

The expansion coefficients q_{lm} are given by

$$q_{lm} = \int Y^*_{lm}(\theta',\phi')(r')^l \rho(\boldsymbol{x}')d^3x' \tag{3.116}$$

and are called *multipole moments*. Spherical harmonics are related to associated Legendre polynomials $P_l^m(\cos\theta)$ by

$$Y_{lm}(\theta,\phi) = (-1)^m \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m(\cos\theta)e^{im\phi}. \tag{3.117}$$

They obey the *orthogonality condition*

$$\int_0^{2\pi} d\phi \int_0^{\pi} Y^*_{l'm'}(\theta,\phi)Y_{lm}(\theta,\phi)\sin\theta d\theta = \delta_{l'l}\delta_{m'm}, \tag{3.118}$$

and *completeness relation*

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi) = \delta(\phi-\phi')\,\delta(\cos\theta-\cos\theta'), \tag{3.119}$$

and behave under complex conjugation as

$$Y^*_{lm}(\theta,\phi) = (-1)^m Y_{l-m}(\theta,\phi). \tag{3.120}$$

The utility of a multipole expansion like Eqs. (3.115) or (3.111) is that, at large distance from the source-charge distribution, the potential may be approximated well by retaining only the first few terms in the multipole expansion, since the terms fall off with distance r as $r^{-(l+1)}$. If we define

1. a total charge Q (monopole moment),
2. the electric dipole moment vector $\boldsymbol{p}$ by

$$\boldsymbol{p} \equiv \int \boldsymbol{x}'\rho(\boldsymbol{x}')d^3x', \tag{3.121}$$

3. and the *traceless quadrupole moment tensor* Q_{ij} by[8]

$$Q_{ij} \equiv \int (3x'_ix'_j - r'^2\delta_{ij})\,\rho(\boldsymbol{x}')d^3x', \tag{3.122}$$

where $r'^2 \equiv |\boldsymbol{x}'|^2$,

the spherical multipole moments in Eq. (3.116) may be expressed in cartesian coordinates

[8] With this definition Q_{ij} is traceless (see Problem 3.7). A quadrupole moment definition without the tracelessness property $Q_{ij} = \frac{1}{2}\int \rho(\vec{x}')x'_ix'_j\,\mathrm{d}^3x'$ is also sometimes encountered. (The traceless form arises naturally in a spherical harmonic multipole expansion; the other form arises in a different multipole expansion.) The quadrupole moment is a rank-2 tensor, which can be expressed as a 3 × 3 matrix, implying nine components. However, it is symmetric in its indices, which reduces the independent components to six, and the traceless constraint reduces the number of independent components to five.

as [15]

$$q_{00} = \frac{1}{\sqrt{4\pi}} \int \rho(\boldsymbol{x}')\, d^3x' = \frac{1}{\sqrt{4\pi}}\, Q, \tag{3.123a}$$

$$q_{11} = -\sqrt{\frac{3}{8\pi}} \int (x' - iy')\, \rho(\boldsymbol{x}')\, d^3x' = -\sqrt{\frac{3}{8\pi}}\, (p_x - ip_y), \tag{3.123b}$$

$$q_{10} = \sqrt{\frac{3}{4\pi}} \int z'\, \rho(\boldsymbol{x}') d^3x' = \sqrt{\frac{3}{4\pi}}\, p_z, \tag{3.123c}$$

$$q_{22} = \frac{1}{4}\sqrt{\frac{15}{2\pi}} \int (x' - iy')^2\, \rho(\boldsymbol{x}')\, d^3x' = \frac{1}{12}\sqrt{\frac{15}{2\pi}}\, (Q_{11} - 2iQ_{12} - Q_{22}), \tag{3.123d}$$

$$q_{21} = -\sqrt{\frac{15}{8\pi}} \int z'(x' - iy')\, \rho(\boldsymbol{x}')\, d^3x' = -\frac{1}{3}\sqrt{\frac{15}{8\pi}}\, (Q_{13} - iQ_{23}), \tag{3.123e}$$

$$q_{20} = \frac{1}{2}\sqrt{\frac{5}{4\pi}} \int (3z'^2 - r'^2)\, \rho(\boldsymbol{x}')\, d^3x' = \frac{1}{2}\sqrt{\frac{5}{4\pi}}\, Q_{33}, \tag{3.123f}$$

where corresponding moments with $m < 0$ may be obtained using $q_{l-m} = (-1)^m q^*_{lm}$. An expansion of $\Phi(\boldsymbol{x})$ in cartesian coordinates may be obtained by a direct Taylor series expansion of $1/\left|\boldsymbol{x} - \boldsymbol{x}'\right|$. The result

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \left[\frac{Q}{r} + \frac{\boldsymbol{p}\cdot\boldsymbol{x}}{r^3} + \frac{1}{2}\sum_{i,j} Q_{ij}\, \frac{x_i x_j}{r^5} + \cdots \right]. \tag{3.124}$$

is quoted without proof.[9]

Example 3.5 Consider a localized charge density

$$\rho(\boldsymbol{r}) = \frac{1}{64\pi} r^2 e^{-r} \sin^2\theta.$$

Let's make a multipole expansion of the potential associated with this charge density and determine all the non-vanishing multipole moments. The charge distribution is axially sym-

[9] The spherical multipole moments like Eq. (3.116) and the corresponding cartesian multipole moments like Eq. (3.121) generally differ in number of components for a given multipole order. For multipole order l there are $(l+1)(l+2)/2$ cartesian components but only $2l+1$ spherical components, which differs for $l > 1$. At a technical group-theoretical level the reason for this difference is that the spherical moments are *irreducible representations* of the rotation group (they are said to transform as irreducible *spherical tensors*) while the cartesian moments are *reducible* under rotational transformations. Physically this means that spherical multipole moments define components of definite angular momentum, while cartesian moments are mixtures of components with different angular momenta. Another technical point is that the coefficients in a multipole moment expansion may depend on choice of origin for the coefficients. In general it can be shown that the value of q_{lm} in Eq. (3.116) for the lowest-order non-vanishing multipole moment of a charge distribution is independent of origin for the coordinates, but higher-order moments will generally depend on the location of the origin for the coordinate system [15].

metric, so only Y_{lm} with $m=0$ are non-zero. From Eq. (3.116), the moments may be written

$$\begin{aligned}
q_{lm} &= \int Y^*_{l0}(\theta,\phi)r^l\rho(\boldsymbol{x})d^3x \\
&= \int Y^*_{l0}(\theta,\phi)r^l\rho(r,\theta)r^2 dr d\phi\, d(\cos\theta) \\
&= 2\pi\sqrt{\frac{2l+1}{4\pi}}\int P_l(\cos\theta)r^l\rho(r,\theta)r^2 dr\, d(\cos\theta) \\
&= \frac{2\pi}{64\pi}\sqrt{\frac{2l+1}{4\pi}}\int_0^\infty r^{l+4}e^{-r}dr\int_{-1}^{+1}P_l(\cos\theta)\sin^2\theta d(\cos\theta) \\
&= \frac{2\pi}{64\pi}\frac{2}{3}\sqrt{\frac{2l+1}{4\pi}}\int_0^\infty r^{l+4}e^{-r}dr\int_{-1}^{+1}P_l(\cos\theta)\left[P_0(\cos\theta)-P_2(\cos\theta)\right]d(\cos\theta) \\
&= \frac{1}{48}\sqrt{\frac{2l+1}{4\pi}}\Gamma(l+5)\left(2\delta_{l0}-\frac{2}{5}\delta_{l2}\right).
\end{aligned}$$

where we have used in the third line,

$$Y_{l0}(\theta,\phi)=\sqrt{\frac{2l+1}{4\pi}}P_l(\cos\theta)$$

and used in the fifth line,

$$\sin^2\theta = 1-\cos^2\theta = \frac{2}{3}\left[P_0(\cos\theta)-P_2(\cos\theta)\right]$$

and used in the last step

$$\int_{-1}^{+1}P_m(x)P_n(x)dx = \frac{2}{2n+1}\delta_{mn},$$

and the radial integral was evaluated using the tabulated definite integral

$$\int_0^\infty r^{n-1}e^{-(a+1)r}dr = \frac{\Gamma(n)}{(a+1)^n}.$$

Thus the multipole moments are

$$q_{lm} = \frac{1}{48}\sqrt{\frac{2l+1}{4\pi}}\Gamma(l+5)\left(2\delta_{l0}-\frac{2}{5}\delta_{l2}\right)$$

and the delta functions allow reading off the only non-vanishing multipole moments as

$$q_{00}=\sqrt{\frac{1}{4\pi}}Q \qquad q_{20}=-6\sqrt{\frac{5}{4\pi}}Q_{33}.$$

where the values were taken from Eq. (3.123).

Example 3.6 A spherical surface of radius R has a uniform surface charge of density $\sigma = Q/4\pi R^2$, except for a spherical cap at the north pole defined by a cone with opening $\theta=\alpha$ where $\sigma=0$, as illustrated in Fig. 3.15. Use the jump condition at a charge layer of Eq. (3.10) for the electric field

$$E_r^{\text{out}}|_{r=R} = E_r^{\text{in}}|_{r=R}+\frac{\sigma}{\varepsilon_0},$$

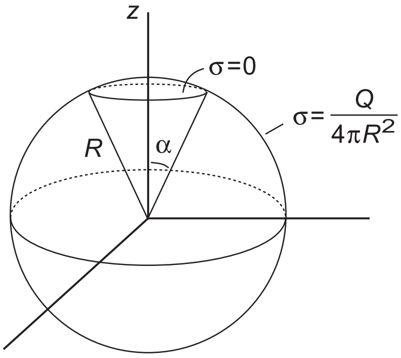


Fig. 3.15 Example 3.6: A spherical surface has a uniform surface charge of density $\sigma = Q/4\pi R^2$, except for a spherical cap at the north pole defined by a cone with opening $\theta = \alpha$ where $\sigma = 0$.

to show that the potential inside the spherical surface can be expressed as

$$\Phi = \frac{Q}{8\pi\varepsilon_0}\sum_{l=0}^{\infty}\frac{1}{2l+1}\left[P_{l+1}(\cos\alpha) - P_{l-1}(\cos\alpha)\right]\frac{r^l}{R^{l+1}}P_l(\cos\theta).$$

What is the potential outside the sphere?

The surface charge density specifies a jump condition on the normal component of the electric field (see Section 3.3),

$$E_r^{\text{out}}|_{r=R} = E_r^{\text{in}}|_{r=R} + \frac{\sigma}{\varepsilon_0},$$

which allows us to solve for the potential $\Phi(r,\theta)$. Because of the axial symmetry about the z axis we may expand the potential in Legendre polynomials according to Eqs. (3.110) and (3.111),

$$\Phi_{\text{in}} = \sum_{l=0}^{\infty} A_l \left(\frac{r}{R}\right)^l P_l(\cos\theta) \qquad \Phi_{\text{out}} = \sum_{l=0}^{\infty} A_l \left(\frac{R}{r}\right)^l P_l(\cos\theta),$$

where the expansion coefficients A_l are the same for Φ_{in} and Φ_{out} because we require Φ to be continuous at the surface $r = R$. The radial components of the interior and exterior electric fields follow from $E_r = -\partial\Phi/\partial r$,

$$E_r^{\text{in}} = -\sum_{l=0}^{\infty}\frac{lA_l}{R}\left(\frac{r}{R}\right)^{l-1}P_l(\cos\theta),$$

$$E_r^{\text{out}} = \sum_{l=0}^{\infty}\frac{(l+1)A_l}{R}\left(\frac{R}{r}\right)^{l+2}P_l(\cos\theta).$$

Substituting this into the jump condition given above for E_r leads to

$$\sigma(\cos\theta) = \varepsilon_0\left[E_r^{\text{out}} - E_r^{\text{in}}\right]_{r=R} = \sum_{l=0}^{\infty}\frac{(2l+1)\varepsilon_0 A_l}{R}P_l(\cos\theta).$$

Multiply both sides by $P_k(\cos\theta)$, integrate over $d(\cos\theta)$, and use

$$\int_{-1}^{+1} P_n(x)P_m(x)dx = \frac{2}{2n+1}\,\delta_{nm}$$

to give

$$\frac{(2l+1)\varepsilon_0 A_l}{R} = \frac{2l+1}{2}\int_{-1}^{+1}\sigma(\cos\theta)P_l(\cos\theta)d(\cos\theta),$$

implying that

$$A_l = \frac{R}{2\varepsilon_0}\int_{-1}^{+1}\sigma(\cos\theta)P_l(\cos\theta)d(\cos\theta).$$

Then using that the surface is covered uniformly with charge except within the cone,

$$\sigma(\cos\theta) = \begin{cases} \dfrac{Q}{4\pi R^2} & (\cos\theta < \cos\alpha), \\ 0 & (\cos\theta > \cos\alpha), \end{cases}$$

leads to

$$A_l = \frac{Q}{8\pi\varepsilon_0 R}\int_{-1}^{\cos\alpha} P_l(\cos\theta)d(\cos\theta).$$

This can be integrated by using [1]

$$P_l(x) = \frac{1}{2l+1}\left[P'_{l+1}(x) - P'_{l-1}(x)\right]$$

(where the primes indicate derivatives), which gives

$$\begin{aligned} A_l &= \frac{Q}{8\pi\varepsilon_0 R}\int_{-1}^{\cos\alpha} P_l(\cos\theta)d(\cos\theta) \\ &= \frac{Q}{8\pi\varepsilon_0 R}\frac{1}{2l+1}\int_{-1}^{\cos\alpha}\left[\frac{dP_{l+1}(\cos\theta)}{d(\cos\theta)} - \frac{dP_{l-1}(\cos\theta)}{d(\cos\theta)}\right]d(\cos\theta) \\ &= \frac{Q}{8\pi\varepsilon_0 R}\frac{1}{2l+1}\int_{-1}^{\cos\alpha}\left[dP_{l+1}(\cos\theta) - dP_{l-1}(\cos\theta)\right] \\ &= \frac{Q}{8\pi\varepsilon_0 R}\frac{1}{2l+1}\left[P_{l+1}(\cos\theta) - P_{l-1}(\cos\theta)\right]_{-1}^{\cos\alpha} \\ &= \frac{Q}{8\pi\varepsilon_0 R}\frac{1}{2l+1}\left[P_{l+1}(\cos\alpha) - P_{l-1}(\cos\alpha)\right], \end{aligned}$$

where in the last step $P_l(-1) = (-1)^l$ was used. Substituting in the original expansions, for the inside solution,

$$\begin{aligned} \Phi_{\text{in}} &= \sum_{l=0}^{\infty} A_l\left(\frac{r}{R}\right)^l P_l(\cos\theta) \\ &= \frac{Q}{8\pi\varepsilon_0}\sum_{l=0}^{\infty}\frac{1}{2l+1}\left[P_{l+1}(\cos\alpha) - P_{l-1}(\cos\alpha)\right]\frac{r^l}{R^{l+1}}P_l(\cos\theta). \end{aligned}$$

and for the outside solution,

$$\begin{aligned}\Phi_{\text{out}} &= \sum_{l=0}^{\infty} A_l \left(\frac{R}{r}\right)^l P_l(\cos\theta) \\ &= \frac{Q}{8\pi\varepsilon_0} \sum_{l=0}^{\infty} \frac{1}{2l+1} \left[P_{l+1}(\cos\alpha) - P_{l-1}(\cos\alpha)\right] \frac{R^{l-1}}{r^l} P_l(\cos\theta).\end{aligned}$$

3.10.4 Multipole Components of the Electric Field

We have expanded the potential in multipole moments so the corresponding electric fields can be expanded in a similar way. It is easiest to express the components of the electric field $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ in spherical coordinates. For a term in Eq. (3.115) with definite (l, m) the spherical electric field components are [15]

$$E_r = \frac{l+1}{(2l+1)\varepsilon_0} q_{lm} \frac{1}{r^{l+2}} Y_{lm}(\theta,\phi), \tag{3.125a}$$

$$E_\theta = -\frac{1}{(2l+1)\varepsilon_0} q_{lm} \frac{1}{r^{l+2}} \frac{\partial}{\partial\theta} Y_{lm}(\theta,\phi), \tag{3.125b}$$

$$E_\phi = \frac{1}{(2l+1)\varepsilon_0} q_{lm} \frac{1}{r^{l+2}} \frac{im}{\sin\theta} Y_{lm}(\theta,\phi), \tag{3.125c}$$

with the multipole moments q_{lm} defined in Eq. (3.123).

Example 3.7 For a dipole $\boldsymbol{p}$ oriented along the z axis, one finds

$$E_r = \frac{2p\cos\theta}{4\pi\varepsilon_0 r^3} \qquad E_\theta = \frac{p\sin\theta}{4\pi\varepsilon_0 r^3} \qquad E_\phi = 0,$$

for the electric-field components (3.125).

3.10.5 Energy of a Charge Distribution in an External Field

If a localized charge distribution $\rho(\boldsymbol{x})$ is subject to an *external potential* $\Phi(\boldsymbol{x})$,[10] the electrostatic energy is

$$W = \int \rho(\boldsymbol{x})\Phi(\boldsymbol{x}) d^3x. \tag{3.126}$$

[10] For example, the charge distribution of an atomic nucleus subject to an external electric field generated by the electrons of that atom, or the charge distribution of an atomic nucleus subject to the external electric field of another nucleus in a collision between the two nuclei.

If the potential varies slowly over the extent of $\rho(\boldsymbol{x})$, it can be expanded in a Taylor series,

$$\begin{aligned}\Phi(\boldsymbol{x}) &= \Phi(0) + \boldsymbol{x}\cdot\boldsymbol{\nabla}\Phi(0) + \frac{1}{2}\sum_i\sum_j x_i x_j \frac{\partial^2\Phi}{\partial x_i \partial x_j}(0) + \cdots \\ &= \Phi(0) - \boldsymbol{x}\cdot\boldsymbol{E}(0) - \frac{1}{2}\sum_i\sum_j x_i x_j \frac{\partial E_j}{\partial x_i}(0) + \cdots \\ &= \Phi(0) - \boldsymbol{x}\cdot\boldsymbol{E}(0) - \frac{1}{6}\sum_i\sum_j \left(3x_i x_j - r^2\delta_{ij}\right)\frac{\partial E_j}{\partial x_i}(0) + \cdots,\end{aligned} \tag{3.127}$$

where in line 2 the definition of the electric field $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ was used and in the last line $\frac{1}{6}r^2\boldsymbol{\nabla}\cdot\boldsymbol{E}(0)$ was subtracted from the last term since $\boldsymbol{\nabla}\cdot\boldsymbol{E} = 0$ for the external field. Inserting the expansion (3.127) in Eq. (3.126), the energy takes the form

$$W = q\Phi(0) - \boldsymbol{p}\cdot\boldsymbol{E}(0) - \frac{1}{6}\sum_i\sum_j Q_{ij}\frac{\partial E_j}{\partial x_i}(0) + \cdots, \tag{3.128}$$

where q is the total charge, the dipole moment $\boldsymbol{p}$ is defined in Eq. (3.121), and the quadrupole moment Q_{ij} is defined in Eq. (3.122).

Example 3.8 Many atomic nuclei have charge distributions exhibiting a quadrupole deformation. Such nuclei will have a contribution to the energy from the quadrupole term in the expansion (3.128) if they are subject to an external electric field. Such an "external" field can be provided by the electrons of the atom containing the nucleus, or by a crystal lattice in which the nucleus is embedded, and coupling to these external field leads to small energy shifts and breaking of degeneracies for nuclear states that can be detected experimentally. (The energy shifts are small and radiofrequency measurements are typically used.) Such methods allow the quadrupole moments of nuclei to be measured, which are important clues to the details of nuclear structure and interactions. In collisions between heavy ions at nuclear accelerators, the electric field of one nucleus can cause excited states to be populated in the other nucleus, in a process called *Coulomb excitation*. The study of the rates at which those excited states are populated (for example by detecting the de-excitation by emission of γ-rays) are another way in which nuclear quadrupole deformations can be measured by analyzing their response to an applied electric field.

The expansion (3.128) manifests the characteristic way in which various multipoles of a charge distribution interact with an external electric field:

1. the charge interacts with the potential in the first term,
2. the dipole moment interacts with the electric field in the second term,
3. the quadrupole moment interacts with the gradient of the electric field in the third term, and so on.

Thus multipole expansions serve a pedagogical as well as practical purpose in understanding electrostatic interactions.

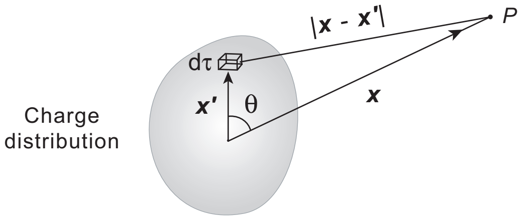
Background and Further Reading

An introduction to the material of this chapter may be found in Griffiths [8] or Purcell and Morin [23]. More advanced treatments may be found in Jackson [15], Garg [6], Chaichian et al [3], and Zangwill [27].

Problems

3.1 A charge q is placed outside a grounded ($\Phi = 0$) conducting sphere. Use the method of images described in Section 3.8 to find the scalar potential Φ in the space outside the sphere, and calculate the force exerted by the sphere on the charge q.

3.2 Assuming that $|\boldsymbol{x}| \gg |\boldsymbol{x}'|$ in the following figure,



prove that $1/|\boldsymbol{x}-\boldsymbol{x}'|$ can be expanded in the power series

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \frac{1}{r}\left(1 - \frac{1}{2}\varepsilon + \frac{3}{8}\varepsilon^2 - \frac{5}{16}\varepsilon^3 + \cdots\right),$$

where $r = |\boldsymbol{x}|$ and $r' = |\boldsymbol{x}'|$, and

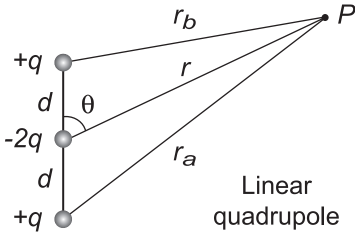
$$\varepsilon \equiv \frac{r'}{r}\left(\frac{r'}{r} - 2\cos\theta\right),$$

and that this is a series expansion in terms of Legendre polynomials that is equivalent to Eq. (3.114). Finally, show that

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\sum_{l=0}^{\infty} r^{-(l+1)} \int (r')^l P_l(\cos\theta)\rho(\boldsymbol{x}')d^3x'$$

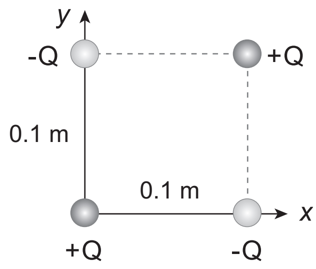
is the scalar potential at the point P.

3.3 The following charge distribution is called a *linear quadrupole*.



What is the potential generated at the point P assuming that $r \gg d$?

3.4 For the following charge distribution,



calculate the components of the dipole moment vector $\boldsymbol{p}$ assuming that $|Q| = 3\,\mu$C.

3.5 A surface charge density specified by a function $\sigma(\theta)$ is pasted onto an empty 3D spherical shell of radius R. Assume axial symmetry and use separation of variables in the Laplace equation to derive a general formula for the potential Φ as a function of $\sigma(\theta)$ inside and outside the shell radius R.

3.6 A point dipole with dipole moment $\boldsymbol{p}$ is located at $\boldsymbol{x}_0$. Show that for calculations of potential Φ or energy density W of a dipole in an external field, the dipole can be described by an effective charge density

$$\rho_{\text{eff}}(\boldsymbol{x}) = -\boldsymbol{p}\cdot\boldsymbol{\nabla}\delta(\boldsymbol{x}-\boldsymbol{x}_0),$$

where $\delta(\boldsymbol{x}-\boldsymbol{x}')$ is the 3D Dirac delta function.

3.7 The quadrupole moment tensor is given by Eq. (3.122) as

$$Q_{ij} \equiv \int (3x'_i x'_j - r'^2\delta_{ij})\,\rho(\boldsymbol{x}')d^3x',$$

where $r'^2 \equiv |\boldsymbol{x}'|^2$. Evaluate Q_{ij} for a discrete set of N static charges q_i and show that it is traceless.

3.8 For the discrete charge distribution displayed in Problem 3.4, calculate the nine *cartesian components* $Q_{ij} = Q_{xx}, Q_{xy}, \cdots$ of the quadruple moment tensor starting from Eq. (3.122). Assume the origin of the coordinate system to be at the lower left charge and that $|Q| = 3\,\mu$C.

4 Electrostatics in Dielectric Matter

To this point we have considered electrostatics primarily in vacuum, except for the presence of electrical charges, and of conducting matter in a few instances. But many important applications of electromagnetic theory involve interactions in non-conducting (dielectric) matter. There are two fundamental differences between conductors and dielectrics, and their electrostatic behavior.

1. Conductors have available many electrons that are not bound to atoms or molecules, and thus are very mobile. In constrast, ideal dielectrics have no free electrons.
2. Dielectrics typically can have electric fields in their interior, but conductors suppress any internal electric fields.

Therefore, in this chapter we begin to address how the properties of dielectric matter influence the equations of electrostatics. We will find that electric fields in matter are largely dipole fields, with the the net dipole moment having two basic sources:

1. Some molecules have an *intrinsic dipole moment*, and an external electric field exerts a torque that tends to align those moments with the field.
2. An electric field produces dipoles by polarizing matter, even if the atoms or molecules have no significant dipole moment in the absence of the applied field. This polarization effect is characterized by a quantity called the *atomic polarizability*.

In either case the material my be characterized in terms of a *polarization P* and an *electric susceptibility*, which is proportional to the ratio of P to the electric field. The primary effect of the polarization is to create a surface charge density in dielectric material. A considerable amount can be learned about the nature of dielectrics by examining the effect of this induced surface charge density on capacitors.

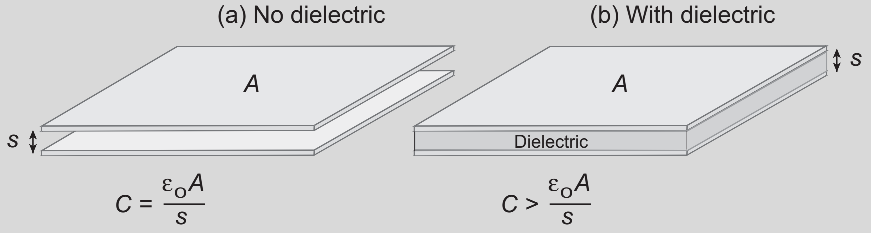
4.1 Dielectrics

As illustrated in Box 4.1, a capacitor with dielectric material between the metal plates has increased capacitance because of this induced surface charge (with the practical implications of reducing the size of the capacitor, or increasing its working voltage). Let us investigate in more depth the effect of a dielectric on a capacitor. A parallel-plate capacitor with no material between the plates has a capacitance C defined by

$$C = \frac{Q}{\Phi_{12}} = \frac{\varepsilon_0 A}{s}, \tag{4.1}$$

Box 4.1 **Capacitors and Dielectrics**

The simplest parallel-plate capacitor has two separated parallel metal plates with nothing in between, as illustrated in Fig. (a) below.


(a) No dielectric
(b) With dielectric
A
A
s
s
Dielectric
$C = \frac{\varepsilon_0 A}{s}$
$C > \frac{\varepsilon_0 A}{s}$

Inserting a dielectric between the plates of the capacitor as in Fig. (b) increases the capacitance because the induced surface charge on the dielectric produced by the electric field between the plates partially cancels the opposite charge on the adjacent plate. Understanding this mechanism leads to fundamental insight into the role that an electric field plays in a dielectric.

where Q is the magnitude of the charge on the plates (positive Q on one plate and negative Q on the other), Φ_{12} is the difference in electrical potential between the two plates, A is the surface area of a plate, and s is the separation of the plates.

4.1.1 Capacitors and Dielectrics

Now suppose that a layer of dielectric material is placed between the capacitor plates, as illustrated in Box 4.1. One will generally find that the capacitance can still be defined by the ratio of charge to potential difference between the plates, $C = Q/\Phi_{12}$, but the actual value of C will be *increased* over that found with no dielectric between the plates.

> This implies that the presence of the dielectric between the plates allows more charge Q on the plates and therefore greater capacitance, for the same potential difference, plate area, and separation of the plates.

Qualitatively, this influence of the dielectric on the capacitance is not difficult to understand. The material of the dielectric consists of atoms or molecules with negatively charged electrons and positively charged nuclei.

1. The electric field between the plates polarizes the charge distribution of the dielectric.
2. If we assume to be definite that the upper plate has a positive charge and the lower plate a negative charge, negative charges in the dielectric will be pulled upward and positive charges pushed down, as illustrated in Fig. 4.1.

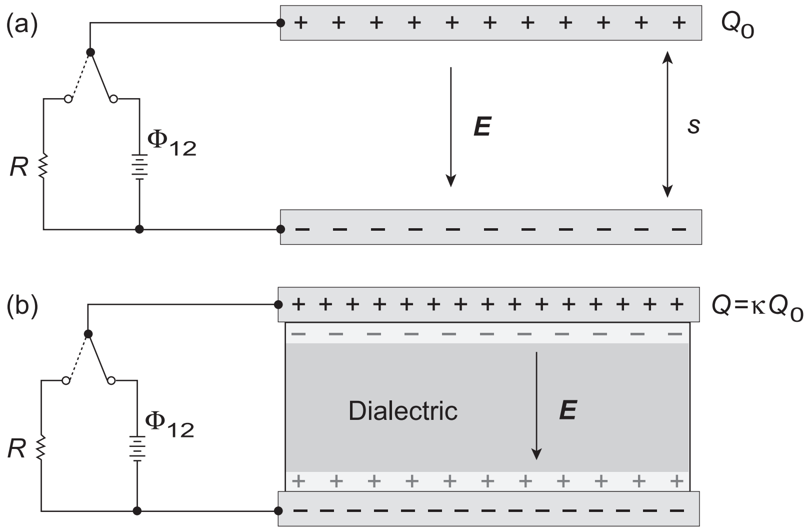


Fig. 4.1 Increase of capacitance by insertion of a dielectric between the plates. (a) No dielectric. (b) With dielectric. Adapted from Ref. [23].

3. This exposes a layer of uncompensated negative charge near the top of the dielectric and a corresponding layer of uncompensated positive charge near the bottom of the dielectric.[1]
4. The charge Q on the upper plate will increase because of the induced top layer of negative charge below it in the dielectric.
5. We shall show later that Q must increase until the algebraic sum of Q and the induced charge layer is equal to the total charge on the top plate Q_0 before the dielectric layer was inserted.
6. Thus, the total charge Q in the top layer is *larger than the charge* Q_0 of the top plate in Fig. 4.1(a) before the dielectric was inserted.
7. Thus, the charge is the Q that appears in Eq. (4.1) and is, in the circuit of Fig. 4.1, the charge supplied by the battery to charge the capacitor while the switch is in the right (charging) position.

If the switch in the circuit of Fig. 4.1 were flipped to the left to discharge the capacitor though the resistance R after the capacitor is fully charged, the charge Q would be dissipated. Notably, *the induced charge layer is not part of* Q, and the induced charge would be absorbed back into the normal structure of the dielectric in the absence of an electric field.

[1] As will be quantified shortly, this displacement of charge is very small. Remember that we are dealing with a dielectric where the electrons are bound up in atoms and molecules. There is no significant population of free charge carriers as would be the case with a metallic conductor.

Table 4.1 dielectric constants κ of some substances

Substance	Conditions	κ
Vacuum		1.00000
Air	gas, 0° C, 1 atm	1.00059
Water vapor	gas, 110° C, 1 atm	1.0126
Liquid water	liquid, 20° C	80.4
Silicon	solid 20° C	11.7
Polyethelene	solid, 20° C	$2.25-2.3$
Porcelain	solid 20° C	$6.0-8.0$

4.1.2 Dielectric Constants

Different dielectric materials would be expected to have different efficiencies for increasing the charge capacity of a capacitor, according to the ease with which electrons can be displaced with respect to the atomic nuclei by the applied electric field. The factor Q/Q_0 by which the charge and thus the capacitance is increased by inserting a particular material between the capacitor plates is called the *dielectric constant* κ of the material

$$Q = \kappa Q_0 \quad \leftrightarrow \quad C = \kappa C_0. \tag{4.2}$$

Dielectric constants are ratios of charges and thus are dimensionless; a few are given in Table 4.1 for some representative substances. Note that the dielectric constant of the vacuum is $\kappa = 1$ exacly (by definition), typical gases have dielectric constants slightly larger than one, and liquids and solids can have dielectric constants varying widely in the range $\kappa \sim 1-100$. The reason for the remarkably large value $\kappa = 80.4$ for water merits an explanation that will be given later.

4.1.3 Bound Charge and Free Charge

In considering the effect of a dielectric on a capacitor, it is useful to introduce some terminology distinguishing between the charge associated with the dielectric itself and the mobile charges that can charge the plates. Those charges associated with the dielectric are termed *bound charges*. They are not mobile because they are attached to the atoms and molecules making up the dielectric. They can be polarized through electric fields causing tiny displacements, but if the capacitor is discharged the bound charges remain with the dielectric, which becomes unpolarized as the electric field vanishes (assuming no permanent dipole moment for the substance of the dielectric); they are *not part of the charge Q on the capacitor plates*. On the other hand, those charges that are not bound in the dielectrics are termed *free charges*. They are the charges that we have some agency over in an experiment (for example, through manipulation of the simple electrical circuit in Fig. 4.1 using the switch to charge or discharge the capacitor).

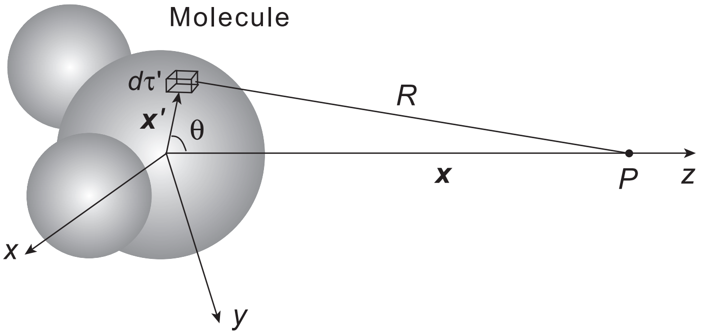


Fig. 4.2 Diagram for calculation of the potential at point P of a molecular charge distribution.

4.2 Moments of a Molecular Charge Distribution

Let's consider the potential at a distant point P due to an electrical charge distribution of a molecule, as illustrated in Fig. 4.2. which will be of the form

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{R}\,d\tau', \tag{4.3}$$

where the integral is over all of the charge distribution. From the law of cosines the distance R is given by

$$R = (r^2 + r'^2 - 2rr'\cos\theta)^{1/2}, \tag{4.4}$$

where we define $r = |\boldsymbol{x}|$ and $r' = |\boldsymbol{x}'|$, and Eq. (4.3) becomes

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\int (r^2 + r'^2 - 2rr'\cos\theta)^{-1/2}\rho(\boldsymbol{x}')\,d\tau'. \tag{4.5}$$

For a distant point P we have $r \gg r'$ and we can expand to give

$$\frac{1}{R} = (r^2 + r'^2 - 2rr'\cos\theta)^{-1/2} = \frac{1}{r}\left[1 + \frac{r'}{r}\cos\theta \right. \\ \left. + \left(\frac{r'}{r}\right)^2 \frac{3\cos^2\theta - 1}{2} + \mathscr{O}\left(\left[\frac{r'}{r}\right]^3\right)\right] \tag{4.6}$$

Then from Eqs. (4.3) and (4.6) the potential can be written as

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\left[\frac{1}{r}\int \rho\,d\tau' + \frac{1}{r^2}\int r'\cos\theta\rho\,d\tau' + \frac{1}{r^3}\int r'^2\frac{3\cos^2\theta - 1}{2}\rho\,d\tau' + \cdots\right], \tag{4.7}$$

where $r = |\boldsymbol{x}|$ is a constant and has been brought outside the integrals (the integration variable is $\boldsymbol{x}'$). Now this is a power series in $1/r$ with coefficients that are constants depending only on integrals over the charge distribution and independent of the distance to P. Thus we can write the potential as a power series (multipole expansion) with constant coefficients

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\left[\frac{C_0}{r} + \frac{C_1}{r^2} + \frac{C_2}{r^3} + \cdots\right], \tag{4.8}$$

where the coefficients C_i are the integrals over the internal charge distribution in Eq. (4.7). The electric field then follows from $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi_P$. This result is valid only along the z axis, but the power series (4.8) illustrates the central point:

> The behavior of the potential at large distance from the source will be dominated by the first term in the multipole expansion (4.8) that has a non-zero coefficient.

Let us look at the coefficients in Eq. (4.8) more closely. The monopole coefficient $C_0 = \int \rho(\boldsymbol{x}')d\tau'$ is the total charge. If we have a neutral atom or molecule, $C_0 = 0$. If it is not neutral (ionized, for example) so that $\rho \neq 0$, the monopole will always dominate at large enough distance. If the atom or molecule is charge-neutral so that $\rho = 0$, the dipole term with $C_1 = \int r' \cos\theta \rho(\boldsymbol{x}')d\tau'$ dominates. Furthermore, if the charge distribution is charge-neutral the value of C_1 is independent of the choice of origin.[2]

> As we shall see, for our main task here of understanding the behavior of dielectrics only the monopole strength (the total charge) and the dipole strength of the molecular building blocks of the dielectric are important in determining its electric-field properties.

Thus, for the properties of dielectrics we will find that all multipole moments of the charge distribution of order greater than the dipole can usually be ignored. As noted in the introduction to this chapter, a net dipole moment can come about because of induced polarization by an electric field, or because of molecules that have a permanent dipole moment. We address induced moments in Section 4.3 and permanent dipole moments in Section 4.4.

4.3 Induced Dipole Moments

The simplest atom is hydrogen, consisting of one nucleus and one electron. The nucleus is so small compared with the electron cloud that in hydrogen and in more complicated atoms and molecules we can approximate the nuclei as point charges. Quantum mechanically the electron in hydrogen must be viewed as a cloud of negative charge with smoothly varying density (with the integrated charge equal to the electron charge e). The density falls off exponentially on the boundaries, so it makes sense to view the charge clouds of atoms and molecules as having approximate radii and shapes.

The electron cloud of the undisturbed hydrogen atom is spherically symmetric but if it is placed in an electric field pointing upward along a z-axis, the charge cloud of the atom will be distorted, with the negative electron charge cloud pulled down and the nucleus pushed

[2] Generally the value of a multipole moment depends on the origin chosen for the coordinate system. But for the special case of a dipole moment in a charge-neutral dielectric, changing the origin will not change the value of the dipole moment.

up. This distorted atom will have an electric dipole moment because the center of mass of the electron cloud will be displaced a small amount Δz from the nucleus, giving a net dipole moment of magnitude $\sim e\Delta z$, where e is the total electron charge. Example 4.1 estimates the size of the distortion in actual atoms.

Example 4.1 We may make a very rough estimate of how much distortion will be caused by a field of strength E by noting that a strong electric field already exists holding the unperturbed H atom together, which may be estimated as

$$E \simeq \frac{e}{4\pi\varepsilon_0 a^2},$$

where a is a characteristic atomic length scale (for example, some multiple of the Bohr radius). If we assume that a field of the same order of magnitude would be required to produce significant distortion, we may estimate the distortion as

$$\frac{\Delta z}{a} \simeq \frac{E}{e/4\pi\varepsilon_0 a^2}.$$

Typically, $e/4\pi\varepsilon_0 a^2 \sim 10^{11}$ volts/m, which is enormous (thousands of times larger than any field produced in a laboratory); the distortion of the atom $\Delta z/a$ will be tiny indeed!

The dipole moment vector $\boldsymbol{p}$ induced by an electric field will point in the direction of $\boldsymbol{E}$. The induced dipole moment is generally proportional to the electric field, at least if the field is not too strong. The factor relating the dipole moment vector $\boldsymbol{p}$ to $\boldsymbol{E}$ is the *atomic polarizability* α,

$$\boldsymbol{p} = \alpha \boldsymbol{E}. \tag{4.9}$$

More generally, the scalar coefficient α in this equation must be replaced by a polarizability tensor It is common to report atomic polarizabilities as the quantity $\alpha/4\pi\varepsilon_0$ (which has units of volume) rather than as α itself.[3] Some atomic polarizabilities are given in Table 4.2, listed in order of increasing total electron number.

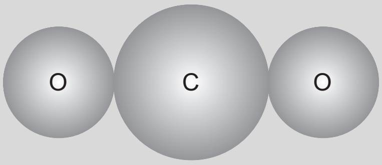
The large variations in polarizabilities seen in Table 4.2 may be attributed to differences in electron number and to differences in valence electronic structure. For example, the noble gas elements He, Ne, and Ar have relatively small polarizabilities because their outer electrons are more tightly bound than for other elements, but the polarizabilities increase in the sequence He to Ne to Ar because the total electron number is increasing. As another example, the elements Na and K each have one electron outside a closed electronic shell and that electron is susceptible to perturbation by an electric field. As a result Na and K have very large polarizabilities ($\kappa = 27$ and $\kappa = 34$, respectively).

Example 4.2 Let's estimate the charge displacement induced by a typical electric field. From the polarizability of atomic hydrogen in Table 4.2, the magnitude of the dipole mo-

[3] Beware: Both α and $\alpha/4\pi\varepsilon_0$ are sometimes called the atomic polarizability in the literature.

Box 4.2 **Asymmetric Polarization**

Molecules can be very asymmetric and may have different polarizabilities along different axes. For example, the linear molecule carbon dioxide (CO_2) illustrated in the following figure



has a polarizability more than twice as large if the field is applied along its long axis than if it is applied perpendicular to that. In the most general case of a highly asymmetric molecule the simple relation (4.9) between the polarization vector and the electric field vector must be replaced by a tensor equation that can be expressed as a matrix–vector multiply,

$$\boldsymbol{p} = \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix},$$

which is equivalent to the simultaneous equations

$$\begin{aligned} p_x &= \alpha_{xx}E_x + \alpha_{xy}E_y + \alpha_{xz}E_z, \\ p_y &= \alpha_{yx}E_x + \alpha_{yy}E_y + \alpha_{yz}E_z, \\ p_z &= \alpha_{zx}E_x + \alpha_{zy}E_y + \alpha_{zz}E_z. \end{aligned}$$

Thus the scalar coefficient α in Eq. (4.9) has been replaced by a polarizability tensor $\mathscr{A}$ that has the components

$$\mathscr{A} = \begin{pmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{pmatrix},$$

expressed as a matrix in the current basis.

ment induced by an electric field of one megavolt / meter is $p < 10^{-34}$ coulomb-meters.

4.4 Permanent Dipole Moments

Some molecules have asymmetric shapes and permanent dipole moments in their normal ground state, even in the absence of an external field; a few examples are shown in Fig. 4.3.

Table 4.2 Atomic polarizabilities $(\alpha/4\pi\varepsilon_0)$ in units of 10^{-30} m^3

Element:	H	He	Li	Be	C	Ne	Na	Ar	K
	0.66	0.21	12	9.3	1.5	0.4	27	1.6	34

Molecules with a permanent dipole moment are termed *polar molecules*. As a rule, intrinsic dipole moments (when they exist) are much larger than dipole moments induced by laboratory electric fields.

Example 4.3 As shown in Fig. 4.3, the magnitude of the (permanent) dipole moment for the molecule HCl is $p = 3.43 \times 10^{-30}$ coulomb-meters (which is equivalent to shifting one electron by a distance of 0.2 angstrom (0.2×10^{-8} cm) [23]. From Example 4.2, the magnitude of the dipole moment in hydrogen induced by an electric field of magnitude one megavolt per meter is $p < 10^{-34}$ coulomb-meters. This result is a characteristic one and permanent dipole moments (when they exist) are typically orders of magnitude larger than those induced by laboratory electric fields.

In a uniform electric field polar molecules will have their dipole vectors partially aligned with the field. The net force on the dipole vanishes, because the force on the negative end exactly cancels the force on the positive end (see Fig. 4.4), but the torque $\boldsymbol{N}$ acting on the dipole is

$$\boldsymbol{N} = \boldsymbol{p} \times \boldsymbol{E}, \tag{4.10}$$

where $\boldsymbol{p}$ is the dipole moment vector. The direction of $\boldsymbol{N}$ is so as to align the dipole moment of a polar molecule with the electric field. Perturbations such as thermal fluctuations tend to inhibit this alignment, so the typical physical outcome is an equilbrium with the dipoles partially aligned.

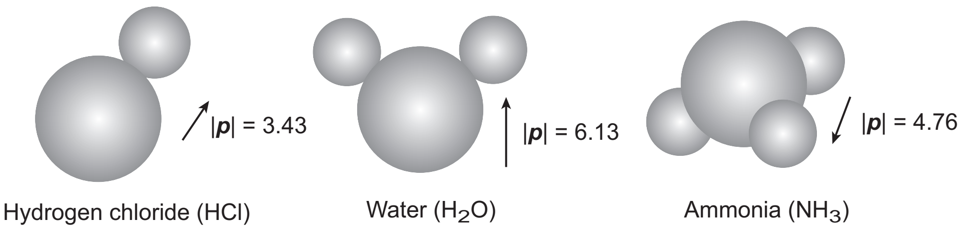

$|\boldsymbol{p}|$ = 3.43
$|\boldsymbol{p}|$ = 6.13
$|\boldsymbol{p}|$ = 4.76
Hydrogen chloride (HCl)
Water (H_2O)
Ammonia (NH_3)

Fig. 4.3

Approximate geometry of the electron cloud and the observed dipole moment vector $\boldsymbol{p}$ for some polar molecules. Magnitudes of the dipole moment vector are in units of 10^{-30} coulomb-meters (adapted from Ref. [23]).

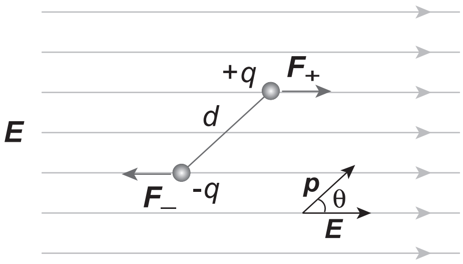


Fig. 4.4 Torque applied to a molecule with an intrinsic dipole moment in a uniform electric field.

4.5 Polarization

We have discussed two basic mechanism that lead to polarization of a dielectric: (1) distortion of the charge distribution by an electric field, which induces many tiny dipoles pointing in the same direction as the field if the dielectric substance consists of atoms or non-polar molecules, and (2) alignment of the existing intrinsic dipole moments by an electric field if the substance consists of polar molecules. The net effect of either is to produce a set of dipoles aligned or partially aligned with the electric field. At this point our primary interest is in the effect of this polarization, without regard to the mechanism by which it was formed. Therefore, a measure of how polarized the matter is can be formed by asking how many dipoles N of average dipole moment $\boldsymbol{p}$ there are in a unit volume, without regard to the source of the dipoles. The total dipole strength of an infinitesimal volume element is then $\boldsymbol{p}Nd\tau$ and the *polarization density* $\boldsymbol{P}$ can be defined by

$$\boldsymbol{P} \equiv \boldsymbol{p}N = \left(\frac{\text{dipole moments}}{\text{unit volume}}\right), \tag{4.11}$$

which has units of $\text{C-m/m}^3 = \text{C/m}^2$. The *polarization-charge density* ρ_{pol} is given by minus the divergence of the polarization,

$$\rho_{\text{pol}}(\boldsymbol{x}) = -\boldsymbol{\nabla}\cdot\boldsymbol{P}(\boldsymbol{x}). \tag{4.12}$$

Let's now estimate the the potential produced by the polarized dielectric material, considering separately the regions outside and inside the dielectric matter.

4.6 Field Outside Polarized Dielectric Matter

If it is assumed that the dielectric was assembled from neutral matter so that there is no net charge, there is no monopole term in the multipole expansion (4.8). Therefore to very good approximation only the dipole moments need be considered as sources of a field. Consider a thin vertical cylinder of dielectric matter in an electric field, as illustrated in Fig. 4.5 [23]. The polarization $\boldsymbol{P}$ is uniform and points in the positive z direction, and we

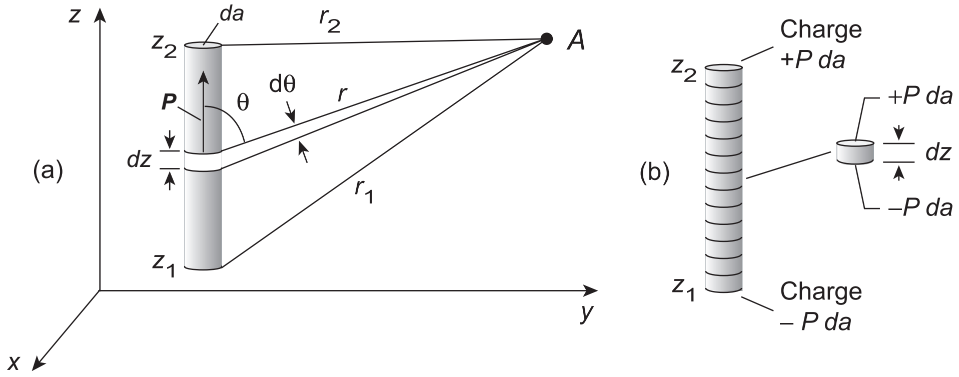


Fig. 4.5 (a) A column of polarized material with cross section da observed at a distant point A produces the same field as two charges, one at each end of (b). Figure adapted from Ref. [23].

wish to calculate the electric potential at the point A. An element of the cylinder of height dz in Fig. 4.5(a) has a dipole moment

$$\boldsymbol{p} = \boldsymbol{P} d\tau = \boldsymbol{P}\, da\, dz, \tag{4.13}$$

and its contribution to the potential at point A is

$$d\Phi_A = \frac{P\cos\theta\, da\, dz}{4\pi\varepsilon_0 r^2}. \tag{4.14}$$

Then the potential at A produced by the entire column of polarized matter is obtained by integration,

$$\begin{aligned}\Phi_A &= \frac{P\,da}{4\pi\varepsilon_0}\int_{z_1}^{z_2}\frac{dz\cos\theta}{r^2}\\ &= \frac{-P\,da}{4\pi\varepsilon_0}\int_{z_1}^{z_2}\frac{dr}{r^2}\\ &= \frac{Pda}{4\pi\varepsilon_0}\left(\frac{1}{r_2}-\frac{1}{r_1}\right).\end{aligned} \tag{4.15}$$

But this is exactly the same result for the potential at A that would be produced by a positive point charge of magnitude Pda at the top of the column at a distance r_2 from A and a negative point charge of the same magnitude placed at the bottom of the column a distance r_1 from A (see the dipole potential example in Box 3.2).

A column of uniformly polarized matter produces the same potential and thus the same electric field at an external point A as a dipole consisting of two concentrated charges of magnitude Pda at the two ends of the column.

We can make Eq. (4.15) plausible by a more heuristic argument illustrated in Fig. 4.5(b). Consider making the cylinder shown in Fig. 4.5(a) by stacking on top of each other small

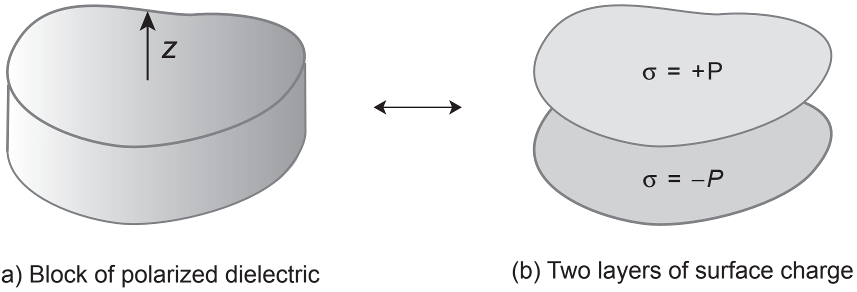

z
σ = +P
σ = −P
(a) Block of polarized dielectric
(b) Two layers of surface charge

Fig. 4.6 The external field due to (a) a block of dielectric polarized by an electric field in the z direction is the same as would be produced by (b) two sheets of charge $\pm P$ at the positions of the top and bottom surfaces of the block.

cylinder segments of height dz, with a charge of $+P\,da$ on its top face and $-P\,da$ on its bottom face. But now within the column the $+$ charge on the top of a segment will be cancelled by the $-$ charge on the bottom of the next segment up, except for the top-most segment, which will have an uncancelled charge on its top face of $+P\,da$, and the bottom-most segment, which will have an uncancelled charge on its bottom face of $-P\,da$. Thus the column of tiny dipoles will appear to be a single large dipole with end charges $+P\,da$ and $-P\,da$, separated by a distance $z_2 - z_1$. Note that nowhere in this derivation have we assumed that A is particularly distant; we have only assumed that the distance to A is much larger than the lengths of the individual microscopic dipoles and much larger than the width of the column in Fig. 4.5(a), both of which are very small.

We can take a slab of dielectric and divide it up infinitesimally into such columns and integrate over them to conclude that the electric field outside the slab is the same as if two sheets of surface charge located where the top and bottom of the slab are located, carrying constant surface charge density $\sigma = +P$ and $\sigma = -P$, respectively. See Fig. 4.6.

4.7 Field Inside Polarized Dielectric Matter

The field inside polarized dielectric matter could be quite complicated. Inside the dielectric we cannot assume that a point is at a much larger distance than the size of the dipoles. However, we may surmise that the electrostatic properties are governed by averages rather than the detailed local microscopic structure. Therefore, let us average over a region that is macroscopically small but microscopically large.[4] The spatial average of $\boldsymbol{E}$ over a volume

[4] That is, a region that is large enough to suppress statistical sampling fluctuations, but small enough that $\boldsymbol{P}$ doesn't vary substantially over the averaging volume. If d is the characteristic size of atoms, L is the averaging scale, and R is the size of the sample, then a suitable macroscopic description of averaged field quantities requires that $d \ll L \ll$. The reason that the averaging procedure preserves important properties of the fields such as $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$ is that derivative operations commute with averaging. A more detailed discussion of the procedure to average microscopic quantities to produce macroscopic quantities may be found in Section 6.6 of Jackson [15] and Section 3.1 of Wald [25].

V inside the polarized matter is given by

$$\langle \boldsymbol{E} \rangle_V = \frac{\int \boldsymbol{E}\, d\tau}{\int d\tau} = \frac{1}{V} \int \boldsymbol{E}\, d\tau, \tag{4.16}$$

where the volume $V = \int d\tau$. This average field $\langle \boldsymbol{E} \rangle_V$ is a *macroscopic quantity* formed from a spatial average of the *microscopic quantity* $\boldsymbol{E}(\boldsymbol{x})$ appearing in the integrand of Eq. (4.16). Let us summarize some important properties of the macroscopic (that is, averaged) electric field.

1. The fundamental relation
$$\boldsymbol{\nabla} \times \boldsymbol{E} = 0 \tag{4.17}$$
of Eq. (2.22) that is obeyed by the microscopic field *remains valid for the macroscopic field.*
2. This implies that the macroscopic electric field is still derivable from a scalar potential in electrostatics (see Section 2.5), through $\boldsymbol{E} = -\boldsymbol{\nabla}\boldsymbol{\Phi}$.

If an electric field is applied to a medium consisting of a large number of atoms or molecules. The dominant multipole mode response to the applied field is the dipole, with an electric *polarization* $\boldsymbol{P}$ corresponding to the density of dipoles given by Eq. (4.11),

$$\boldsymbol{P}(\boldsymbol{x}) = \sum_i N_i \langle \boldsymbol{p}_i \rangle,$$

where $\boldsymbol{p}_i$ is the dipole moment of the ith species (atoms or molecules), and the average $\langle\ \rangle$ is taken over a small volume centered on $\boldsymbol{x}$ and N_i is the average number per unit volume of the ith species at $\boldsymbol{x}$.

We can build up the macroscopic potential or field by linear superposition of contributions from each macroscopically small volume element ΔV at the variable point $\boldsymbol{x}'$. If there are no macroscopic multipole moments higher than dipole, then from Eq. (3.124), the macroscopically averaged potential is

$$\Delta\Phi(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{4\pi\varepsilon_0}\left[\frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\Delta V + \frac{\boldsymbol{P}(\boldsymbol{x}')\cdot(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\Delta V\right], \tag{4.18}$$

assuming $\boldsymbol{x}$ to lie outside ΔV. Setting $\Delta V \to d^3x'$ and integrating over all space gives the potential

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int d^3x' \left[\frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} + \boldsymbol{P}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)\right], \tag{4.19}$$

where the prime on $\boldsymbol{\nabla}'$ indicates that the $\boldsymbol{\nabla}$ operator acts on $\boldsymbol{x}'$ instead of $\boldsymbol{x}$. An integration of the second term by parts leads to

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int d^3x' \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\left[\rho(\boldsymbol{x}') - \boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')\right], \tag{4.20}$$

which represents the potential generated by an effective charge distribution $\tilde{\rho}(\boldsymbol{x}') = \rho(\boldsymbol{x}') -$

$\boldsymbol{\nabla}' \cdot \boldsymbol{P}(\boldsymbol{x}')$.[5] Then Gauss's law (1.1a) reads

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{1}{\varepsilon_0}\left[\rho - \boldsymbol{\nabla}' \cdot \boldsymbol{P}(\boldsymbol{x}')\right], \tag{4.21}$$

which reduces to Eq. (1.1a) in the absence of polarization.

4.8 Displacement and Constitutive Relations

If we define the *electric displacement* $\boldsymbol{D}$ by

$$\boldsymbol{D} \equiv \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}, \tag{4.22}$$

Eq. (4.21) can be written as

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho. \tag{4.23}$$

Solution for potentials or fields requires also that the *constitutive relations* that connect $\boldsymbol{D}$ and $\boldsymbol{E}$ be specified.

1. A common assumption is that the response to the applied field is linear, $\boldsymbol{P} \propto \boldsymbol{E}$.
2. A second common assumption is that the medium is isotropic, so that $\boldsymbol{P}$ is parallel to $\boldsymbol{E}$ and the coefficient of proportionality has no angular dependence,

$$\boldsymbol{P} = \varepsilon_0 \chi_{\mathrm{e}} \boldsymbol{E}, \tag{4.24}$$

   where χ_{e} is termed the *electric susceptibility* of the medium.

With these assumptions the displacement $\boldsymbol{D}$ and the electric field $\boldsymbol{E}$ are related by

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} \qquad \varepsilon \equiv \varepsilon_0(1 + \chi_{\mathrm{e}}), \tag{4.25}$$

where ε is the *electric permittivity* and

$$\kappa \equiv \frac{\varepsilon}{\varepsilon_0} = 1 + \chi_{\mathrm{e}} \tag{4.26}$$

is called the *dielectric constant*. Then the polarization can be written

$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E}. \tag{4.27}$$

If the dielectric medium is *uniform in addition to isotropic*, then ε is independent of position and the divergence equation (4.23) can be written

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon}, \tag{4.28}$$

which is Gauss's law (1.1a) with ε_0 replaced by $\varepsilon = \varepsilon_0(1 + \chi_{\mathrm{e}})$. Equations (4.23) and (4.17) are the macroscopic counterparts of the microscopic equations (1.1a) and (2.22), respectively.

[5] The divergence term appears in the effective charge density $\tilde{\rho}$ because for non-uniform polarization there can be a net increase or decrease of charge within any small volume. See Section 4.9. As we have noted in Eq. (4.12), the polarization-charge density is given by $-\boldsymbol{\nabla} \cdot \boldsymbol{P}$.

> For a medium fulfilling all the conditions given above, solutions for electrostatics problems in that medium are equivalent to the corresponding solutions found previously for vacuum, except that the electric fields must be reduced by a factor $\varepsilon_0/\varepsilon$. Physically this reduction is a consequence of polarized atoms producing fields that oppose the applied field.

One immediate consequence relevant to our prior discussion is that the capacitance of a capacitor is increased by a factor $\varepsilon/\varepsilon_0$ if the empty space between the electrodes is filled with a material having dielectric constant $\varepsilon/\varepsilon_0$ (and the effect of fringing fields may be neglected). Example 4.4 illustrates.

Example 4.4 From Eq. (4.25), the effect of inserting a dielectric layer in a parallel-plate capacitor is to alter the electric field, which alters the capacitance from its value C_0 without the dielectric,

$$C = \kappa C_0 = \frac{\varepsilon}{\varepsilon_0} C_0 = (1 - \chi_e) C_0. \tag{4.29}$$

For the parallel-plate capacitor described in Section 3.2.1, the capacitance with a dielectric layer between the plates is

$$C = \frac{\varepsilon}{\varepsilon_0} C_0 = \frac{\varepsilon}{\varepsilon_0} \frac{A\varepsilon_0}{d} = \frac{A\varepsilon}{d}. \tag{4.30}$$

The corresponding work required to charge a parallel-plate capacitor given in Section 3.2.2 is modified to

$$W = \frac{Q^2}{2C} = \frac{1}{2} CV^2 = \frac{1}{2} \frac{A\varepsilon}{d} V^2, \tag{4.31}$$

which is $\varepsilon/\varepsilon_0$ times the work required without the dielectric layer that was given in Eq. (3.8).

If a system contains different media juxtaposed, the question of boundary conditions must be considered for $\boldsymbol{D}$ and $\boldsymbol{E}$ at interfaces between media. This will be addressed in Section 4.11. The results for electrostatics are that the normal components of $\boldsymbol{D}$ and the tangential components of $\boldsymbol{E}$ on either side of an interface between medium 1 and medium 2 must satisfy the boundary conditions (for static or time-varying fields)

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n} = \sigma, \tag{4.32a}$$

$$(\boldsymbol{E}_2 - \boldsymbol{E}_1) \times \boldsymbol{n} = 0, \tag{4.32b}$$

where $\boldsymbol{n}$ is a unit normal to the surface pointing from medium 1 to medium 2, and σ is the macroscopic surface-charge density on the boundary surface (which does not include the polarization charge discussed in following sections).

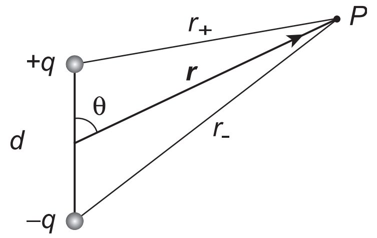


Fig. 4.7 Coordinate system for an electric dipole. The distance from the dipole to P is $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$ and $r \equiv |\boldsymbol{r}|$.

4.9 Surface and Volume Bound Charges

The potential created by a single dipole can be written

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \frac{\boldsymbol{p} \cdot \hat{\boldsymbol{r}}}{r^2}, \tag{4.33}$$

where from Fig. 4.7, $r = |\boldsymbol{r}|$, $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$, and $\boldsymbol{p}$ is the dipole moment vector. The total potential contributed by dipoles then follows by integration,

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \int_V \frac{\boldsymbol{P}(\boldsymbol{x}') \cdot \hat{\boldsymbol{r}}}{r^2}\, d\tau', \tag{4.34}$$

where $d\tau' \equiv d^3x'$ and $\boldsymbol{P}$ is the dipole moment density. This can be rewritten using

$$\boldsymbol{\nabla}' \left(\frac{1}{r} \right) = \frac{\hat{\boldsymbol{r}}}{r^2} \tag{4.35}$$

in the form

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \int_V \boldsymbol{P} \cdot \boldsymbol{\nabla}' \left(\frac{1}{r} \right) d\tau'. \tag{4.36}$$

This can be integrated by parts using the product rule in Eq. (A.18c)

$$\boldsymbol{\nabla}' \cdot (f\boldsymbol{A}) = f(\boldsymbol{\nabla}' \cdot \boldsymbol{A}) + \boldsymbol{A} \cdot (\boldsymbol{\nabla}' f).$$

Setting $f = 1/r$ and $\boldsymbol{A} = \boldsymbol{P}$, and integrating both sides over the volume V, the product rule becomes

$$\int_V \boldsymbol{\nabla}' \cdot \left(\frac{\boldsymbol{P}}{r} \right) d\tau' = \int_V \frac{1}{r} (\boldsymbol{\nabla}' \cdot \boldsymbol{P}) d\tau' + \int_V (\boldsymbol{P} \cdot \boldsymbol{\nabla}') \frac{1}{r} d\tau' \tag{4.37}$$

Now apply the divergence theorem (A.33)

$$\oint_S \boldsymbol{A} \cdot \boldsymbol{n}\, da = \int_V \boldsymbol{\nabla} \cdot \boldsymbol{A}\, d^3x,$$

to the term on the left side of Eq. (4.37) to give

$$\oint_S \frac{1}{r} \boldsymbol{P} \cdot \boldsymbol{n}\, da' = \int_V \frac{1}{r} (\boldsymbol{\nabla}' \cdot \boldsymbol{P}) d\tau' + \int_V (\boldsymbol{P} \cdot \boldsymbol{\nabla}') \frac{1}{r} d\tau'. \tag{4.38}$$

Multiply both sides by $1/4\pi\varepsilon_0$ and rearrange to give

$$\frac{1}{4\pi\varepsilon_0}\int_V (\boldsymbol{P}\cdot\boldsymbol{\nabla}')\frac{1}{r}d\tau' = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{r}\boldsymbol{P}\cdot\boldsymbol{n}\,da' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau', \tag{4.39}$$

and finally, comparing with Eq. (4.36),

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{r}\boldsymbol{P}\cdot\boldsymbol{n}\,da' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau'. \tag{4.40}$$

Thus we see that

1. The first term of Eq. (4.40) looks like the potential generated by a surface charge of magnitude

$$\sigma_{\mathrm{b}} \equiv \boldsymbol{P}\cdot\hat{\boldsymbol{n}}, \tag{4.41}$$

with $\hat{\boldsymbol{n}}$ the unit vector normal to the surface and the subscript "b" indicates that it originates in the bound charges.

2. The second term of Eq. (4.40) looks like the potential generated by a volume charge of magnitude

$$\rho_{\mathrm{b}} = -\boldsymbol{\nabla}\cdot\boldsymbol{P}. \tag{4.42}$$

Equation (4.40) can thus be written

$$\Phi = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{\sigma_{\mathrm{b}}}{r}da' + \frac{1}{4\pi\varepsilon_0}\int_V \frac{\rho_{\mathrm{b}}}{r}d\tau'. \tag{4.43}$$

The potential Φ and the field $\boldsymbol{E}$ of a polarized object is equivalent to that produced by a volume charge density $\rho_{\mathrm{b}} = -\boldsymbol{\nabla}\cdot\boldsymbol{P}$ plus a surface charge density $\sigma_{\mathrm{b}} = \boldsymbol{P}\cdot\hat{\boldsymbol{n}}$. Notice that the volume charge density ρ_{b} involves derivatives of the polarization $\boldsymbol{P}$. Thus it contributes only if the polarization is spatially non-uniform [see Eq. (4.20)].

Example 4.5 Let's use the result of Eq. (4.43) to determine the potential inside and outside a uniformly polarized sphere of radius R. Since we assume uniform polarization the second (volume-charge) term, which is non-zero only if the derivative of the polarization density is finite, makes no contribution and the potential is generated entirely by the surface charge defined in Eq. (4.41),

$$\sigma_{\mathrm{b}} = \boldsymbol{P}\cdot\hat{\boldsymbol{n}} = P\cos\theta, \tag{4.44}$$

where we have chosen the z-axis as the direction of polarization. Thus, we need to evaluate the potential for a sphere with the surface charge (4.44) painted on it. This problem was already solved in Problems 3.5 and 4.2, with the result that

$$\Phi(r,\theta) = \begin{cases} \dfrac{Pr}{3\varepsilon_0}\cos\theta & (r \le R), \\ \dfrac{PR^3}{3\varepsilon_0 r^2}\cos\theta & (r \ge R). \end{cases} \tag{4.45}$$

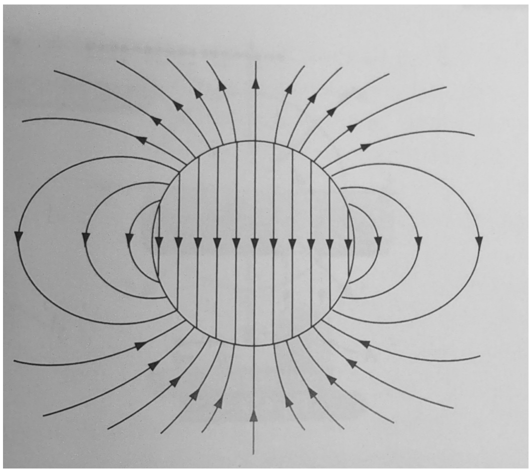


Fig. 4.8 Electric field for uniformly polarized sphere.

The electric field is then given by $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$. Note that since $r\cos\theta = z$, inside the sphere the field is uniform:

$$\boldsymbol{E} = -\boldsymbol{\nabla}\Phi = -\frac{P}{3\varepsilon_0}\hat{\boldsymbol{z}} = -\frac{1}{3\varepsilon_0}\boldsymbol{P} \qquad (r < R). \tag{4.46}$$

Outside the sphere the field has a dipole form

$$\Phi = \frac{1}{4\pi\varepsilon_0}\frac{\boldsymbol{p}\cdot\hat{\boldsymbol{r}}}{r^2} \qquad (r \geq R),$$

where the dipole moment is equal to the total dipole moment of the sphere

$$\boldsymbol{p} = \frac{4}{3}\pi R^3\boldsymbol{P}.$$

The field lines for $\boldsymbol{E}$ of the uniformly polarized sphere are plotted in Fig. 4.8.

4.10 Average Electric Fields in Matter

The influence of matter on electrostatics will in most cases require some amount of averaging over the detailed and complex microscopic interactions in the matter. One important concept will be to compute the average electric field in some volume of matter. Let's begin by consider a sphere containing a single point charge Q located a distance r' from the origin along the z-axis, as illustrated in Fig. 4.9 [4]. By symmetry the average field over the entire volume must be along the z-axis The average field is then

$$\langle E_z\rangle = \frac{\int_\tau E_z\,d\tau}{\int_\tau d\tau} = \frac{1}{\tau}\int_\tau E_z\,d\tau,$$

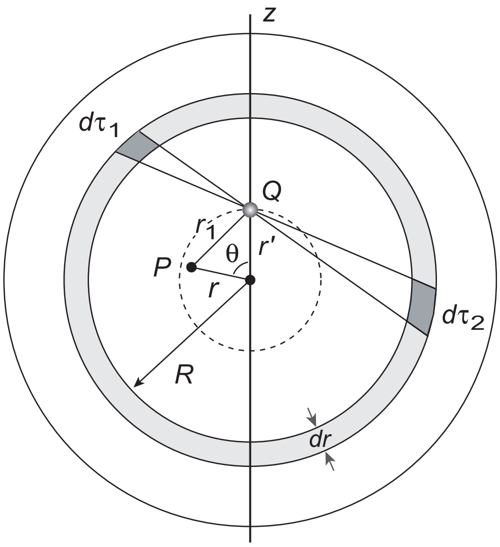


Fig. 4.9 A point charge Q located at a distance r' from the center of a sphere of radius R.

where τ is the volume of the sphere. It is convenient to separate the integral into two parts: one over the spherical shell from radii r' to R (outside the dashed circle in the above diagram) and one over the sphere of radius r' (inside the dashed circle).

The integral over the outer volume vanishes, by the following qualitative argument. Consider the concentric shell at radius R with thickness dr. The solid angle element intercepts the volume elements $d\tau_1$ and $d\tau_2$ in the shaded shell of thickness dr. The value of E_z decreases quadratically with distance from Q but $d\tau$ increases quadratically with distance, so their product remains constant. But E_z is positive at $d\tau_1$ but negative at $d\tau_2$, so the two contributions cancel. A similar argument can be make for all shells and the entire outer shell with $r > r'$ contributes zero.

To calculate the integral over the inner volume (inside the dashed circle), consider the point P in Fig. 4.9. The potential at P is

$$\Phi = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r_1} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{|\boldsymbol{x}-\boldsymbol{x}'|},$$

where from Fig. 4.9 that $r_1 = |\boldsymbol{x}-\boldsymbol{x}'|$. Thus, we may expand in the multipole expansion given in Eq. (3.111),

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \sum_{l=0}^{\infty}\frac{r_<^l}{r_>^{l+1}}P_l(\cos\theta) = \sum_{l=0}^{\infty}\frac{r^l}{(r')^{l+1}}P_l(\cos\theta)$$

where we've used from the diagram that $r < r'$. Writing this expansion out and substituting the explicit values for the Legendre polynomials from Table 3.1,

$$\Phi = \frac{Q}{4\pi\varepsilon_0}\frac{1}{r'}\left[1+\frac{r}{r'}\cos\theta+\frac{1}{2}\frac{r^2}{r'^2}(3\cos^2\theta-1)+\frac{1}{2}\frac{r^3}{r'^3}(5\cos^3\theta-3\cos\theta+\cdots\right].$$

The electric field is minus the gradient of the potential, so we need to evaluate $E_z = -\partial\Phi/\partial z$. Utilizing

$$r\cos\theta = z \qquad \frac{\partial r}{\partial z} = \frac{\partial}{\partial z}(x^2+y^2+z^2)^{1/2} = \frac{z}{r} = \cos\theta$$

the preceding multipole expansion can be written in terms of z and the derivative taken to give the expansion for the electric field

$$E_z = -\frac{Q}{4\pi\varepsilon_0 r'^2}\left[1+\frac{2z}{r'}+\frac{3}{2r'^2}(3z^2-r^2)+\cdots\right]$$

Then we may compute the average by integrating term by term in this series. The first term gives

$$\langle E_z\rangle_1 = -\frac{Q}{4\pi\varepsilon_0 r'^2}\int_0^\pi\int_0^{r'} 2\pi r^2\sin\theta\, dr\, d\theta = -\frac{Qr'}{3\varepsilon_0\tau}.$$

All of the higher-order terms give zero, so we obtain

$$\langle E_z\rangle = -\frac{Qr'}{3\varepsilon_0\tau} = -\frac{Qr'}{3\varepsilon_0}\frac{3}{4\pi R^3} = -\frac{Qr'}{4\pi\varepsilon_0 R^3} = -\frac{p}{4\pi\varepsilon_0 R^3},$$

where the volume is $\tau = \frac{4}{3}\pi R^3$ and the dipole moment p of the charge Q is $p = Qr'$. This result was for a single charge on the z-axis. For an arbitrary charge distribution the same result is obtained, except that

$$\langle E_z\rangle = -\frac{p_{\text{total}}}{4\pi\varepsilon_0 R^3},$$

where p_{total} is the total dipole moment of the arbitrary charge distribution within the sphere of radius R.

4.11 Boundary Conditions at Interfaces

We must often consider problems in which media with different properties are adjacent and boundary conditions at interfaces must be evaluated. In a medium that is possibly polarized by electric or magnetic fields, the vacuum Maxwell equations (1.1) are modified to the Maxwell equations in medium,

$$\nabla\cdot\boldsymbol{D} = \rho \qquad \text{(Gauss's law)}, \qquad (4.47a)$$

$$\nabla\times\boldsymbol{E} + \frac{\partial\boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \qquad (4.47b)$$

$$\nabla\cdot\boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \qquad (4.47c)$$

$$\nabla\times\boldsymbol{H} - \frac{\partial\boldsymbol{D}}{\partial t} = \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \qquad (4.47d)$$

where $\boldsymbol{D}$ is defined in Eq. (4.22) and $\boldsymbol{H}$ is defined in Eq. (6.7).[6] The Maxwell equations (4.47) in differential form can be cast in integral form using the divergence theorem and Stokes' theorem.

Let V be a finite volume bounded by a closed surface (or surfaces) S, let da be an area element of that surface, and let $\boldsymbol{n}$ be a unit normal at da, pointing out of the volume. The divergence theorem (2.15)

$$\oint_S \boldsymbol{A}\cdot\boldsymbol{n}\,da = \int_V \boldsymbol{\nabla}\cdot\boldsymbol{A}\,d^3x,$$

applied to Eq. (4.47a) yields

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\,d^3x, \tag{4.48}$$

with Eq. (4.48) being an integral form of Gauss's law requiring that the total flux $\boldsymbol{D}$ out through the surface be equal to the charge contained in the volume. Likewise, applying the divergence theorem to Eq. (4.47c) yields the integral equation

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0, \tag{4.49}$$

with Eq. (4.49) being the magnetic analog of Eq. (4.48) requiring no net flux of B through the closed surface because (as far as we know) magnetic charges do not exist.

In a similar manner, suppose that C is a closed contour spanned by an open surface S', $\boldsymbol{dl}$ is a line element on C, da is an area element on S', and $\boldsymbol{n}'$ is a unit vector pointing in a direction given by the right-hand rule (see Box 2.3). Applying Stokes' theorem (2.20),

$$\int_S (\boldsymbol{\nabla}\times\boldsymbol{A})\cdot\boldsymbol{n}\,da = \oint_C \boldsymbol{A}\cdot\boldsymbol{dl},$$

to Eq. (4.47b) gives

$$\oint_C \boldsymbol{E}\cdot\boldsymbol{dl} = -\int_S \frac{\partial\boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}'da, \tag{4.50}$$

which is an integral form of Faraday's law of magnetic induction. Likewise, applying Stokes' theorem to Eq. (4.47d) gives

$$\oint_C \boldsymbol{H}\cdot\boldsymbol{dl} = \oint_{S'}\left(\boldsymbol{J}+\frac{\partial\boldsymbol{D}}{\partial t}\right)\cdot\boldsymbol{n}'da \tag{4.51}$$

which is an integral form of the Ampère–Maxwell law of magnetic fields. Summarizing equations (4.48)-(4.51),

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\,d^3x \qquad \text{(Gauss's law)} \tag{4.52a}$$

$$\oint_C \boldsymbol{E}\cdot\boldsymbol{dl} = -\int_S \frac{\partial\boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}'da, \qquad \text{(Faraday's law)} \tag{4.52b}$$

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0, \qquad \text{(No magnetic charges)} \tag{4.52c}$$

$$\oint_C \boldsymbol{H}\cdot\boldsymbol{dl} = \oint_{S'}\left(\boldsymbol{J}+\frac{\partial\boldsymbol{D}}{\partial t}\right)\cdot\boldsymbol{n}'da \qquad \text{(Ampère–Maxwell law)} \tag{4.52d}$$

[6] The vacuum Maxwell equations (1.1) can be recovered by substituting $\boldsymbol{D} = \varepsilon_0\boldsymbol{E}$ and $\boldsymbol{H} = \boldsymbol{B}/\mu_0$ into Eqs. (4.47), remembering from Eq. (2.4) that $\mu_0\varepsilon_0 = 1/c^2$.

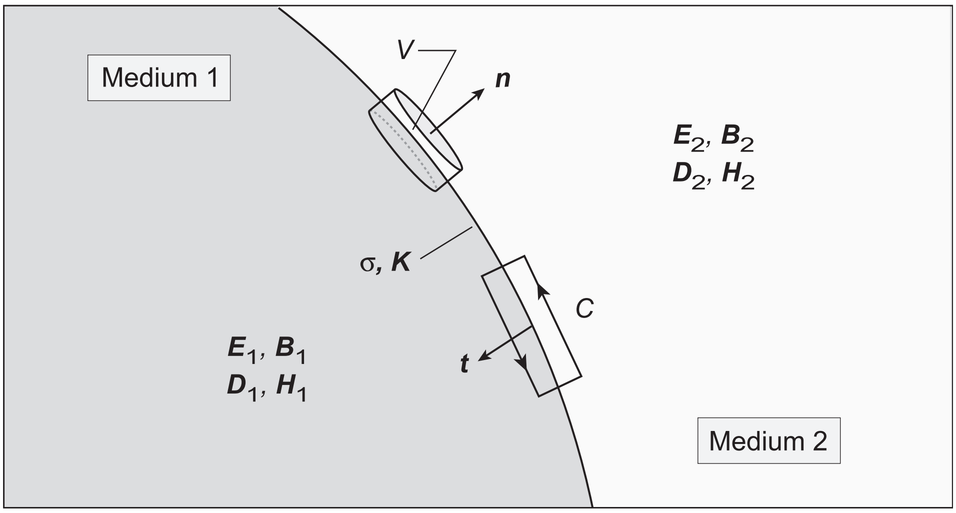


Fig. 4.10 Schematic illustration of a boundary surface between different media that is assumed to carry surface charge σ and surface current density $\boldsymbol{K}$. The infinitesimal cylinder of volume V is half in one medium and half in the other, with the normal $\boldsymbol{n}$ to its surface pointing from medium 1 into medium 2. The infinitesimal rectangular contour C is partly in one medium and partly in the other, with its plane perpendicular to the surface so that its normal $\boldsymbol{t}$ (pointing out of the page) is tangent to the surface. Adapted from Ref. [15].

define Maxwell's equations in medium in integral form.

These equations may be used to determine the relationship of normal and tangential components of the fields on either side of an interface between media having different electromagnetic properties, as we now discuss [15]. Consider Fig. 4.10, where there is a boundary between medium 1 and medium 2, and we allow the possibility that there is a surface charge σ and a current density $\boldsymbol{K}$ at the interface. To facilitate our analysis, an infinitesimal cylindrical Gaussian pillbox of volume V straddles the surface between the two media. In addition, an infinitesimal rectangular contour C has long sides on either side of the boundary and is oriented so that the normal to the rectangular surface points out of the page and is tangent to the interface.

Let us first apply equations (4.52a) and (4.52c) to the cylindrical pillbox in Fig. 4.10. In the limit that the pillbox is very shallow the side of the cylinder does not contribute to the integrals on the left side of Eqs. (4.52a) and (4.52c). If the top and bottom of the cylinder are parallel to the interface and have area Δa, then the integral on the left side of Eq. (4.52a) is

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = (\boldsymbol{D}_2 - \boldsymbol{D}_1)\cdot\boldsymbol{n}\Delta a.$$

Applying a similar argument to the left side of Eq. (4.52c), the integral is

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = (\boldsymbol{B}_2 - \boldsymbol{B}_1)\cdot\boldsymbol{n}\Delta a.$$

If the charge density ρ is singular at the interface and produces an idealized surface charge density σ, the integral on the right side of Eq. (4.52a) evaluates to

$$\int_V \rho\, d^3x = \sigma \Delta a,$$

and the normal components of $\boldsymbol{D}$ and $\boldsymbol{B}$ on the two sides of the interface are related by

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n} = \sigma, \tag{4.53a}$$

$$(\boldsymbol{B}_2 - \boldsymbol{B}_1) \cdot \boldsymbol{n} = 0. \tag{4.53b}$$

Thus, in words

1. the normal component of $\boldsymbol{B}$ is continuous across the interface but
2. the discontinuity of the normal component of $\boldsymbol{D}$ at any point is equal to the surface charge density σ at the point.

In a similar manner, the infinitesimal rectangular contour C in Fig. 4.10 can be used in conjunction with Stokes' theorem to determine the discontinuities in the tangential components of $\boldsymbol{E}$ and $\boldsymbol{H}$. In the limit that the short sides of the rectangular loop may be neglected and each long side is of length Δl and parallel to the interface, the integral on the left side of Eq. (4.52b) is

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{l} = (\boldsymbol{t} \times \boldsymbol{n}) \cdot (\boldsymbol{E}_2 - \boldsymbol{E}_1)\Delta l.$$

Likewise, the integral on the left side of Eq. (4.52d) is

$$\oint_C \boldsymbol{H} \cdot d\boldsymbol{l} = (\boldsymbol{t} \times \boldsymbol{n}) \cdot (\boldsymbol{H}_2 - \boldsymbol{H}_1)\Delta l.$$

The right side of Eq. (4.52b) vanishes because in the limit of vanishing length of the short side of the rectangular contour $\partial \boldsymbol{B}/\partial t$ is finite while $\Delta t \to 0$. Because there is an idealized surface current density $\boldsymbol{K}$ flowing exactly on the boundary, the integral on the right side of Eq. (4.52d) is equal to

$$\oint_{S'} \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \right) \cdot \boldsymbol{t}\, da = \boldsymbol{K} \cdot \boldsymbol{t} \Delta l,$$

where the second term vanishes by the same argument as for the right side of Eq. (4.52b). Therefore, the tangential components of $\boldsymbol{E}$ and $\boldsymbol{H}$ on either side of the media interface are related by

$$(\boldsymbol{t} \times \boldsymbol{n}) \cdot (\boldsymbol{E}_2 - \boldsymbol{E}_1) = 0 \qquad (\boldsymbol{t} \times \boldsymbol{n}) \cdot (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{K} \cdot \boldsymbol{t},$$

and using the identity (A.4), this implies that

$$\boldsymbol{n} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1) = 0, \tag{4.54a}$$

$$\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{K}, \tag{4.54b}$$

where in Eq. (4.54b) it is understood that the surface current has only components parallel to the interface at every point. Thus,

1. the tangential component of $\boldsymbol{E}$ is continuous across an interface, but
2. the tangential component of $\boldsymbol{H}$ is discontinuous across the interface by an amount with magnitude equal to $|\boldsymbol{K}|$ and direction given by the direction of $\boldsymbol{K} \times \boldsymbol{n}$.

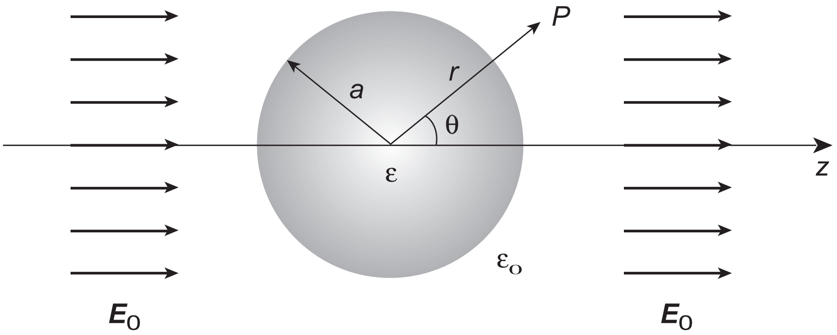


Fig. 4.11 Delectric ball of radius a and dielectric constant $\kappa = \varepsilon/\varepsilon_0$ in an initially uniform external electric field $\boldsymbol{E}_0$ directed along the z axis. Used in Example 4.6.

The discontinuity equations (4.53) and (4.54) for the fields $\boldsymbol{B}$, $\boldsymbol{E}$, $\boldsymbol{D}$, and $\boldsymbol{H}$ allow solving the Maxwell equations in different regions having potentially different electromagnetic properties, and then connecting the solutions to obtain the fields over all of the space.

4.12 Dielectric Boundary Value Problems

The methods developed in previous chapters may be adapted to handle the presence of dielectrics. Example 4.6 illustrates.

Example 4.6 A dielectric ball with dielectric constant $\kappa = \varepsilon/\varepsilon_0$ is placed in an initially uniform electric field directed along the z axis, as illustrated in Fig. 4.11. Let's determine the potential and the electric field inside and outside the ball, assuming no free charges inside or outside of the ball. There are no free charges so the solution will involve solving the Laplace equation with appropriate boundary conditions. We assume axial symmetry about the z axis. Solving the Laplace equation in spherical coordinates by separation of variables (see Section 3.10.2), general solutions are of the form given by Eq. (3.97),

$$\Phi(r,\theta) = \sum_{l=0}^{\infty} (A_l r^l + B_l r^{-(l+1)}) P_l(\cos\theta).$$

However, there are no charges at the origin and the requirement that $\Phi(r,\theta)$ be finite there demands that for the interior solution, $B_l = 0$, and the interior solution is of the form

$$\Phi_{\text{in}}(r,\theta) = \sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta), \tag{4.55}$$

and the exterior solution is of the form

$$\Phi_{\text{out}}(r,\theta) = \sum_{l=0}^{\infty} (B_l r^l + C_l r^{-(l+1)}) P_l(\cos\theta),$$

with the constants A_l, B_l, and C_l to be determined by imposing boundary conditions. At infinity, we must have [see Eq. (3.101)],

$$\Phi(r \to \infty) = -E_0 z = -E_0 r\cos\theta = -E_0 r P_1(\cos\theta).$$

This requires that

$$-E_0 r P_1(\cos\theta) = [B_l r^l + C_l r^{-(l+1)}]P_l(\cos\theta) \simeq B_l r^l P_l(\cos\theta),$$

since the C_l term vanishes as $r \to \infty$. Multiply both sides by $P_1(\cos\theta)$ and integrate,

$$-E_0 r \int P_1(\cos\theta)P_1(\cos\theta)d(\cos\theta) = B_l r^l \int P_1(\cos\theta)P_l(\cos\theta)d(\cos\theta).$$

Using the orthogonality relation (3.98),

$$\int_{-1}^{+1} P_k(\cos\theta)P_l(\cos\theta)d(\cos\theta) = \frac{2}{2l+1}\delta_{kl},$$

all terms vanish except for $l = 1$ and we obtain $B_1 = -E_0$, with all other B_l equal to zero. Thus, the exterior solution becomes,

$$\Phi_{\text{out}}(r,\theta) = -E_0 r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)} P_l(\cos\theta), \tag{4.56}$$

Now let's use the boundary conditions at the edge of the sphere to fix the other constants, by requiring the matching at the surface $r = a$ for tangential and normal components of Φ,

$$-\frac{1}{a}\frac{\partial\Phi_{\text{in}}}{\partial\theta}\bigg|_{r=a} = -\frac{1}{a}\frac{\partial\Phi_{\text{out}}}{\partial\theta}\bigg|_{r=a} \qquad (\text{Tangential}), \tag{4.57}$$

$$-\varepsilon\frac{\partial\Phi_{\text{in}}}{\partial r}\bigg|_{r=a} = -\varepsilon_0\frac{\partial\Phi_{\text{out}}}{\partial r}\bigg|_{r=a} \qquad (\text{Normal}). \tag{4.58}$$

Now substitute the expansions (4.55) and (4.56) into these matching equations and solve to determine the constants.

First consider the tangential matching. Substituting (4.55) and (4.56) into Eq. (4.57) gives

$$-\frac{1}{a}\frac{\partial}{\partial\theta}\sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta)\bigg|_{r=a} = -\frac{1}{a}\frac{\partial}{\partial\theta}\left[B_l r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)}\right]_{r=a} P_l(\cos\theta),$$

which simplifies to

$$\sum_{l=0}^{\infty} a^l A_l \frac{\partial}{\partial\theta}P_l(\cos\theta) = -aE_0\frac{\partial}{\partial\theta}P_1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \frac{\partial}{\partial\theta}P_l(\cos\theta). \tag{4.59}$$

Let's convert the derivatives of Legendre polynomials $P_n(x)$ to associated Legendre polynomials $P_n^m(x)$ using the general relationship [1],

$$P_n^m(x) = (1-x^2)^{m/2}\frac{d^m}{dx^m}P_n(x), \tag{4.60}$$

which upon specializing to $m = 1$ yields

$$\frac{dP_n(x)}{dx} = (1-x^2)^{-1/2} P_n^1(x), \tag{4.61}$$

or letting $x = \cos\theta$,

$$\frac{dP_n(\cos\theta)}{d\theta} = -P_n^1(\cos\theta), \tag{4.62}$$

Then Eq. (4.59) becomes

$$-\sum_{l=0}^{\infty} a^l A_l P_l^1(\cos\theta) = -aE_0 P_1^1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)} C_l P_l^1(\cos\theta) \tag{4.63}$$

We will now exploit the orthogonality properties of the associated Legendre polynomials, which obey the relationship [1],

$$\int_{-1}^{+1} P_p^m(x) P_q^m(x) dx = K_{qm} \delta_{pq} \qquad K_{qm} \equiv \frac{2}{2q+1} \frac{(q+m)!}{(q-m)!} \tag{4.64}$$

or in spherical coordinates

$$\int_0^{\pi} P_p^m(\cos\theta) P_q^m(\cos\theta) \sin\theta d\theta = K_{qm}\, \delta_{pq} \tag{4.65}$$

There are two solutions, a special solution for $l = 1$ and a general solution for all $l \neq 1$. To obtain the special solution let $x = \cos\theta$, multiply both sides by $P_1^1(x)$, and integrate.

$$-\sum_{l=0}^{\infty} a^l A_l \int P_1^1(x) P_l^1(x) dx = -aE_0 \int P_1^1(x) P_1^1(x) dx + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \int P_l^1(x) P_1^1(x) dx$$

Utilizing Eq. (4.64), this gives

$$-\sum_{l=0}^{\infty} a^l A_l \delta_{1l} = -aE_0 + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \delta_{1l}$$

and finally

$$aA_1 = -aE_0 + a^{-2} C_1$$

or equivalently,

$$A_1 = -E_0 + \frac{C_1}{a^3}. \tag{4.66}$$

Now let's find the general solution for $l \neq 1$. Let $x = \cos\theta$ and multiply both sides of Eq. (4.63) by $P_k^1(x)$ and integrate,

$$-\sum_{l=0}^{\infty} a^l A_l \int P_k^1(x) P_l^1(x) dx = -aE_0 \int P_k^1(x) P_1^1(x) dx + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \int P_k^1(x) P_l^1(x) dx$$

Utilizing the orthogonality relation (4.64), this gives

$$a^l A_l = a^{-(l+1)} C_l.$$

which rearranges to

$$A_l = \frac{C_l}{a^{2l+1}}. \tag{4.67}$$

We now have two relations [Eqs. (4.66) and (4.67)] among the coefficients obtained from the tangential matching condition (4.57). Now let's use the normal matching condition (4.58) to obtain additional constraints. From Eq. (4.58), upon substituting the expansions (4.55) and (4.56),

$$\varepsilon \sum_{l=0}^{\infty} l a^{l-1} A_l P_l(x) = \varepsilon_0 B_1 P_1(x) + \sum_{l=0}^{\infty} -(l+1) C_l a^{-(l+2)} P_l(x). \tag{4.68}$$

As before, there are two solutions, a special one when $l = 1$, and a general one when $l \neq 1$. Lets find the special solution by multiplying Eq. (4.68) by $P_1(x)$ and integrating. Upon exploiting the orthogonality properties the result is another relationship among coefficients:

$$\frac{\varepsilon}{\varepsilon_0} A_1 = -E_0 - 2\frac{C_1}{a^3} \qquad (l = 1). \tag{4.69}$$

The general solution may be obtained by multiplying both sides of Eq. (4.68) by $P_k(x)$, integrating, and exploiting the orthogonality constraints (3.98). The result is

$$\frac{\varepsilon}{\varepsilon_0} l A_l = -(l+1)\frac{C_l}{a^{2l+1}} \qquad (l \neq 1). \tag{4.70}$$

We now have four equations, (4.66), (4.67), (4.69), and (4.70), to solve simultaneously for the unknown coefficients. Equations (4.67) and (4.70) can be satisfied only if $A_l = C_l = 0$ for all $l \neq 1$. Solving the other two equations (4.66) and (4.69) simultaneously for $l = 1$ gives

$$A_1 = -\left(\frac{3}{\varepsilon/\varepsilon_0 + 2}\right) E_0 \qquad C_1 = \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) a^3 E_0. \tag{4.71}$$

Inserting these results into Eqs. (4.55) and (4.56) gives for the interior and exterior potentials

$$\Phi_{\text{in}} = -\left(\frac{3}{2 + \varepsilon/\varepsilon_0}\right) E_0\, r\cos\theta = -\left(\frac{3}{2 + \varepsilon/\varepsilon_0}\right) E_0\, z, \tag{4.72a}$$

$$\Phi_{\text{out}} = -E_0\, z + \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) E_0\, a^3 \frac{\cos\theta}{r^2}, \tag{4.72b}$$

The electric fields then follow from $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$. For the interior,

$$E_{\text{in}} = -\frac{\partial}{\partial z}\left[\left(\frac{-3}{\varepsilon/\varepsilon_0 + 2}\right) E_0\, z\right] = \left(\frac{3}{\varepsilon/\varepsilon_0 + 2}\right) E_0, \tag{4.73}$$

from which we conclude that

1. the interior field E_{in} is constant, proportional to the constant exterior field,
2. if $\varepsilon = \varepsilon_0$, then the interior field is equal to the exterior field, $E_{\text{in}} = E_0$, and
3. if $\varepsilon > \varepsilon_0$, then the interior field is reduced relative to the exterior field, $E_{\text{in}} < E_0$.

In Eq. (4.72b) it is clear that the first term is due to the unperturbed exterior field $\boldsymbol{E}_0$ and the second term is a correction associated with a potential generated by the polarized dielectric sphere. If the induced electric dipole moment $\boldsymbol{p}$ [see Eq. (3.116)] of the polarized sphere is taken to have magnitude

$$p = 4\pi\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) a^3 E_0, \tag{4.74}$$

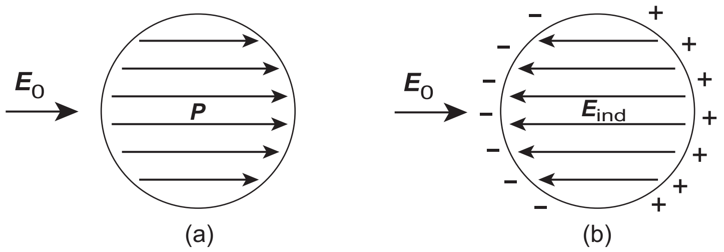
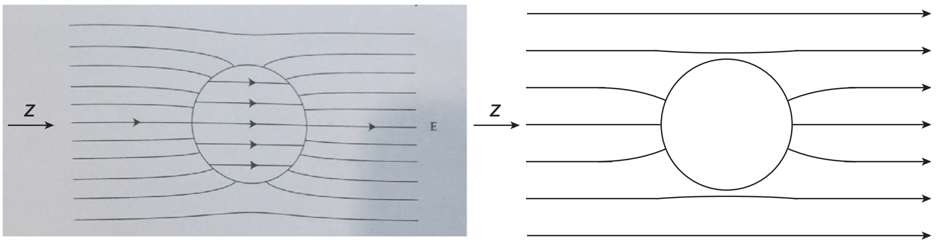


Fig. 4.12 (a) Polarization of a dielectric ball by an external electric field $\boldsymbol{E}_0$. (b) The polarization of charge induces a field $\boldsymbol{E}_{\text{ind}}$ that opposes the applied field $\boldsymbol{E}_0$.



(a) Dielectric filled sphere

(a) Metal filled sphere

Fig. 4.13 (a) Electric field lines for dielectric filled sphere in an external electric field $\boldsymbol{E}_0$ directed along the z axis. (b) Electric field lines for a conducting ball in an external electric field $\boldsymbol{E}_0$ directed along the z axis. Original calculation in Fig. 3.13.

then the exterior potential (4.72b) can be written

$$\Phi_{\text{out}} = -E_0 z + \frac{p}{4\pi\varepsilon_0}\frac{\cos\theta}{r^2}, \tag{4.75}$$

making clear that the external potential is the unperturbed external potential (first term) modified by a potential corresponding to an induced electric dipole associated with the polarized dielectric sphere. From Eqs. (4.24) and (4.26), the polarization $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E}, \tag{4.76}$$

and from Eq. (4.11) the polarization may also be defined as the density of dipole moments. Then from Eq. (4.74) the polarization of the dielectric sphere is given by

$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E} = \left(\frac{\text{dipole moments}}{\text{unit volume}}\right) = \frac{p}{\frac{4}{3}\pi a^3} = 3\varepsilon_0\left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right)\boldsymbol{E}_0. \tag{4.77}$$

The results of this example are displayed in Fig. 4.12 and Fig. 4.13.

! In Fig. 4.12 the results of Example 4.6 are illustrated. As indicated schematically in Fig. 4.12(a), the external field E_0 aligned in the z direction polarizes the dielectric sphere, with the polarization vector $\boldsymbol{P}$ given in Eq. (4.77) and oriented in the direction of the external field. The polarization is the density of induced dipole moments (4.74). This corresponds to the polarization of charge illustrated in Fig. 4.12(b) with positive charge accumulating on the right side of the sphere and negative charge on the left side. The polarization of charge indicated in Fig. 4.12(b) induces an electric field $\boldsymbol{E}_{\text{ind}}$ that opposes the applied field $\boldsymbol{E}_0$ but does not completely cancel it as would happen for a conducting filled sphere. The electric field $\boldsymbol{E}_{\text{in}}$ inside the sphere is given by Eq. (4.73) and is constant. If $\varepsilon > \varepsilon_0$ the electric field inside acts in the opposite direction as the applied field and is reduced in strength by a factor $3/(\varepsilon/\varepsilon_0 + 2)$ relative to the applied field. This is indicated schematically in Fig. 4.13(a), where we notice that

1. The field inside $\boldsymbol{E}_{\text{in}}$ is in the same direction as the applied field $\boldsymbol{E}_0$, but is reduced in strength by the induced field $\boldsymbol{E}_{\text{ind}}$ acting against the applied field.
2. This reduction in strength for the interior field is indicated by the decreased density of field lines inside the sphere relative to the outside.
3. There is a discontinuity of the electric field lines at the boundary of the sphere because of the surface charge that has accumulated there because of polarization.

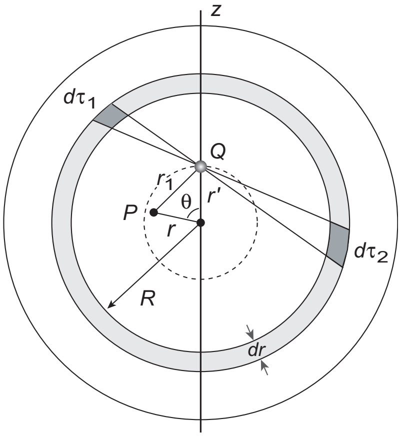
As indicated in Eqs. (4.74) and (4.75), the external potential consists of the original potential contributed by the external field plus a correction term that may be viewed as the potential generated by a set of electric dipoles centered on the sphere because of the charge polarization. Far from the sphere the external potential is that of the original applied field $\boldsymbol{E}_0$, but near the sphere the field lines in Fig. 4.13(a) are strongly distorted by the increasing importance of the second term in Eq. (4.75), which scales as r^{-2} and therefore grows rapidly as the sphere is approached.

Background and Further Reading

Good introductions to the material of this chapter may be found in Griffiths [8]. More advanced treatments may be found in Jackson [15], Garg [6], Chaichian et al [3], and Zangwill [27].

Problems

4.1 Consider a sphere containing a single point charge Q located a distance r' from the origin along the z-axis.



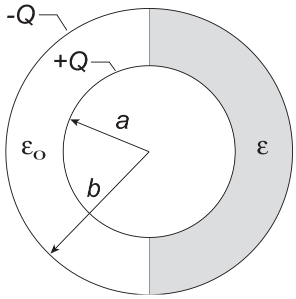
Calculate the average electric field inside the sphere. *Hint*: Separate the volume of the sphere τ into the part outside Q plus the part inside Q, as indicated by the dashed circle in the figure, and show by qualitative argument that the contribution from outside the dashed circle is zero.

4.2 Assume a sphere of radius R parameterized by spherical coordinates (r, θ, ϕ), with a surface charge layer of density

$$\sigma(\theta) = kP_1(\cos\theta) = k\cos\theta,$$

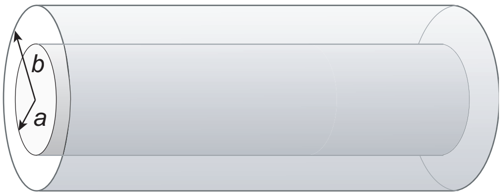
where k is a constant and $P_1(\cos\theta)$ is a Legendre polynomial. Use the Laplace equation assuming axial symmetry (no ϕ dependence) to find expressions for the potential $\Phi(r, \theta)$ inside and outside of R. *Hint*: This will be easy if you do Problem 3.5 first.

4.3 Two concentric conducting spheres of inner and outer radii a and b, respectively, carry charges $\pm Q$. The empty space between the spheres is half-filled by a hemispherical shell of dielectric having dielectric constant $\varepsilon/\varepsilon_0$.



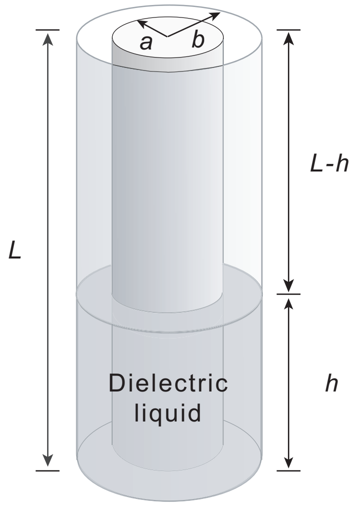
(a) Find the electric field in the region between the spheres. (b) Calculate the surface-charge distribution on the inner sphere. (c) Calculate the polarization-charge density induced on the surface of the dielectric at $r = a$.

4.4 (a) A capacitor consists of concentric cylinders with radii a for the inner cylinder and b for the outer cylinder.



Derive a formula for the capacitance per unit length assuming the inner cylinder to be positively charged with a charge density of λ coulombs per unit length.

(b) Two long coaxial conducting cylinders of radii a and b are lowered vertically into a dielectric liquid.



If the liquid rises an average height of h between the electrodes when a potential difference V is established between them, show that the susceptibility of the dielectric liquid is given by

$$\chi_e = \frac{(b^2 - a^2)\rho g h \ln(b/a)}{\varepsilon_0 V^2},$$

where ρ is the density of the liquid, g is the acceleration due to gravity, and the susceptibility of the air is neglected.

4.5 A parallel-plate capacitor has plates of area A separated by a distance d and the plates are charged to a potential difference V using a battery.

(a) With the charging battery *disconnected* a dielectric sheet that has exactly the same width and length as a plate is inserted between the plates. Find the work done on the dielectric sheet; is it pulled in or must it be pushed in?

(b) Repeat the experiment and analysis of part (a), but with the charging battery *connected* to the plates.

4.6 A point charge q is located in free space a distance d from the center of a dielectric sphere of radius a, with $a < d$. Find the potential at all points in space as an expansion in spherical harmonics with expansion coefficients evaluated.

5 Magnetostatics in Vacuum

Chapters 1-4 have dealt with the subject of electrostatics. The basic goal of electrostatics is that we have some set of discrete source charges $\{q_i\}$, or a localized continuous distribution of charge $\rho(\boldsymbol{x})$, that are *stationary* with respect to a chosen reference frame, and we desire to calculate the electric field at arbitrary locations in space produced by those charges. This can be done using the principle of superposition: calculate the contribution of each source charge q_i or infinitesimal piece of a continuous charge $d\rho(\boldsymbol{x})$ to the electric field at some point, and sum them to get the total electric field at that point. We have developed a number of sophisticated ways to do this beyond brute force summation or integration, but that is the essential idea. In this chapter we wish to expand upon this idea by the (seemingly) elemental extension of allowing the source charges to move.

5.1 Magnetostatics Versus Electrostatics

The introduction of charges in motion (currents) may seem an innocuous change on its surface, but it adds to the electric field associated with the stationary source charges a new *magnetic field* associated with their motion, and a host of associated phenomena (*magnetism*) having a phenomenology that is often very different from that of electrostatics; so much so that it took centuries after electrostatic and magnetic phenomena were first identified in nature to realize that they are not separate subjects but are in fact different manifestations of the same basic physical principles.

Magnetostatics is more subtle and complex than electrostatics. From Maxwell's equations (1.1), a time-independent current distribution $\boldsymbol{J}(\boldsymbol{x})$ is a source of a vector field $\boldsymbol{B}(\boldsymbol{x})$ called the *magnetic field* that satisfies the differential equations[1]

$$\boldsymbol{\nabla}\cdot\boldsymbol{B}(\boldsymbol{x}) = 0, \tag{5.1a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B}(\boldsymbol{x}) = \mu_0\boldsymbol{J}(\boldsymbol{x}). \tag{5.1b}$$

Applying the divergence operation to both sides of Eq. (5.1b) using the vector identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0$ of Eq. (A.6) gives

$$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{B}(\boldsymbol{x})) = 0 = \mu_0\boldsymbol{\nabla}\cdot\boldsymbol{J}(\boldsymbol{x}). \tag{5.2}$$

Thus, magnetostatic current densities satisfy $\boldsymbol{\nabla}\cdot\boldsymbol{J}(\boldsymbol{x}) = 0$. In electrostatics stationary charges produce *electric fields constant in time*; in magnetostatics stationary (unchanging) currents

[1] Equations (5.1) specify the divergence and the curl of the vector field $\boldsymbol{B}(\boldsymbol{x})$. Therefore, this should be sufficient to specify $\boldsymbol{B}(\boldsymbol{x})$ uniquely by the Helmholtz theorem of Box 3.1.

produce *magnetic fields that are constant in time*. The more complex nature of magnetostatics relative to electrostatics arises in part because magnetostatics involves a current density that is a vector and the force law [given in Eq. (5.4) below] involves a cross product of vectors, in contrast to the scalar quantities and operations inherent in electrostatics. From a physics perspective magnetism is also complicated by there being two fundamentally different types of currents that can produce magnetic effects:

1. currents that results from moving charges, and
2. currents that result from the quantum spins of point-like particles (*magnetization currents*), which have no suitable classical analogs.[2]

It follows that *magnetizable matter* exhibits much greater variety than *polarizable matter* (Ch. 4), both in its fundamental attributes and in the way that it responds to external fields.

> For example, permanent magnetism (*ferromagnetism*), caused by spontaneous breaking of angular momentum symmetry by macroscopic alignment of spins, is much more common than than the dielectric counterpart of *ferroelectricity* [27].

From the static (all time derivatives set to zero) Maxwell equations (1.1), the basic equations governing electrostatics are

$$\nabla \times \boldsymbol{E}(\boldsymbol{x}) = 0 \tag{5.3a}$$

$$\nabla \cdot \boldsymbol{E}(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x})}{\varepsilon_0}. \tag{5.3b}$$

Comparing with the basic equations (5.1) governing magnetostatics, we see that the formal roles of the divergence and curl operators are interchanged between electrostatics and magnetostatics. As a purely practical matter, this leads to methods and corresponding results for magnetostatics that are quite different from the methods and corresponding results in electrostatics.

Finally, though, let us note that from a fundamental point of view special relativity tells us that (despite the differences described above) the distinction between electric $\boldsymbol{E}$ fields and magnetic $\boldsymbol{B}$ fields amounts to nothing more than a choice of observer reference frame:

> What looks like an electric field from one inertial frame looks like a magnetic field from a different inertial frame.

We shall take that up in later chapters when we discuss the special theory of relativity, Lorentz transformations, and the formulation of the Maxwell equations in a manifestly Lorentz-covariant manner. For now we note that in the low-energy world[3] there is utility in using a formalism that distinguishes between electric and magnetic phenomena through the use of equations that are (secretly) Lorentz covariant, but are not manifestly so.

[2] Quantum-mechanical orbital angular momentum can be viewed semiclassically as charge in motion on a classical orbit (think of the Bohr model of the hydrogen atom), but quantum spin angular momentum has no corresponding classical analog that is completely faithful to the underlying quantum physics.

[3] Although it is not completely apparent from looking at the equations in SI units without examining the con-

5.2 Magnetic Forces

A magnetic field $\boldsymbol{B}$ generated by source charges in motion and the electric field $\boldsymbol{E}$ associated with those charges lead to a force $\boldsymbol{F}$ that is found to act on a test charge q having relative velocity v according to the *Lorentz force law*,

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{5.4}$$

As we will find, the very different phenomenology of magnetic effects relative to electrostatic effects is partially due to the nature of this force law, which mixes the effect of the electric field and magnetic field in a non-trivial way. The characteristic motion of a charged particle in a magnetic field is circular, as illustrated in Box 5.1 and Fig. 5.1. Two physical consequences of the force law (5.4) are illustrated in Box 5.1: cyclotron orbits and the classical Hall effect.

How much work can be done by the Lorentz force? We calculated previously in Eq. (2.38) that the work done if a test charge is moved in an electric field is the product of the charge and the difference in electric potential over the path. If we repeat that calculation using the Lorentz force (5.4) with $\boldsymbol{B} = 0$ the same result is obtained (which is reassuring!). On the other hand, let's set $\boldsymbol{E} = 0$ in Eq. (5.4) and calculate the work done by the magnetic part of the Lorentz force on a test charge. If a charge Q moves a distance $d\boldsymbol{L} = \boldsymbol{v}dt$, then the amount of work done by the magnetic field is

$$dW = \boldsymbol{F} \cdot d\boldsymbol{L} = Q(v \times \boldsymbol{B}) \cdot \boldsymbol{v}\, dt = 0. \tag{5.5}$$

The magnetic field does no work, which is obviously because the cross product $\boldsymbol{v} \times \boldsymbol{B}$ is perpendicular to the velocity $\boldsymbol{v}$, so the scalar product $(v \times \boldsymbol{B}) \cdot \boldsymbol{v}$ must vanish. As exemplified in the cyclotron motion discussed in Box 5.1, magnetic forces can *alter the direction* of charged-particle motion, but they *cannot change its speed* (magnitude of the velocity).

5.3 The Law of Biot and Savart

The magnetic field $\boldsymbol{B}(\boldsymbol{x})$ of a steady current described by the density $\boldsymbol{J}(\boldsymbol{x})$ is given by the empirical *Biot–Savart law*,[4]

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{J}(\boldsymbol{x}') \times (\boldsymbol{x} - \boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|^3}\, d^3x', \tag{5.6}$$

where the *permeability of free space* μ_0 is defined in Eq. (1.2). When a steady current is flowing in a 1D wire the magnitude of the current I must be constant. If $\boldsymbol{L}$ is a vector that

stants carefully, the magnetic field is intrinsically much weaker than the electric field under typical laboratory conditions. But this is largely because the typical speeds of charged particles are much less than the speed of light in those circumstances. If the velocities of particles approach the speed of light magnetic forces approach electric forces in intrinsic strength.

[4] The SI units for $\boldsymbol{B}$ are *teslas* (T), where $1\,\mathrm{T} \equiv 1\,\mathrm{N\,A^{-1}\,m^{-1}}$. It is common also to use the CGI unit of *gauss* for $\boldsymbol{B}$, where $1\,\text{tesla} = 10^4\,\text{gauss}$.

Box 5.1 Motion of Charged Particles in Magnetic and Electric Fields

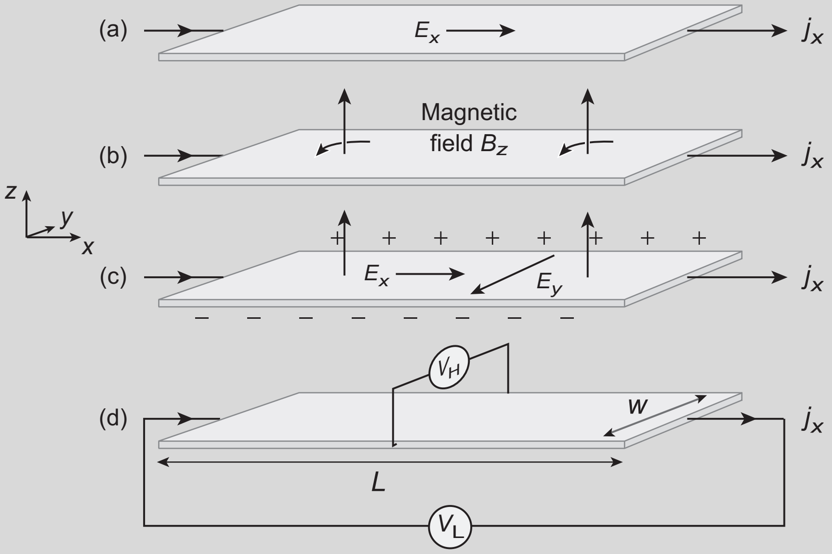
Characteristic *cyclotron motion* of a charged particle in a magnetic field is circular with the centripetal acceleration that curves the path provided by the magnetic force.

Cyclotron Motion

As illustrated in Fig. 5.1(a), the magnetic field $\boldsymbol{B}$, oriented into the page, produces through the $q\boldsymbol{v} \times \boldsymbol{B}$ component of the Lorentz force (5.4) a centripetal force directed toward the center of the circle for a particle of charge q and velocity $\boldsymbol{v}$, causing the particle to be deflected in a circular path without changing its speed (implying that the magnetic field does no work). The motion becomes more complex if an electric field is present also. For an electric field perpendicular to the magnetic field, the circular cyclotron motion due to the magnetic field is converted into a the spiral motion depicted in Fig. 5.1(b) by the $q\boldsymbol{E}$ component of the Lorentz force.

The Hall Effect

A scientifically important consequence of the Lorentz force is exemplified by the *classical Hall effect*, depicted in the following diagram.



(a) For a 2D sample an electric field E_x causes a current density j_x in the x direction. (b) A uniform magnetic field $\boldsymbol{B}$ placed on the sample in the $+z$ direction deflects electrons in the $-y$ direction. (c) Negative charge accumulate on one edge and a positive charge excess on the other edge, producing a transverse electric field E_y called the *Hall field* that just cancels the force from the magnetic field; thus in equilibrium current flows only in the x direction. (d) Typically the longitudinal voltage V_{L} and the transverse Hall voltage V_{H} are measured.

The *Hall resistance* R_{H} (inferred from $R_{\mathrm{H}} = V_{\mathrm{H}}/wj_x$) depends on the sign and density of charge carriers. Thus the classical Hall effect is important as a diagnostic for charge carriers in materials samples. At very high magnetic fields, quantum effects become significant. These *quantum Hall effects* are of even greater importance, since they first revealed topological effects that were the harbingers of the modern topological matter revolution in condensed matter and materials science.

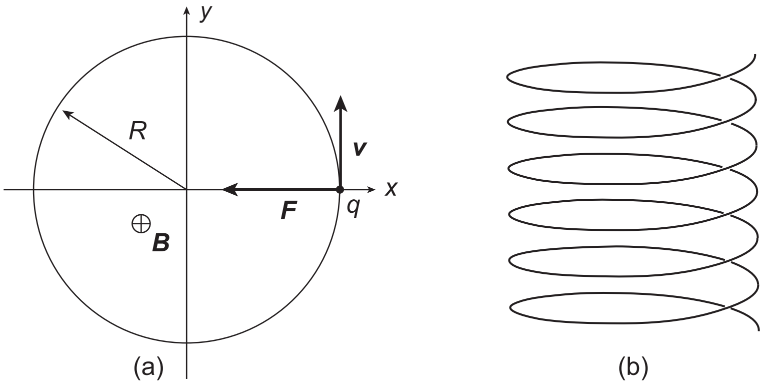


Fig. 5.1 (a) Cyclotron motion caused by magnetic field $\boldsymbol{B}$ pointing into the page along the z axis (not shown) acting on a charge q with a velocity $\boldsymbol{v}$, as explained in Box 5.1. (b) An electric field $\boldsymbol{E}$ perpendicular to the magnetic field converts circular cyclotron motion into spiral motion because of the Lorentz force law (5.4).

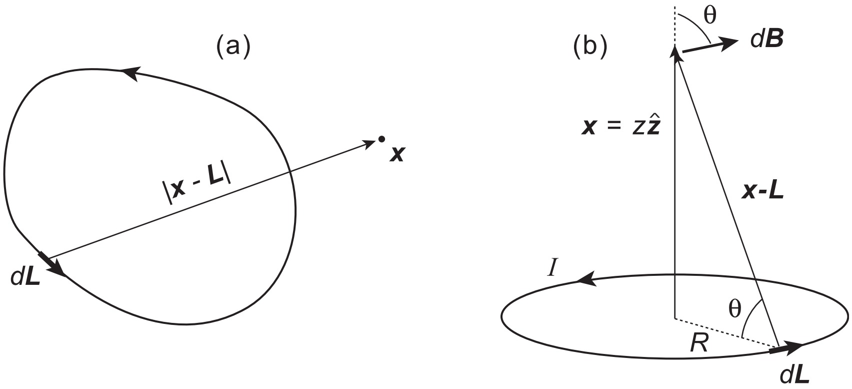


Fig. 5.2 (a) Current loop carrying a steady current I for Eq. (5.7). (b) Geometry for Biot–Savart calculation of the magnetic field on the symmetry axis of a circular current loop.

points to a line element $d\boldsymbol{L}$ of the wire as in Fig. 5.2(a), substitution of $Id\boldsymbol{L}$ for $\boldsymbol{j}d^3x$ in Eq. (5.6) yields the Biot–Savart law in the form

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0 I}{4\pi}\int \frac{d\boldsymbol{L}\times(\boldsymbol{x}-\boldsymbol{L})}{|\boldsymbol{x}-\boldsymbol{L}|^3}. \tag{5.7}$$

for a steady current.

Example 5.1 Let us use the Biot–Savart law (5.7) to calculate the magnetic field produced along the z axis by the circular current loop in Fig. 5.2(b). The components of $d\boldsymbol{B}$ that are perpendicular to the z symmetry axis cancel one another when the entire loop is traversed, but the z components add with the same magnitude for each line increment $d\boldsymbol{L}$, which gives

$$\boldsymbol{B}(z)=\hat{\boldsymbol{z}}\frac{\mu_0}{4\pi}\frac{\cos\theta}{R^2+z^2}\oint dL=\hat{\boldsymbol{z}}\frac{\mu_0 I}{2}\frac{R^2}{(R^2+z^2)^{3/2}} \tag{5.8}$$

This non-zero value of $\boldsymbol{B}$ on the symmetry axis contrasts with the value of zero for the electric field $\boldsymbol{E}$ found for a uniformly charged ring. The difference follows from the cross product in the Biot–Savart formula that isn't present in the electric field formula. This causes contributions to $\boldsymbol{B}(z=0)$ from opposite sides of the ring to add to each other for the magnetic field, but to subtract and cancel each other for the electric field. This example is one illustration of the basic differences between the way that stationary charges produce electric fields and the way that charges in motion produce magnetic fields.

Both Coulomb's law and the Biot–Savart law are empirical, with each tailored to account for the corresponding electrostatic and magnetostatic data, respectively.

> The Biot–Savart law may be viewed as the starting point for magnetostatics, just as Coulomb's law may be viewed as the starting point for electrostatics.

Both laws exhibit a one over distance squared dependence, but otherwise they differ substantially because of the vector character of the magnetic law.

5.4 Differential Form of the Biot–Savart Law

The Biot–Savart law (5.6) in integral form,

$$\boldsymbol{B}(\boldsymbol{x})=\frac{\mu_0}{4\pi}\int\frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x',$$

contains (in principle) a complete description of magnetostatics, but is not always the most convenient form for solving problems. In many situations a differential equation is more convenient to use. Let us find a form of the Biot–Savart law expressed as a differential equation, following the presentation in Jackson [15]. From Eq. (A.11),

$$\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}=-\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)=\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right), \tag{5.9}$$

which allows converting the Biot–Savart equation (5.6) into

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{x})&=\frac{\mu_0}{4\pi}\int\frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x'\\&=\frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',\end{aligned} \tag{5.10}$$

where $\boldsymbol{\nabla}$ has been pulled out of the integral because it operates on $\boldsymbol{x}$ but not $\boldsymbol{x}'$.[5] Now take the divergence of Eq. (5.10)

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'.$$

But from Eq. (A.6) we have the identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0$ so

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0, \tag{5.11}$$

which may be termed the *first law of magnetostatic*s [and is the third Maxwell equation (1.1c), corresponding to the absence of magnetic charges].

Next, take the curl of $\boldsymbol{B}$ in Eq. (5.10),

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{5.12}$$

Using the vector identity $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \nabla^2\boldsymbol{A}$ of Eq. (A.8), this may be written as

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{B} &= \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int\boldsymbol{J}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x' - \frac{\mu_0}{4\pi}\int\boldsymbol{J}(\boldsymbol{x})\nabla^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x' \\ &= -\frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int\boldsymbol{J}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x' + \mu_0\boldsymbol{J}(\boldsymbol{x}),\end{aligned} \tag{5.13}$$

where we have used the identities of Eq. (A.11),

$$\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) \qquad \nabla^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}') \tag{5.14}$$

(where $\boldsymbol{\nabla}$ operates on $\boldsymbol{x}$ and $\boldsymbol{\nabla}'$ operates on $\boldsymbol{x}'$) in the last step. Integrating the remaining integral in Eq. (5.13) by parts then gives

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J} + \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int\frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{5.15}$$

But for steady-state magnetism $\boldsymbol{\nabla}\cdot\boldsymbol{J} = 0$ and we obtain finally

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J}, \tag{5.16}$$

which may be termed the *second law of magnetostatics* [and is the fourth Maxwell equation (1.1d) if electric fields don't depend on time (Ampère's law)].

The integral equivalent of Ampère's law (5.16) may be obtained from Stokes' theorem (2.20),

$$\int_S(\boldsymbol{\nabla}\times\boldsymbol{A})\cdot\boldsymbol{n}\,da = \oint_C\boldsymbol{A}\cdot d\boldsymbol{l},$$

for the vector field $\boldsymbol{A}$ where S is an arbitrary open surface bounded by a closed curve C and where $\boldsymbol{n}$ is the normal to S. Figure 5.3 illustrates. Applying Stokes' theorem to Eq. (5.16),

[5] Remember that in these manipulations the integrations are over the *primed coordinates*, but applied gradient, divergence, and curl operations [$\boldsymbol{\nabla}$, $\boldsymbol{\nabla}\cdot$, and $\boldsymbol{\nabla}\times$, respectively, without primes] are taken with respect to the *unprimed coordinates*. This is made explicit in the notation exemplified in Eqs. (A.11) of Appendix A, where $\boldsymbol{\nabla}\equiv\boldsymbol{\nabla}_x$ operates on the $\boldsymbol{x}$ coordinates while $\boldsymbol{\nabla}'\equiv\boldsymbol{\nabla}_{x'}$ operates on the $\boldsymbol{x}'$ coordinates.

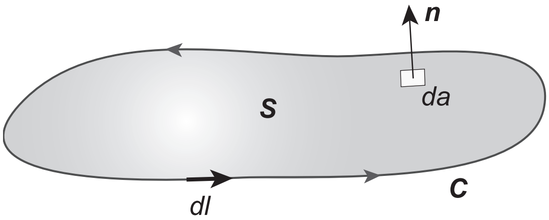


Fig. 5.3 Surface S and contour C bounding the surface for Stokes' theorem. An infinitesimal line element on C is indicated by dl and an infinitesimal surface element on S is indicated by da, with the normal to da indicated by $\boldsymbol{n}$. The path in the line integration is traversed in a right-hand screw sense relative to $\boldsymbol{n}$, as indicated by arrows.

$$\int_S (\nabla \times \boldsymbol{B}) \cdot \boldsymbol{n}\, da = \oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \boldsymbol{n}\, da \tag{5.17}$$

and therefore Eq. (5.16) becomes

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \boldsymbol{n}\, da. \tag{5.18}$$

Using that the total current I passing through the closed curve C is given by the surface integral on the right side of Eq. (5.18),

$$I = \int_S \boldsymbol{J} \cdot \boldsymbol{n}\, da, \tag{5.19}$$

we finally write Ampère's law in the integral form

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I, \tag{5.20}$$

We found in our study of electrostatics that Gauss's law can often be used to find the electric field in highly symmetric cases. Ampére's law can be employed in an analogous way for magnetostatic problems with high symmetry. The following example illustrates.

Example 5.2 Let's determine the magnetic field for the long solenoid illustrated in Fig. 5.4, with n closely wound turns per unit length on a cylinder of radius R and carrying a steady current I. Because the solenoid is tightly wound, we expect that each coil can be assumed to be perpendicular to the symmetry axis of the cylinder. We expect then on symmetry and general grounds that the magnetic field field is oriented along the cylinder axis, and that it drops to zero at large distances from the solenoid. Let's apply Ampère's law (5.20) to the two rectangular loops shown in the diagram. Loop 1 is completely outside the solenoid and encloses no current, $I_{\text{enc}} = 0$, with its left side a distance a and its right side a distance b from the central axis of the cylinder. Applying Ampère's law to it

$$\oint \boldsymbol{B} \cdot dl = [B(a) - B(b)]L = \mu_0 I_{\text{enc}} = 0,$$

where $B = |\boldsymbol{B}|$. So $B(a) = B(b)$ and the magnetic field outside is independent of the distance from the solenoid. But the boundary conditions require that $B = 0$ at infinity, so the

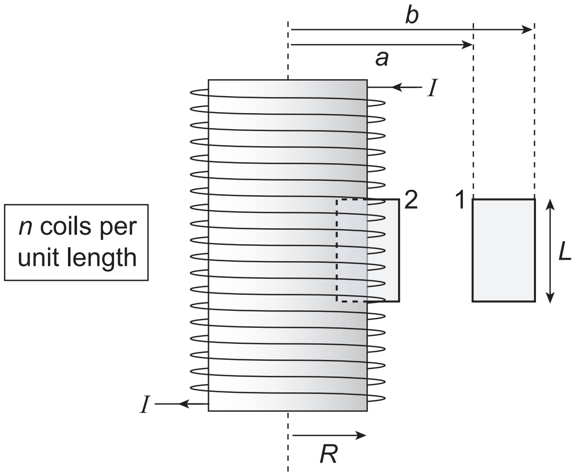


Fig. 5.4 Analysis of a long, tightly wound solenoid using Ampère's law. Two rectangular loops are shown, number 1 outside the solenoid and number 2 overlapping the solenoid.

magnetic field must *vanish everywhere outside the solenoid*. Loop 2 is halfway inside the solenoid and Ampère's law gives

$$\oint \boldsymbol{B}\cdot dl = BL = \mu_0 I_{\text{enc}} = \mu_0 nIL,$$

where B is the field inside the solenoid (there is no contribution from the half rectangle outside the cylinder since $B = 0$ there). Thus inside the solenoid the field is uniform, $\boldsymbol{B} = \mu_0 nI\hat{\boldsymbol{z}}$, where $\hat{\boldsymbol{z}}$ is a unit vector along the cylinder axis, and outside the solenoid the magnetic field vanishes.

Like Gauss's law, Ampère's law is generally valid (for steady currents), but it is *useful* only if a problem has sufficient symmetry to allow B to be pulled out of the integral $\oint \boldsymbol{B}\cdot d\boldsymbol{l}$, allowing Eq. (5.20) to be solved easily for the magnetic field..

5.5 The Vector Potential and Gauge Invariance

We have seen in the preceding section that the basic laws of magnetostatics are defined in differential form through Eqs. (5.16) and (5.11),

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{B} &= \mu_0\boldsymbol{J},\\ \boldsymbol{\nabla}\cdot\boldsymbol{B} &= 0,\end{aligned} \tag{5.21}$$

which must be solved for the magnetic field $\boldsymbol{B}$. In electrostatics we found that the electric potential Φ was an extremely useful quantity because the electric field can be derived from it by taking the gradient, $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$. The electric potential Φ is a scalar quantity, and it is often termed the *scalar potential* [we have indeed used the terms "(electric) potential" and "scalar potential" interchangeably]. For the special case that the current density is zero in a region, it is possible to define a *magnetic scalar potential* Φ_{M} such that the magnetic field is given by $\boldsymbol{B} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}}$. Then Eq. (5.16) with $\boldsymbol{J} = 0$ reduces to the Laplace equation for Φ_{M}. Therefore, the methods for solving Laplace's equation developed for electrostatics in Chs. 2-4 become applicable. However, this approach has limited applicability because it is valid only if the current density vanishes. As we now discuss, an approach with more general applicability may be developed by exploiting the second equation above, $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$. This will allow defining a *vector potential* $\boldsymbol{A}$, from which the magnetic field $\boldsymbol{B}$ can be derived by taking the curl of $\boldsymbol{A}$.

We begin by noting that if $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ is to hold everywhere, than $\boldsymbol{B}$ must be the curl of some vector field $\boldsymbol{A}$ (the vector potential),

$$\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{\nabla}\times\boldsymbol{A}(\boldsymbol{x}), \tag{5.22}$$

since then the identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0$ ensures that the divergence of $\boldsymbol{B}$ vanishes identically under all conditions. In fact, $\boldsymbol{B}$ was already written in this form in Eq. (5.10),

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \tag{5.23}$$

and upon comparing Eqs. (5.22) and (5.23), a vector potential $\boldsymbol{A}$ consistent with the phenomenology of magnetism takes the general form

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x' + \boldsymbol{\nabla}\psi(\boldsymbol{x}), \tag{5.24}$$

where the addition of the *arbitrary scalar function* $\psi(\boldsymbol{x})$ has no effect on $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ because of the identity $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}f) = 0$ [Eq. (A.7)] for an arbitrary scalar function f. Since $\psi(\boldsymbol{x})$ is an arbitrary scalar function of $\boldsymbol{x}$, the vector potential $\boldsymbol{A}$ can be transformed freely according to

$$\boldsymbol{A}\to\boldsymbol{A}+\boldsymbol{\nabla}\psi, \tag{5.25}$$

which has no effect on the magnetostatic equation $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ because identically $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{\nabla}\chi) = 0$.

Gauge transformations: The addition of the gradient of an arbitrary scalar function ψ to the vector potential $\boldsymbol{A}\to\boldsymbol{A}+\boldsymbol{\nabla}\psi$ is called a *gauge transformation* and the invariance of the laws of electromagnetism under this transformation is called *gauge invariance*. The *gauge symmetry* associated with this invariance has large implications for classical electromagnetism and quantum electrodynamics, and a generalization of this gauge symmetry is of fundamental importance in relativistic quantum field theories, particularly for the Standard Model of elementary particle physics. We will elaborate on electromagnetism as the prototype gauge field theory in Ch. 10.

From the Helmholtz theorem (Box 3.1), a vector field with suitable boundary conditions is

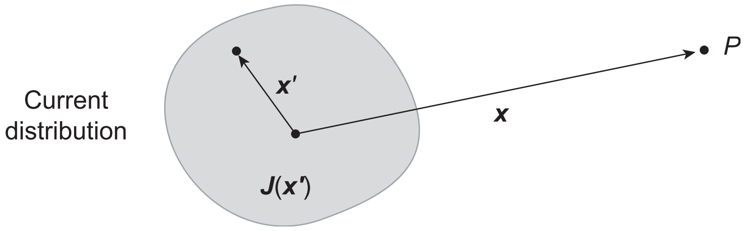


Fig. 5.5 A localized current density $\boldsymbol{J}(\boldsymbol{x}')$ that produces a vector potential $\boldsymbol{A}(\boldsymbol{x})$ and corresponding magnetic field $\boldsymbol{B}(\boldsymbol{x})$ at the point $\boldsymbol{x}$, with $\boldsymbol{x} \gg \boldsymbol{x}'$.

specified uniquely by its curl and its divergence. From Eq. (5.22), specifying the magnetic field requires only the curl of $\boldsymbol{A}$. Thus we are free to make gauge transformations in the vector potential such that $\boldsymbol{\nabla}\cdot\boldsymbol{A}$ has any convenient functional form without affecting the magnetic field and thus without altering the physics of electromagnetism.

Substituting $\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A}$ into Eq. (5.16) gives

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) = \mu_0\boldsymbol{J},$$

which becomes

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \nabla^2\boldsymbol{A} = \mu_0\boldsymbol{J}, \tag{5.26}$$

upon invoking the identity $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \nabla^2\boldsymbol{A}$ given in Eq. (A.8). Now we exploit the freedom to make a gauge transformation (5.25) by choosing a gauge where

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0 \qquad \text{(Coulomb gauge condition)}, \tag{5.27}$$

which defines the *Coulomb gauge* (also know as the *radiation gauge* or the *transverse gauge*, for reasons that will be explained later). In Coulomb gauge, Eq. (5.26) is transformed into the Poisson equation,

$$\nabla^2\boldsymbol{A} = -\mu_0\boldsymbol{J} \qquad \text{(Coulomb gauge)}. \tag{5.28}$$

From application of the Poisson equation in electrostatics we expect that the solution for $\boldsymbol{A}$ in Coulomb gauge is given by the first term of Eq. (5.24), with $\psi = \text{constant}$,

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x' \tag{5.29}$$

if the $\boldsymbol{J}(\boldsymbol{x}') \to 0$ sufficiently rapidly at infinity. Gauge transformations will be discussed in more depth in Ch. 7.

5.6 Magnetic Fields of Localized Current Distributions

We now consider a current distribution that is localized in a region of space small compared to the length scale of interest to an observer. Figure 5.5 illustrates. A full treatment of this problem by analogy with electric multipole expansions is possible using *vector spherical* *harmonics*, which are described briefly in Box 5.2. We shall avoid using them and confine ourselves instead to lowest-order expansions.

Let us assume that $\boldsymbol{x} \gg \boldsymbol{x}'$ and expand the denominator of Eq. (5.29) in powers of $\boldsymbol{x}'$ relative to a suitable origin located in the current distribution in Fig. 5.5,

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \frac{1}{|\boldsymbol{x}|} + \frac{\boldsymbol{x}\cdot\boldsymbol{x}'}{|\boldsymbol{x}|^2} + \cdots . \tag{5.30}$$

Thus the *i*th components of the vector potential $\boldsymbol{A}(\boldsymbol{x})$ has the expansion

$$A_i(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left[\frac{1}{|\boldsymbol{x}|}\int J_i(\boldsymbol{x}')\,d^3x' + \frac{\boldsymbol{x}}{|\boldsymbol{x}|^2}\cdot\int J_i(\boldsymbol{x}')\boldsymbol{x}'\,d^3x' + \cdots\right]. \tag{5.31}$$

That the current *current*$(\boldsymbol{x}')$ is localized and has zero divergence permits simplification of this expansion. The first term vanishes because the integral of the current vector over all orientations will average to zero. (This argument is intuitive; a more formal proof that the first term vanishes is given in Jackson [15], Section 5.6.) This reflects that the first term in the multipole expansion is the monopole term, which measures the total “charge”; but there are no magnetic monopoles because there is no magnetic charge consistent with the Maxwell equations ($\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$). The integral in the second term of Eq. (5.31) can be manipulated into

$$\begin{aligned}
\boldsymbol{x}\cdot\int \boldsymbol{x}' J_i(\boldsymbol{x}')\,d^3x' &= \sum_i x_i \int x'_j J_i\,d^3x' \\
&= -\frac{1}{2}\sum_i x_i \int (x'_i J_i - x'_j J_i)\,d^3x' \\
&= -\frac{1}{2}\sum_{j,k} \varepsilon_{ijk}\, x_j \int (\boldsymbol{x}'\times\boldsymbol{J})_k\,d^3x' \\
&= -\frac{1}{2}\left[\boldsymbol{x}\times\int (\boldsymbol{x}'\times\boldsymbol{J})\,d^3x\right]_i .
\end{aligned} \tag{5.32}$$

The *magnetic moment density* or *magnetization* is defined by

$$\boldsymbol{M}(\boldsymbol{x}) = \frac{1}{2}\left[\boldsymbol{x}\times\boldsymbol{J}(\boldsymbol{x})\right], \tag{5.33}$$

and the *magnetic moment* $\boldsymbol{m}$ is defined by

$$\boldsymbol{m} = \frac{1}{2}\int \boldsymbol{x}'\times\boldsymbol{J}(\boldsymbol{x}')\,d^3x'. \tag{5.34}$$

Then the multipole expansion (5.31) can be written as

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\boldsymbol{x}}{|\boldsymbol{x}|^3} + \text{higher-order multipole terms.}$$

The first non-vanishing term in this multipole expansion for a steady-state current distribution has the form of a *magnetic dipole*. The higher-order multipoles have increasingly larger powers of $|\boldsymbol{x}|$ in their denominators and at large distances from the current source

Box 5.2 Vector Spherical Harmonics

Vector spherical harmonics are an extension of regular (scalar) spherical harmonics designed for use with vector fields. Generally the components of vector spherical harmonics are complex-valued functions expressed in terms of spherical basis vectors. Following the conventions of Ref. [2], three vector spherical harmonics may be defined

$$\boldsymbol{Y}_{lm} = Y_{lm}\hat{\boldsymbol{r}} \qquad \boldsymbol{\Psi}_{lm} = r\boldsymbol{\nabla} Y_{lm} \qquad \boldsymbol{\Phi}_{lm} = \boldsymbol{r}\times\boldsymbol{\nabla} Y_{lm},$$

where $\boldsymbol{r}$ is the radial vector in spherical coordinates. The purpose of the radial factors is to ensure that the vector spherical harmonics have the same dimensions as ordinary spherical harmonics, and that they do not depend on the radial coordinate. These new vector fields facilitate separation of radial from angular coordinates in spherical coordinates, permitting a a multipole expansion of the electric field $\boldsymbol{E}$ of the form

$$\boldsymbol{E} = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l}\left(E^r_{lm}(r)\boldsymbol{Y}_{lm} + E^{(1)}_{lm}(r)\boldsymbol{\Psi}_{lm} + E^{(2)}_{lm}(r)\boldsymbol{\Phi}_{lm}\right),$$

The component labels indicate that $E^r_{lm}(r)$ is the radial component, and $E^{(1)}_{lm}(r)$ and $E^{(2)}_{lm}(r)$ are transverse components of the vector field (with respect to the vector $\boldsymbol{r}$).

the vector potential can be approximated well by

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\boldsymbol{x}}{|\boldsymbol{x}|^3}. \tag{5.35}$$

Then the magnetic field may be found by taking the curl of $\boldsymbol{A}$, giving

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left(\frac{3\boldsymbol{n}(\boldsymbol{n}\cdot\boldsymbol{m})-\boldsymbol{m}}{|\boldsymbol{x}|^3}\right), \tag{5.36}$$

where $\boldsymbol{n}$ is a unit vector in the direction of $\boldsymbol{x}$. Because the dipole term dominates the multipole expansion in typical situations, it is common to refer to the magnetic dipole moment as simply the magnetic moment.

If the current as assumed to correspond to an arbitrary closed loop confined to a plane the magnitude of the magnetic moment takes a simple form that is independent of the shape of the loop,

$$|\boldsymbol{m}| = I\times A, \tag{5.37}$$

where A is the area enclosed by the planar current loop. If the current distribution is due to charged particles with charges q_i, masses M_i, and velocities $\boldsymbol{v}_i$, the current density is

$$\boldsymbol{J} = \sum_i q_i\boldsymbol{v}_i\delta(\boldsymbol{x}-\boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ is the position of the *i*th particle, and the magnetic moment is

$$\boldsymbol{m} = \frac{1}{2}\sum_i q_i(\boldsymbol{x}_i \times \boldsymbol{v}_i).$$

But the orbital angular momentum of particle i is $\boldsymbol{L}_i = M_i(\boldsymbol{x}_i \times \boldsymbol{v}_i)$ and the magnetic moment becomes

$$\boldsymbol{m} = \sum_i \frac{q_i}{2M_i}\boldsymbol{L}_i. \tag{5.38}$$

If all the particles have the same charge-to-mass ratio $q_i/M_i = e/M$,

$$\boldsymbol{m} = \frac{e}{2M}\sum_i \boldsymbol{L}_i = \frac{e}{2M}\boldsymbol{L}, \tag{5.39}$$

where $\boldsymbol{L}$ is the total orbital angular momentum. Equation (5.39) is the well-known classical connection between angular momentum and magnetic moment. It holds for orbital angular momentum even on the atomic scale, but fails for intrinsic moments arising from particle spin, which is an intrinsically quantum effect that is beyond the scope of our present discussion of classical electromagnetism.

Background and Further Reading

Good introductions to the material of this chapter may be found in Griffiths [8]. More advanced treatments may be found in Jackson [15], Garg [6], Chaichian et al [3], and Zangwill [27].

Problems

5.1 A localized current distribution of cylindrical symmetry has a current flowing only in the azimuthal direction, with a current density $\boldsymbol{J} = J(r,\theta)\hat{\boldsymbol{\phi}}$. The current is zero both outside the cylinder and near the origin. Working in Coulomb gauge, show that the corresponding vector potential $\boldsymbol{A}$ has only an azimuthal component with

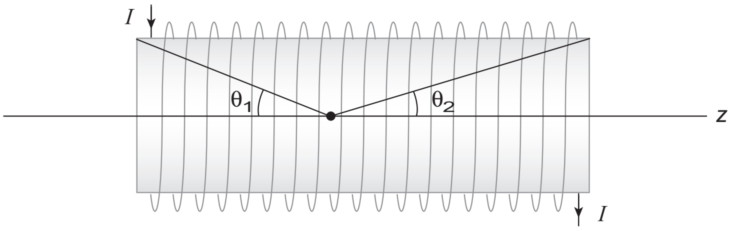
$$A_\phi^{\text{in}}(r,\theta) = -\frac{\mu_0}{4\pi}\sum_l m_l r^l P_l^1(\cos\theta),$$

where $P_l^1(\cos\theta)$ is an associated Legendre polynomial and the multiple moments are given by

$$m_l = \begin{cases} -\dfrac{1}{l(l+1)}\displaystyle\int r^{-l-1}P_l^1(\cos\theta)J(r,\theta)\,d^3x & (\text{ interior}), \\ -\dfrac{1}{l(l+1)}\displaystyle\int r^l P_l^1(\cos\theta)J(r,\theta)\,d^3x & (\text{exterior}), \end{cases}$$

for the interior and exterior solutions, respectively.

5.2 A cylindrical solenoid of length L and radius a carries a current I through N turns per unit length.



Show that in the limit $NL \to \infty$, the magnetic field is

$$B_z = \frac{\mu_0 NI}{2}(\cos\theta_1 + \cos\theta_2)$$

at the point defined on the cylinder axis by the angles θ_1 and θ_2 in the preceding figure.

5.3

5.4

6 Magnetic Fields in Matter

Just as electric fields can polarize matter, magnetic fields can magnetize matter. In classical electromagnetism all magnetic phenomena have their origin in the motion of electrical charges. At the microscopic level, we may view magnetism as being produced by small current loops (for example, electrons in orbits around nuclei in atoms) that may be modeled as tiny magnetic dipoles. Ordinarily the effects of these dipoles cancel out because of random orientation of atoms, but if a magnetic field is applied to the matter it can cause a net alignment of the dipoles so that the matter becomes magnetically polarized (*magnetized*). For the electric polarization of matter described in Ch. 4, the polarization is usually in the direction of the $\boldsymbol{E}$ field. Magnetic polarization of matter is more complex and varied than electric polarization of matter. For example,

1. *paramagnetic materials* acquire a magnetization in the *same direction* as the applied field $\boldsymbol{B}$, while
2. the magnetization of *diamagnetic materials* is in the direction *opposite the applied field*, and
3. *ferromagnetic materials* can become *permanent magnets*, by retaining their magnetization after the polarizing field has been removed, with the retained magnetization depending on the *entire magnetic history* of the material.

As a consequence, the discussion of magnetized matter is more involved than the discussion of electrically polarized matter.

6.1 Ampère's Law in Magnetically Polarized Matter

In the previous chapter we have dealt with steady-state magnetic fields in a microscopic manner, assuming that the current density $\boldsymbol{J}$ is a known function of position. In macroscopic problems dealing with magnetic effects in materials, this will often not be true. The atomic currents in matter produce rapidly fluctuating current densities on a microscopic scale. Just as discussed in Section 4.7 for electric fields in matter, only averages over small macroscopic volumes are known and only these enter into the classical equations of electromagnetism.

6.1.1 Macroscopic Averaging

The first step in introducing averaging in matter for electric fields in Section 4.7 was to note that the averaging procedure preserves the crucial relation $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$, which ensures that the macroscopic electric field is still derivable from a scalar potential through $\boldsymbol{E} = -\nabla\Phi$. In a similar manner, for magnetic fields the macroscopic averaging procedure leads to the same equation $\nabla \cdot \boldsymbol{B} = 0$ that we introduced in Eq. (5.11). Thus, the averaging procedure preserves the notion of a vector potential $\boldsymbol{A}(\boldsymbol{x})$, from which we can derive the macroscopic magnetic field by taking the curl, $\boldsymbol{B} = \nabla \times \boldsymbol{A}$.

The basic effect of magnetization is to establish currents within a material and on its surface such that these currents produce the field due to magnetization. The average macroscopic magnetization or magnetic (dipole) moment density is

$$\boldsymbol{M}(\boldsymbol{x}) = \sum_i N_i \langle \boldsymbol{m}_i \rangle, \tag{6.1}$$

where $\langle \boldsymbol{m}_i \rangle$ is the magnetic moment averaged over a small volume around the point $\boldsymbol{x}$. In addition to the bulk magnetization we may assume that there is a macroscopic current density $\boldsymbol{J}(\boldsymbol{x})$ produced by the flow of free charge in the medium. Then the vector potential resulting from averaging over a small volume ΔV around $\boldsymbol{x}'$ will take the form

$$\Delta\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left[\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} + \frac{\boldsymbol{M}(\boldsymbol{x}') \times (\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\right]\Delta V, \tag{6.2}$$

where the first term represents the contribution of the flow of free charge and the second term represents the contribution from the magnetic dipoles in the medium described by Eq. (5.35). Letting ΔV tend to d^3x', the total vector potential at $\boldsymbol{x}$ is given by an integral over all space,

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int\left[\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} + \frac{\boldsymbol{M}(\boldsymbol{x}') \times (\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\right]d^3x'. \tag{6.3}$$

The second (magnetization) term can be cast in another form by utilizing Eq. (5.9) to write

$$\int \frac{\boldsymbol{M}(\boldsymbol{x}') \times (\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\, d^3x' = \int \boldsymbol{M}(\boldsymbol{x}') \times \nabla'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) d^3x'.$$

Integration by parts then allows Eq. (6.3) to be written

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\left[\boldsymbol{J}(\boldsymbol{x}') + \nabla' \times \boldsymbol{M}(\boldsymbol{x}')\right]}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x', \tag{6.4}$$

where a surface term has been discarded by assuming $\boldsymbol{M}(\boldsymbol{x}')$ to be localized and well behaved.

6.1.2 The Auxiliary Field $\boldsymbol{H}$ and Constituitive Relations

From the results in the preceding section we see that the magnetization contributes an effective current density

$$\boldsymbol{J}_{\mathrm{M}} = \nabla \times \boldsymbol{M}. \tag{6.5}$$

Then $\boldsymbol{J} + \boldsymbol{J}_{\mathrm{M}}$ plays the role of the effective current such that

$$\nabla \times \boldsymbol{B} = \mu_0(\boldsymbol{J} + \nabla \times \boldsymbol{M}). \tag{6.6}$$

It is then convenient to define a new macroscopic field[1]

$$\boldsymbol{H} \equiv \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M}. \tag{6.7}$$

The introduction of $\boldsymbol{H}$ in the presence of magnetically polarized materials is a matter of convenience, analogous to the introduction in Section 4.7 of $\boldsymbol{D}$ to account for electrically polarized materials in electrostatics. The *fundamental fields* in classical electromagnetism are the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$.[2] The *derived fields* $\boldsymbol{D}$ and $\boldsymbol{H}$ are introduced as a convenient way to take into account the average contributions to charge density and current of the atomic-level charges and currents. Then the macroscopic fields in medium that replace the microscopic fields of Eqs. (5.21) are

$$\begin{aligned} \boldsymbol{\nabla} \times \boldsymbol{H} &= \boldsymbol{J}, \\ \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0, \end{aligned} \tag{6.8}$$

which are analogous to the macroscopic fields for electrostatics in medium

$$\begin{aligned} \boldsymbol{\nabla} \times \boldsymbol{E} &= 0, \\ \boldsymbol{\nabla} \cdot \boldsymbol{D} &= \rho, \end{aligned} \tag{6.9}$$

that were derived in Sections 4.7 and 4.8.

Just as for Eqs. (6.9), the description of macroscopic magnetostatics in Eqs. (6.8) requires constituitive relationships, in this case between the fundamental field $\boldsymbol{B}$ and the derived field $\boldsymbol{H}$. For paramagnetic and diamagnetic materials that are isotropic, the relationship may be assumed linear,

$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{6.10}$$

where the constant μ is characteristic of the medium and is called the *magnetic permeability*. It is also common to characterize the magnetic properties of the medium in terms of the *magnetic susceptibility* χ, which is related to the magnetic permeability by

$$\chi = \left(\frac{\mu}{\mu_0} - 1\right). \tag{6.11}$$

Typically for paramagnetic materials $\mu > 1$ and for diamagnetic materials $\mu < 1$, with μ/μ_0 differing from unity by a part in $\sim 10^5$ for either case. The physical reasons for paramagnetism and diamagnetism are discussed in Box 6.1.

[1] We have routinely termed $\boldsymbol{B}$ the magnetic field. Some authors instead call $\boldsymbol{H}$ the magnetic field, which requires finding another name for $\boldsymbol{B}$. For example, Jackson [15] calls $\boldsymbol{H}$ the *magnetic field* and $\boldsymbol{B}$ the *magnetic induction*. This is a question of terminology, not physics, and arguments can be made from a theoretical or experimental perspective in favor of either naming convention. Because the fundamental fields are $\boldsymbol{E}$ and $\boldsymbol{B}$, and the auxiliary fields $\boldsymbol{D}$ and $\boldsymbol{H}$ are derived quantities, we have adopted the convention of calling $\boldsymbol{B}$ the magnetic field, with no specific name for $\boldsymbol{H}$. Likewise $\boldsymbol{E}$ is termed the electric field and $\boldsymbol{D}$ the displacement field (but the common name "displacement" for $\boldsymbol{D}$ survives for historical, not descriptive, reasons).

[2] This is a classical statement. As we shall discuss briefly in Ch. 7, in quantum mechanics one finds that the scalar potential Φ and the vector potential $\boldsymbol{A}$ (or a 4-vector having Φ and $\boldsymbol{A}$ as components in a Lorentz-invariant theory) should be viewed as more fundamental than the electric and magnetic fields, for two basic reasons. (1) There are experiments where probes that never see the magnetic field are influenced by the vector potential (the Aharonov–Bohm effect), and (2) the fundamental coupling of electromagnetism to charged particles is through the vector and scalar potentials (the *minimal coupling prescription*).

Box 6.1 Diamagnetism and Paramagnetism

Materials characterized by small magnetic susceptibilities $|\chi| \ll 1$ are called *paramagnetic* if $\chi > 0$, and *diamagnetic* if $\chi < 0$. Magnetization in diamagnetic and paramagnetic media typically depends linearly on the applied magnetic field.

Physical Origin of Diamagnetism

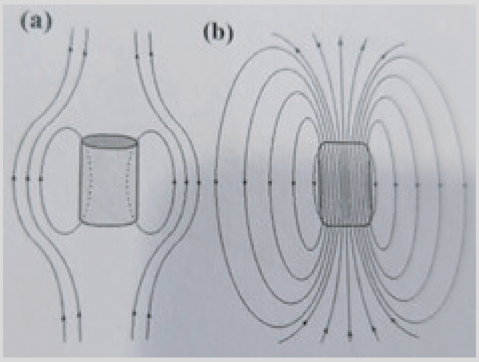
In diamagnetic media an external field $\boldsymbol{H}$ creates a magnetization $\boldsymbol{M}$ opposite to $\boldsymbol{H}$, so $\chi < 0$. Physically, external fields induce currents associated with orbital motion in the diamagnetic material. (The external field also interacts with electron spins, but this is a much smaller effect than the interaction with orbital currents.) The field created by the moving charges *opposes* the applied field $\boldsymbol{H}$ (*Lenz's law*). Thus, the magnetic field is decreased inside the material and magnetic lines of force are expelled from the medium. This diamagnetic effect is particularly dramatic in *superconductors*, where the magnetic field is expelled completely (*Meissner effect*), except for a thin surface layer where the magnetic field decays exponentially over a distance called the *London penetration depth*).[a]

Physical Origin of Paramagnetism

Some materials have magnetic dipoles associated with intrinsic spins (a quantum-mechanical effect). Application of an external field $\boldsymbol{H}$ then partially aligns the dipoles with the applied field, thereby *enhancing the internal field*. (This ordering is opposed by thermal fluctuations, so the fraction of alignment is temperature dependent.) Such materials are called *paramagnetic*. The net effect is that the lines of magnetic force are "drawn in" to paramagnetic material.

Magnetic Field Lines for Diamagnets and Paramagnets

The contrasting behavior of diamagnetic and paramagnetic matter in magnetic fields can be characterized by their magnetic field lines, as in the following figure [3].



The diamagnetic matter in (a) expels the magnetic field but the paramagnetic matter in (b) enhances the density of field lines within the sample.

[a] Diamagnetic effects are greatly amplified in a superconductor because the induced currents flow without resistance. It may be shown in quantum field theory that the (normally massless) photon gains an effective mass through interaction with the dielectric medium, which causes it to penetrate the superconductor with exponentially decaying probability. This acquisition of effective mass by photons is a non-relativistic model of the *Higgs mechanism*, whereby a massless gauge boson (the photon) acquires a mass. As we shall discuss further in Ch. 10, this *Higgs mode of spontaneous symmetry breaking* is fundamental for the Standard Model of elementary particle physics.

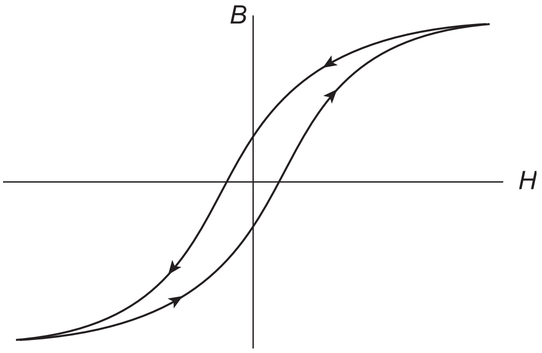


Fig. 6.1 Schematic illustration of hysteresis.

The case of ferromagnetic materials is more involved. In ferromagnetic substances the magnetic susceptibility χ (or the magnetic permeability μ) is positive but not constant; it depends on the applied field $\boldsymbol{H}$, and typically can take large values. When an external field is applied to a ferromagnetic material the system acquires a magnetization $\boldsymbol{M}$ aligned with the applied field. The origin of permanent ferromagnetism (magnetism that remains in the absence of an external field) is an essentially quantum-mechanical effect where many spins align spontaneously to form a highly collective state.[3] Ferromagnets may exhibit *hysteresis*, where the magnetic field $\boldsymbol{B}$ is not a single-valued function of $\boldsymbol{H}$, and the state of the system may depend on its preparation history; Fig. 6.1 illustrates. These complex behaviors lead to a constituitive relationship

$$\boldsymbol{B} = \boldsymbol{F}[\boldsymbol{H}], \tag{6.12}$$

where $\boldsymbol{F}[\boldsymbol{H}]$ is a *non-linear function of* $\boldsymbol{H}$. Clearly the complex relationship of $\boldsymbol{B}$ and $\boldsymbol{H}$ in ferromagnetic materials means that magnetic boundary-value problems are generally more difficult to deal with than corresponding problems in electrostatics.

6.1.3 Magnetic Boundary-Value Conditions

Boundary conditions for $\boldsymbol{B}$ and $\boldsymbol{H}$ at media interfaces were considered in Section 4.11; Eqs. (4.53b) and (4.54b) indicate that normal components of $\boldsymbol{B}$ and tangential components of $\boldsymbol{H}$ on opposite sides of a boundary between medium 1 and medium 2 are related by

$$(\boldsymbol{B}_2 - \boldsymbol{B}_1) \cdot \boldsymbol{n} = 0, \tag{6.13a}$$

$$\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{K}, \tag{6.13b}$$

[3] This is an example of quantum-mechanical *spontaneous symmetry breaking*, which is a phenomenon where the ground state wavefunction does not have the full symmetry of the Hamiltonian for a system. In the case of permanent ferromagnetism, the correct Hamiltonian is rotationally invariant (conservation of angular momentum) but the ferromagnetic wavefunction has spins aligned in a preferred direction, even if no external field is present, which breaks rotational invariance. Thermal fluctuations can destroy this collective alignment of spins, so in permanent ferromagnets the collective magnetic state disappears above a critical temperature called the *Curie temperature* T_C. For $T > T_\text{C}$, a ferromagnet then behaves as a paramagnet. An introduction to spontaneously broken symmetry may be found in Chs. 17 and 18 of Ref. [12].

where $\boldsymbol{n}$ is a unit normal vector pointing from region 1 into region 2, and $\boldsymbol{K}$ is the surface current density. If media satisfy the linear constituitive relation (6.10) and have finite conductivities so that $\boldsymbol{J} = \sigma \boldsymbol{E}$ and $\boldsymbol{K} = 0$, boundary conditions can be expressed as [15],

$$\boldsymbol{B}_2 \cdot \boldsymbol{n} = \boldsymbol{B}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{B}_2 \times \boldsymbol{n} = \frac{\mu_2}{\mu_1} \boldsymbol{B}_1 \times \boldsymbol{n}, \tag{6.14}$$

or as

$$\boldsymbol{H}_2 \cdot \boldsymbol{n} = \frac{\mu_1}{\mu_2} \boldsymbol{H}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{H}_2 \times \boldsymbol{n} = \boldsymbol{H}_1 \times \boldsymbol{n}. \tag{6.15}$$

If $\mu_1 \gg \mu_2$ the boundary conditions on $\boldsymbol{H}$ for highly-permeable material are essentially the same as for the electric field at the surface of a conductor, which permits electrostatic potential theory to be applied to magnetic field problems.

6.2 Solving Magnetostatic Boundary-Value Problems

Magnetostatic boundary value problems require solving Eqs. (6.8) subject to constituitive relations as in Eqs. (6.10) and (6.12). Especially because of the range of constituitive relations that are possible for magnetic and magnetized materials, a variety of situations can occur and a survey of possible methods of attack is useful. We summarize some approaches following the presentation in Jackson [15].

6.2.1 Methods Using the Vector Potential

Because $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$, it is always possible introduce a vector potential $\boldsymbol{A}(\boldsymbol{x})$ such that $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$. Then if we have a non-linear constituitive relation (6.12) the resulting differential equation

$$\boldsymbol{\nabla} \times \boldsymbol{H}[\boldsymbol{\nabla} \times \boldsymbol{A}] = \boldsymbol{J}$$

is generally very difficult to solve. However, if the constituitive relation is linear, $\boldsymbol{B} = \mu \boldsymbol{H}$, the preceding equation becomes

$$\boldsymbol{\nabla} \times \left(\frac{1}{\mu} \boldsymbol{\nabla} \times \boldsymbol{A} \right) = \boldsymbol{J}. \tag{6.16}$$

For a region of space in which μ is constant, the vector identity $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}$ of Eq. (A.8) may be used to write this as

$$\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} = \mu \boldsymbol{J}. \tag{6.17}$$

Invoking the Coulomb gauge condition (5.27) by setting $\boldsymbol{\nabla} \cdot \boldsymbol{A} = 0$ we obtain the Poisson equation

$$\nabla^2 \boldsymbol{A} = -\mu \boldsymbol{J}. \tag{6.18}$$

Comparing with Eq. (5.28) in vacuum, this is a Poisson equation with a current density modified by the medium, which is similar to a Poisson equation for uniform dielectric media with an effective charge density $(\varepsilon/\varepsilon_0)$. Matching of solutions for Eq. (6.18) across interfaces between different (linear) media can be implemented using the boundary conditions (6.14) or (6.15).

6.2.2 Using a Magnetic Scalar Potential if $\boldsymbol{J} = 0$

As mentioned in Section 5.5, for the special case $\boldsymbol{J} = 0$ where the current density vanishes in a region of interest, $\boldsymbol{\nabla} \times \boldsymbol{H} = 0$, suggesting the introduction of a *magnetic scalar potential* Φ_{M} such that

$$\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}}. \tag{6.19}$$

If the medium is linear and μ is constant the magnetic scalar potential satisfies the Laplace equation

$$\nabla^2\Phi_{\mathrm{M}} = 0, \tag{6.20}$$

for which the boundary conditions (6.14) are appropriate. This method is of limited utility since it applies only if $\boldsymbol{J} = 0$. One use case is the magnetic field external to a closed loop of current. Another is the hard ferromagnet that will be considered in Section 6.2.3.

6.2.3 Hard Ferromagnets

A *hard ferromagnet* is one having a magnetization that is largely independent of applied field for moderate field strengths. This suggests an approximation where the ferromagnet can be assumed to have a specified fixed magnetization $\boldsymbol{M}(\boldsymbol{x})$, using either magnetic scalar potential methods or vector potential methods.

Solve Using the Magnetic Scalar Potential

Since $\boldsymbol{J} = 0$ for a hard ferromagnetic, the magnetic scalar potential method of Section 6.2.2 is applicable. Then $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ becomes

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = \mu_0 \boldsymbol{\nabla} \cdot (\boldsymbol{H} + \boldsymbol{M}) = 0, \tag{6.21}$$

where Eq. (6.7) was used, and since $\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}}$ from Eq. (6.19), this becomes a *magnetic Poisson equation*,

$$\nabla^2\Phi_{\mathrm{M}} = \boldsymbol{\nabla} \cdot \boldsymbol{M} = -\rho_{\mathrm{M}}, \tag{6.22}$$

where an *effective magnetic-charge density*

$$\rho_{\mathrm{M}} \equiv -\boldsymbol{\nabla} \cdot \boldsymbol{M} \tag{6.23}$$

has been introduced. By analogy with the second term of Eq. (4.20) for electrostatics in electrically polarized matter, if there are no boundary surfaces the solution is expected to be [15]

$$\Phi_{\mathrm{M}}(\boldsymbol{x}) = -\frac{1}{4\pi}\int \frac{\boldsymbol{\nabla}' \cdot \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x = -\frac{1}{4\pi}\boldsymbol{\nabla} \cdot \int \frac{\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x, \tag{6.24}$$

where the second form results from an integration by parts and Eq. (5.9), which is justified if $\boldsymbol{M}$ is localized and well behaved.

Although physical magnetization distribution generally don't have discontinuities, it is sometimes useful to idealize a problem and treat $\boldsymbol{M}(\boldsymbol{x})$ as if it is discontinuous. We can then model a hard ferromagnet as having a volume V and a surface S, with $\boldsymbol{M}(\boldsymbol{x})$ finite inside but falling to zero at the surface S. Application of the divergence theorem to the surface indicates that in this idealization there is an effective magnetic surface-charge density given by [15]

$$\sigma_{\mathrm{M}} = \boldsymbol{n} \cdot \boldsymbol{M}, \tag{6.25}$$

where $\boldsymbol{n}$ is the outward normal at the surface. Then the first form of the potential (6.24) is modified to

$$\Phi_{\mathrm{M}}(\boldsymbol{x}) = -\frac{1}{4\pi} \int_V \frac{\boldsymbol{\nabla}' \cdot \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} d^3x + \frac{1}{4\pi} \oint_S \frac{\boldsymbol{n}' \cdot \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} da'. \tag{6.26}$$

An important special case is for uniform magnetization of the volume. Then $\boldsymbol{J}' = 0$ inside the volume and the first term of Eq. (6.26) vanishes, leaving only the surface term,

$$\Phi_{\mathrm{M}}(\boldsymbol{x}) = \frac{1}{4\pi} \oint_S \frac{\boldsymbol{n}' \cdot \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} da' = \frac{1}{4\pi} \oint_S \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} da', \tag{6.27}$$

where Eq. (6.25) was used. Thus, introduction of a sharp boundary for the volume of a uniformly magnetized object induces a surface charge on the boundary given by Eq. (6.25).

Solve Using the Vector Potential

If we choose to satisfy $\boldsymbol{\nabla} \cdot \boldsymbol{B}$ automatically by introducing a vector potential $\boldsymbol{A}$ and defining $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$, then the first of Eqs. (6.8) is

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{\nabla} \times \left(\frac{1}{\mu_0} \boldsymbol{B} - \boldsymbol{M} \right) = 0, \tag{6.28}$$

which becomes upon introduction of $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$,

$$\frac{1}{\mu_0} (\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla} \times \boldsymbol{M},$$

and upon using the identity $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}$ on the left side and Eq. (6.5) on the right side,

$$\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} = \mu_0 \boldsymbol{J}_{\mathrm{M}}, \tag{6.29}$$

where $\boldsymbol{J}_{\mathrm{M}}$ is the effective magnetic current density defined in Eq. (6.5). Transforming to Coulomb gauge (5.27) by setting $\boldsymbol{\nabla} \cdot \boldsymbol{A} = 0$ then leads to a Poisson equation for the vector potential,

$$\nabla^2 \boldsymbol{A} = -\mu_0 \boldsymbol{J}_{\mathrm{M}}. \tag{6.30}$$

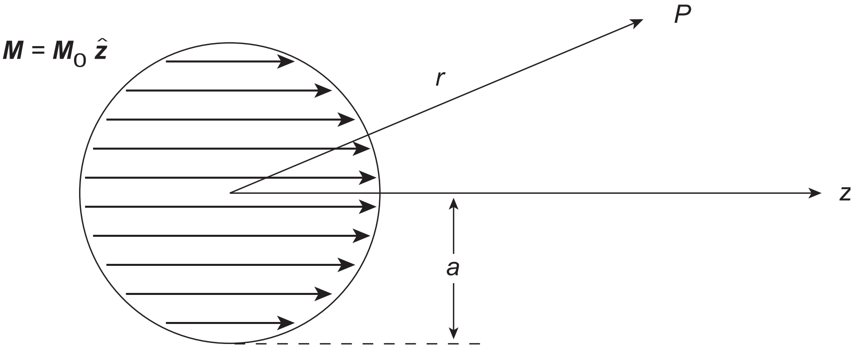


Fig. 6.2 Uniformly magnetized sphere of radius a, with a sharp surface and surface magnetic-charge density $\sigma_{\rm M}(\theta)$ with $\boldsymbol{M} = \boldsymbol{M}_0\hat{\boldsymbol{z}}$ parallel to the z axis, so that $\sigma_{\rm M} = \boldsymbol{n}\cdot\boldsymbol{M} = M_0\cos\theta$.

If there are no bounding surfaces, the solution is

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x \tag{6.31}$$

[compare Eq. (6.4) with $\boldsymbol{J} = 0$]. If the magnetization is discontinuous [as in the uniformly magnetized sphere with sharp surface in Section 6.2.4] it is necessary to add a surface integral to Eq. (6.24). For the case of $\boldsymbol{M}$ falling to zero at the surface S bounding the volume V, starting from Eq. (6.3), the proper generalization of Eq. (6.31) may be shown to be [15]

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x + \frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}')\times\boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da'. \tag{6.32}$$

For the special case that the magnetization is constant over the volume V, only the surface integral can contribute in Eq. (6.32).

6.2.4 Example: Uniformly Magnetized Sphere

As an example of solving boundary problems in magnetostatics using the methods just discussed, let's consider a sphere with uniform permanent magnetization, with a sharp transition at the surface from magnetized to non-magnetized matter; Fig. 6.2 illustrates.

Solution Using the Magnetic Scalar Potential

Since the sphere is uniformly magnetized, there are no volume currents and the scalar magnetic potential method of Section 6.2.2 is applicable. As discussed in Section 6.2.3, there will be a surface charge $\sigma_{\rm M} = \boldsymbol{n}\cdot\boldsymbol{M}$ associated with the sharp transition from magnetized to un-magnetized material. Assuming that $\boldsymbol{M} = M_0\hat{\boldsymbol{z}}$ so that $\sigma_{\rm M} = \boldsymbol{n}\cdot\boldsymbol{M} = \boldsymbol{M}_0\cos\theta$ using spherical coordinates, from Eq. (6.24) the scalar magnetic potential is

$$\Phi_{\mathrm{M}}(r,\theta) = \frac{1}{4\pi}\oint_S \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da' = \frac{M_0 a^2}{4\pi}\int \frac{\cos\theta}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega'.$$

Inserting the multipole expansion

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \sum_{l=0}^{\infty}\frac{r_<^l}{r_>^{l+1}}P_l(\cos\theta)$$

of Eq. (3.111) and using that $P_1(\cos\theta) = \cos\theta$ gives

$$\begin{aligned}\Phi_{\mathrm{M}}(r,\theta) &= \frac{M_0 a^2}{4\pi}\int \frac{P_1(\cos\theta)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega' \\ &= \frac{M_0 a^2}{4\pi}\sum_{l=0}^{\infty}\frac{r_<^l}{r_>^{l+1}}\int P_l(\cos\theta)P_1(\cos\theta)\,d\Omega' \\ &= \frac{1}{3}M_0 a^2\frac{r_<}{r_>^2}P_1(\cos\theta) = \frac{1}{3}M_0 a^2\frac{r_<}{r_>^2}\cos\theta, \end{aligned} \tag{6.33}$$

where the orthogonality condition of Eq. (3.99) has been used to eliminate all terms in the sum except for $l = 1$, and $(r_<, r_>)$ are the smaller and larger of r and a, respectively. Inside the sphere, $r_< = r$ and $r_> = a$, so the magnetic scalar potential is

$$\Phi_{\mathrm{M}}^{\mathrm{in}} = \frac{1}{3}M_0 r\cos\theta = \frac{1}{3}M_0 z.$$

Then from Eq. (6.19)

$$\boldsymbol{H}_{\mathrm{in}} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}}^{\mathrm{in}} = -\frac{\partial}{\partial z}\left(\frac{1}{3}M_0 z\right)\hat{\boldsymbol{z}} = -\frac{1}{3}M_0\hat{\boldsymbol{z}} = -\frac{1}{3}\boldsymbol{M},$$

and from Eq. (6.7),

$$\boldsymbol{B}_{\mathrm{in}} = \mu_0(\boldsymbol{H}_{\mathrm{in}} + \boldsymbol{M}) = \mu_0\left(-\frac{1}{3}\boldsymbol{M} + \boldsymbol{M}\right) = \frac{2\mu_0}{3}\boldsymbol{M}.$$

Therefore, the interior solution is

$$\Phi_{\mathrm{M}}^{\mathrm{in}} = \frac{1}{3}M_0 z \qquad \boldsymbol{H}_{\mathrm{in}} = -\frac{1}{3}\boldsymbol{M} \qquad \boldsymbol{B}_{\mathrm{in}} = \frac{2\mu_0}{3}\boldsymbol{M}, \tag{6.34}$$

where we see that the fields are constant inside the sphere, with $\boldsymbol{B}$ parallel and $\boldsymbol{H}$ antiparallel to $\boldsymbol{M}$.

For the exterior solution, $r_< = a$ and $r_> = r$, which gives from Eq. (6.33) for the exterior potential,

$$\Phi_{\mathrm{M}}^{\mathrm{out}} = \frac{1}{3}M_0 a^3\frac{\cos\theta}{r^2}, \tag{6.35}$$

which is a dipole potential with dipole moment

$$\boldsymbol{m} = \frac{4\pi a^3}{3}\boldsymbol{M}. \tag{6.36}$$

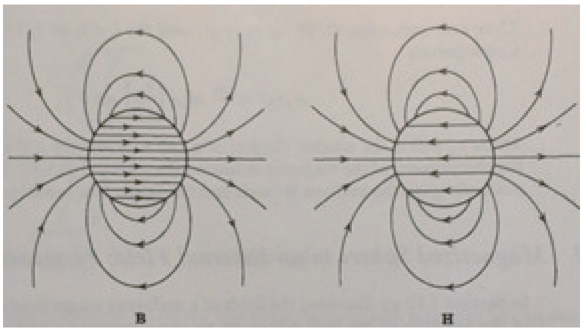


Fig. 6.3 Lines of ***B*** and ***H*** for a uniformly magnetized sphere having a sharp boundary [15].

Then, proceeding as above

$$\boldsymbol{H}_{\text{out}} = -\boldsymbol{\nabla}\Phi_{\text{M}}^{\text{out}} = -\frac{1}{3}\frac{a^3}{r^3}\boldsymbol{M} \qquad \boldsymbol{B}_{\text{out}} = \mu_0(\boldsymbol{H}+\boldsymbol{M}) = \left(\frac{3r^3-a^3}{3r^3}\right)\mu_0\boldsymbol{M}. \tag{6.37}$$

The ***B*** and ***H*** fields are plotted in Fig. 6.3

Solution Using the Vector Potential

Alternatively, the vector potential and Eq. (6.32) can be used to obtain the solution for the problem posed in Fig. 6.2. The magnetization is assumed uniform so the volume current $\boldsymbol{J}_{\text{M}} = 0$ and the first (volume) term in Eq. (6.32) make no contribution. However, there is a surface charge due to the sharp boundary on magnetization so the second term in Eq. (6.32) will be non-zero and

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}')\times\boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da'. \tag{6.38}$$

Using the notation $(\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_3) = (\hat{\boldsymbol{x}},\hat{\boldsymbol{y}},\hat{\boldsymbol{z}})$ for cartesian basis vectors and $(\boldsymbol{\varepsilon}_r,\boldsymbol{\varepsilon}_\theta,\boldsymbol{\varepsilon}_\phi) = (\hat{\boldsymbol{r}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}})$ for spherical basis vectors, since $\boldsymbol{M} = M_0\boldsymbol{\varepsilon}_3$ we have

$$\begin{aligned}\boldsymbol{M}(\boldsymbol{x}')\times\boldsymbol{n}' &= M_0\sin\theta'\boldsymbol{\varepsilon}_\phi \\ &= M_0\sin\theta'(-\sin\phi\boldsymbol{\varepsilon}_1+\cos\phi'\boldsymbol{\varepsilon}_2),\end{aligned}$$

where Eq. (A.41) was used. The problem has azimuthal (ϕ) symmetry about the z-axis. If the observing point P is chosen to lie in the x-y plane, then only the y component of $\boldsymbol{M}\times\boldsymbol{n}$ survives integration over the azimuth so $(-\sin\phi'\boldsymbol{\varepsilon}_1+\cos\phi'\boldsymbol{\varepsilon}_2) \to \cos\phi'\boldsymbol{\varepsilon}_2$, which gives an azimuthal component of the vector potential

$$A_\phi = \frac{\mu_0}{4\pi}M_0a^2\int\frac{\sin\theta'\cos\phi'}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega', \tag{6.39}$$

where the components of $\boldsymbol{x}'$ are $(r'=a,\theta',\phi')$, confining the integration to the surface of the sphere. Since

$$Y_{11}(\theta',\phi') = -\sqrt{\frac{3}{8\pi}}\sin\theta' e^{i\phi'} = -\sqrt{\frac{3}{8\pi}}\sin\theta'(\cos\phi'+i\sin\phi'),$$

we may write

$$\sin\theta'\cos\phi' = -\sqrt{\frac{8\pi}{3}}\,\mathrm{Re}\,[\mathrm{Y}_{11}(\theta',\phi')]$$

where Re (x) denotes the real part of x. Thus,

$$A_\phi = -\sqrt{\frac{3}{8\pi}}\frac{\mu_0}{4\pi}M_0a^2\int\frac{\mathrm{Re}\,[\mathrm{Y}_{11}(\theta',\phi')]}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega'.$$

Then if the denominator is expanded using Eq. (3.114),

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = 4\pi\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{1}{2l+1}\frac{r_<^l}{r_>^{l+1}}Y_{lm}^*(\theta',\phi')Y_{lm}(\theta,\phi),$$

the spherical harmonic orthogonality relation (3.118),

$$\int_0^{2\pi}d\phi\int_0^{\pi}Y_{l'm'}^*(\theta,\phi)Y_{lm}(\theta,\phi)\sin\theta d\theta = \delta_{l'l}\delta_{m'm},$$

ensures that only the $l=1, m=1$ term survives the summation and the vector potential is

$$A_\phi(\boldsymbol{x}) = \frac{\mu_0}{3}M_0a^2\left(\frac{r_<}{r_>^2}\right)\sin\theta \tag{6.40}$$

For the inside solution, $r_< = r$ and $r_> = a$, so

$$A_\phi^{\text{in}}(\boldsymbol{x}) = \frac{\mu_0}{3}M_0r\sin\theta \tag{6.41}$$

You are asked to calculate the corresponding $\boldsymbol{B}$ and $\boldsymbol{H}$ fields for the interior of the sphere in Problem 6.1.

By placing a shell of permeable matter in a magnetic field, it is possible to shield the interior of the shell from the magnetic field. Example 6.1 illustrates.

Example 6.1 Consider Fig. 6.4, where the shell contains material of permeability μ. Let us find the fields $\boldsymbol{B}$ and $\boldsymbol{H}$ for this arrangement. There are no currents so the magnetic scalar potential method of Section 6.2.2 is applicable and from Eq. (6.19)

$$\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}},$$

and since $\boldsymbol{B} = \mu\boldsymbol{H}$, $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ becomes $\boldsymbol{\nabla}\cdot\boldsymbol{H} = 0$ in all regions. Therefore, the magnetic potential Φ_{M} satisfies the Laplace equation in all regions and the problem reduces to solving the Laplace equation subject to the boundary conditions of Eq. (6.15) at $r=a$ and $r=b$. If $r>b$, the potential is of the form

$$\Phi_{\mathrm{M}} = -H_0r\cos\theta + \sum_{l=0}^{\infty}\frac{\alpha_l}{r^{l+1}}P_l(\cos\theta) \qquad (r>b), \tag{6.42}$$

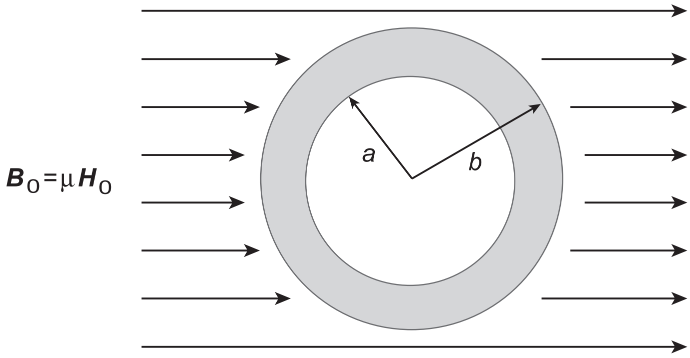


Fig. 6.4 A shell of material with permeability μ is placed in a previously uniform magnetic field $\boldsymbol{B}_0 = \mu_0 \boldsymbol{H}_0$. Example 6.1 demonstrates that if the shell contains highly permeable material the cavity inside the shell is shielded strongly from the magnetic field.

which gives a uniform field $\Phi_{\mathrm{M}} = \boldsymbol{H}_0$ at large distance. Likewise, in the interior regions the potential must take the form

$$\Phi_{\mathrm{M}} = \sum_{l=0}^{\infty} \left(\beta_l r^l + \gamma_l \frac{1}{r^{l+1}} \right) P_l(\cos\theta). \qquad (a < r < b), \qquad (6.43a)$$

$$\Phi_{\mathrm{M}} = \sum_{l=0}^{\infty} \delta_l r^l P_l(\cos\theta) \qquad (r < a). \qquad (6.43b)$$

The boundary conditions at $r = a$ and $r = b$ require the components H_θ and B_r be continuous, which implies that

$$\frac{\partial \Phi_{\mathrm{M}}}{\partial \theta}(b_+) = \frac{\partial \Phi_{\mathrm{M}}}{\partial \theta}(b_-) \qquad \frac{\partial \Phi_{\mathrm{M}}}{\partial \theta}(a_+) = \frac{\partial \Phi_{\mathrm{M}}}{\partial \theta}(a_-), \qquad (6.44a)$$

$$\mu_0 \frac{\partial \Phi_{\mathrm{M}}}{\partial r}(b_+) = \mu \frac{\partial \Phi_{\mathrm{M}}}{\partial r}(b_-) \qquad \mu \frac{\partial \Phi_{\mathrm{M}}}{\partial r}(a_+) = \mu_0 \frac{\partial \Phi_{\mathrm{M}}}{\partial r}(a_-), \qquad (6.44b)$$

where the notation b_+ means the limit $r \to b$ approached from $r > b$ and the notation b_- means means the limit $r \to b$ approached from $r < b$, with a similar convention for $a_\pm$. The four conditions (6.44) determine the unknown constants in Eqs. (6.42) and (6.43). One finds that all coefficients with $l \neq 1$ vanish and the $l = 1$ coefficients satisfy the simultaneous equations

$$\begin{gathered} \alpha_1 - b^3\beta_1 - \gamma_1 = b^3 H_0, \\ 2\alpha_1 + \mu' b^3 \beta_1 - 2\mu'\gamma_1 = -b^3 H_0, \\ a^3\beta_1 + \gamma_1 - a^3\delta_1 = 0, \\ \mu' a^3 \beta_1 - 2\mu'\gamma_1 - a^3\delta_1 = 0, \end{gathered} \qquad (6.45)$$

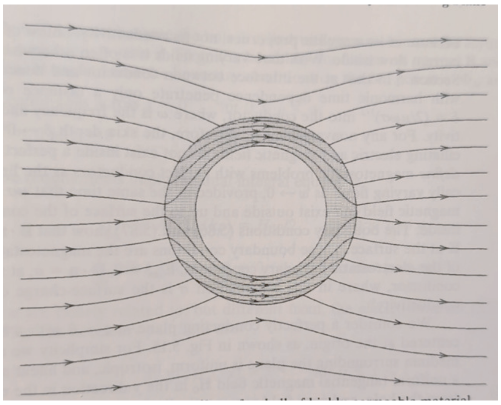


Fig. 6.5 The magnetic shielding effect of a shell of highly permeable material. Lines for the magnetic field $\boldsymbol{B}$ are shown [15]. Note the absence of field lines in the cavity inside the shell. Calculation described in Example 6.1.

where $\mu' \equiv \mu/\mu_0$. The solution of Eqs. (6.45) for α_1 and δ_1 are [15],

$$\begin{aligned} \alpha_1 &= \left(\frac{(2\mu'+1)(\mu'-1)}{(2\mu'+1)(\mu'+2) - 2\dfrac{a^3}{b^3}(\mu'-1)^2} \right)(b^3 - a^3)H_0, \\ \delta_1 &= -\left(\frac{9\mu'}{(2\mu'+1)(\mu'+2) - 2\dfrac{a^3}{b^3}(\mu'-1)^2} \right) H_0. \end{aligned} \tag{6.46}$$

The corresponding magnetic field lines are shown in Fig. 6.5. For this solution the potential outside the shell corresponds to the original uniform field $\boldsymbol{H}_0$ plus a dipole field with dipole moment α_1 parallel to $\boldsymbol{H}_0$. In the cavity inside the shell there is a uniform field parallel to $\boldsymbol{H}_0$ and equal in magnitude to $-\delta_1$. If $\mu \gg \mu_0$, the dipole moment α_1 and the inner field $-\delta_1$ tend to

$$\alpha_1 \to b^3 H_0 \qquad -\delta_1 \to \frac{9\mu_0}{2\mu(1 - a^3/b^3)} H_0. \tag{6.47}$$

Thus the field in the inner cavity is proportional to μ^{-1} and a shield made of highly permeable material causes a large reduction of the field inside. Even thin shells of material with $\mu/\mu_0 \sim 10^3 - 10^6$ can greatly reduce the interior field, as is clear from Fig. 6.5.

6.3 Faraday and the Law of Induction

We have considered primarily static charges (no relative motion) as sources of electric fields and steady-state (no change in time) charge currents as the source of magnetic fields

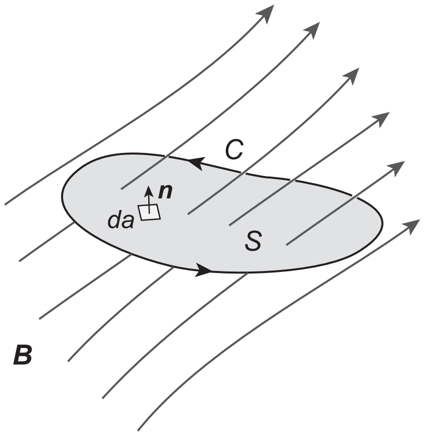


Fig. 6.6 Magnetic flux through a circuit C bounding a surface S with a unit normal to the surface $\boldsymbol{n}$ and surface element da. The magnetic field in the neighborhood of the circuit is $\boldsymbol{B}$.

to this point. However, many important electromagnetic phenomena involve the motion of charges and/or non-steady electrical currents. Michael Faraday performed the first quantitative experiments on time-dependent electric and magnetic field, and their relationship, beginning in 1831. Faraday's essential observations were that

1. a transient current is produced in a test circuit if a steady current in a nearby circuit is turned off;
2. a transient current is also produced in a test circuit if a nearby circuit with a steady current is moved relative to the first circuit;
3. a transient current is produced in a test circuit if a permanent magnet is moved into or out of the circuit.

In summary, a current flows in the test circuit if a the current in a nearby circuit changes in time, or if the two circuits move with respect to each other.

Faraday interpreted these results as originating in a *changing magnetic flux* that produced an electric field around the test circuit, which leads to an *electromotive force* $\mathscr{E}$ (EMF) that drives a current in the test circuit governed by Ohm's law (see Box 6.2). Thus, with Faraday's results and interpretation begins the unification of electricity and magnetism in a single theory of *electromagnetism*.

Let us express the results of Faraday's observations mathematically. Consider Fig. 6.6; the magnetic flux linking the circuit is

$$\mathscr{F} = \int_S \boldsymbol{B} \cdot \boldsymbol{n} da \tag{6.48}$$

and the electromotive force around the circuit is

$$\mathscr{E} = \oint_C \boldsymbol{E}' \cdot d\boldsymbol{l}, \tag{6.49}$$

Box 6.2 Ohm's Law

To make a current flow some force, typically electromagnetic in origin, must push on the charges (continuously if, as is usual, there is any resistance to the charge flow). For most materials the current density $\boldsymbol{J}$ is proportional to the force per unit charge $\boldsymbol{f}$,

$$\boldsymbol{J} = \sigma \boldsymbol{f},$$

where the constant of proportionality σ depends on the nature of the medium and is called the *conductivity*. Often the reciprocal of the conductivity, $\rho = \sigma^{-1}$, which is called the *resistivity*, is used instead. If the force is electromagnetic in origin,

$$\boldsymbol{J} = \sigma(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}),$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, and $\boldsymbol{v}$ is the velocity. Except for special circumstances such as in plasmas, the velocity is relatively small and the $\boldsymbol{v} \times \boldsymbol{B}$ term can be neglected, leaving

$$\boldsymbol{J} = \sigma \boldsymbol{E}.$$

This is called *Ohm's law*.[a] For a cylindrical wire of cross-sectional area A and conductivity σ, if the potential difference between the ends is V the electric field and current density are uniform and the total current is

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

In this and similar examples the total current flow in the wire from one point to another is proportional to the potential difference between the points,

$$V = IR,$$

where the constant of proportionality R is called the *resistance*. This is the most familiar form of Ohm's law. The resistance depends on the geometry and the conductivity of the medium; in the example given above $R = L/\sigma A$, where L is the length.

[a] Ohm's law is not a true "law" in the same sense as say the law of gravity. It is a "rule of thumb" that is often (but not always) obeyed. Conductors that do obey Ohm's law are called *ohmic conductors*.

where $\boldsymbol{E}'$ is the electric field at element $d\boldsymbol{l}$ of the circuit C. Faraday's observations are summarized by the relationship

$$\mathscr{E} = -k\frac{d\mathscr{F}}{dt}.$$

between the rate of change of the magnetic flux and the EMF, where k is a constant of proportionality. The constant k can be determined by requiring Galilean invariance at low velocities[4] This indicates that $k = 1$ in SI units (and $k = c^{-1}$ in Gaussian units). Therefore, in SI units

$$\mathscr{E} = -\frac{d\mathscr{F}}{dt}, \tag{6.50}$$

where the sign is specified by *Lenz's law*.

> ***Lenz's law:*** The induced current and corresponding magnetic field are in a direction so as to oppose changing the flux through the circuit.

Combining Eq. (6.50) with Eqs. (6.48) and (6.49),

$$\oint_C \boldsymbol{E}' \cdot d\boldsymbol{l} = -\frac{d}{dt} \int_S \boldsymbol{B} \cdot \boldsymbol{n}\, da \qquad \text{(Faraday's law)} \tag{6.51}$$

indicating that the induced EMF is proportional to the *total time derivative* of the flux.[5] Equation (6.51) represents a form of Faraday's law with broad implications. If we view C as a geometrical closed path not necessarily coincident with an electrical circuit, Eq. (6.51) may be viewed as a relationship among the fields $\boldsymbol{B}$ and $\boldsymbol{E}'$.

If the circuit C is moving with some velocity, the total time derivative must take that motion into account, since the flux through the circuit can change for two reasons:

1. the flux is changing with time at a given point, or
2. translation of the circuit changes location of the circuit in an external field.

This can be taken into account by computing the *convective derivative*,

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla}, \tag{6.52}$$

where $\boldsymbol{v}$ is the velocity and the partial derivative in the first term on the right side has the usual meaning that it is the derivative with respect to a variable (time in this case) with all other variables held constant. Then the total time derivative of the moving circuit is [15],

$$\frac{d}{dt} \int_S \boldsymbol{B} \cdot \boldsymbol{n}\, da = \int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}\, da + \oint_C (\boldsymbol{B} \times \boldsymbol{v}) \cdot d\boldsymbol{l}. \tag{6.53}$$

Then this result can be used to write Eq. (6.51) in the form

$$\oint_C \left[\boldsymbol{E}' - (\boldsymbol{v} \times \boldsymbol{B})\right] \cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}\, da. \tag{6.54}$$

This is Faraday's law applied to the moving circuit C, but if we think of the circuit C and

[4] As we shall see in Ch. 8, the Maxwell equations are invariant under Lorentz transformations, not Galilean transformations. But Lorentz transformations reduce to Galilean transformations in the limit $v \ll c$, so requiring Galilean invariance for low velocities is a legitimate way to determine the constant k.

[5] The flux can be changed by changing the magnetic field, or the shape, position, or orientation of the circuit. The total time derivative accounts for all such possibilities. Note that $\boldsymbol{E}'$ is the electric field at $d\boldsymbol{l}$ *in the coordinate system where* $d\boldsymbol{l}$ *is at rest*.

surface S being instantaneously at a certain point, then application of Eq. (6.51) to that circuit at fixed location gives,

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n} \, da, \tag{6.55}$$

where $\boldsymbol{E}$ is now the electric field in the laboratory frame. Galilean invariance (valid at low velocities) then requires that the left sides of Eqs. (6.54) and (6.55) must be equivalent.

If the circuit is held fixed in a reference frame so that the electric and magnetic fields are defined in the same frames, Faraday's law (6.51) in integral form can be transformed into a differential equation. From Stoke's theorem,

$$\oint_C \boldsymbol{E}' \cdot d\boldsymbol{l} = \int_S (\boldsymbol{\nabla} \times \boldsymbol{E}') \cdot \boldsymbol{n} \, da$$

and Eq. (6.51) may be written as

$$\int_S \left(\boldsymbol{\nabla} \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} \right) \cdot \boldsymbol{n} \, da = 0.$$

But the circuit C and surface S are arbitrary, so the integrand must vanish identically, giving

$$\boldsymbol{\nabla} \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0, \tag{6.56}$$

which is Faraday's law (1.1b) in differential form. We note that this is the time-dependent generalization of the equation $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$ from electrostatics.

Background and Further Reading

The presentation in this chapter has followed Jackson [15] rather closely.

Problems

6.1 Find the fields $\boldsymbol{B}_{\text{in}}$ and $\boldsymbol{H}_{\text{in}}$ corresponding to the vector potential of Eq. (6.41),

$$A_{\phi}^{\text{in}}(\boldsymbol{x}) = \frac{\mu_0}{3} M_0 r \sin\theta,$$

inside the sphere of Fig. 6.2.

7 Maxwell's Equations

The preceding chapters have provided a systematic understanding and validation of the various pieces of electromagnetic theory that James Clerk Maxwell synthesized into the four concise *Maxwell equations*. Up to this point we have treated electricity and magnetism largely as separate subjects. As Faraday's discovery of induction that was discussed in Section 6.3 makes clear, this distinction begins to fail when we move from static to time-dependent phenomena, and we will now begin to address a unified picture of electromagnetism. In this chapter we take the Maxwell equations to be the basis for all classical understanding of electromagnetism and address systematically their broader scientific and technical implications.

7.1 The Almost-but-Not-Quite Maxwell's Equations

The basic equations of electricity and magnetism that we have studied in Chs. 2-6 can be summarized (in medium, in differential form) as

$$\nabla \cdot \boldsymbol{D} = \rho \qquad \text{(Gauss's law)}, \qquad (7.1a)$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \qquad (7.1b)$$

$$\nabla \cdot \boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \qquad (7.1c)$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} \qquad \text{(Ampère's law)}, \qquad (7.1d)$$

In modern notation, these were the fundamental equations of electromagnetism as they stood in the mid-1800s when James Clerk Maxwell set about his synthesis of electromagnetic understanding. A comparison shows that these are almost, but not quite, the Maxwell equations (in medium, in differential form) that were introduced in Eqs. (4.47). The difference lies in the fourth equation (Ampère's law), where a term $-\partial \boldsymbol{D}/\partial t$ is missing relative to the fourth Maxwell equation (4.47d).

7.1.1 Ampère's Law and the Displacement Current

It is important to recognize that all but Faraday's law in Eqs. (7.1) were derived from data taken under steady-state conditions, so we should expect that modifications might be required in the face of data for time-dependent fields. Maxwell first realized that the fundamental equations of electromagnetism (7.1) in his time were inconsistent as they then stood. Since the divergence of a curl vanishes by a basic vector-calculus identity,

$\nabla\cdot(\nabla\times\boldsymbol{B})=0$, it follows from Ampère's law (7.1d) for steady currents that

$$\nabla\cdot\boldsymbol{J}=\nabla\cdot(\nabla\times\boldsymbol{H})=0. \tag{7.2}$$

But from the continuity equation (1.3) ensuring conservation of charge,

$$\frac{\partial\rho}{\partial t}=-\nabla\cdot\boldsymbol{J},$$

so $\nabla\cdot\boldsymbol{J}=0$ can hold only if the charge density doesn't change with time. Maxwell modified Ampère's law to accomodate a time dependence by introducing a *displacement current* term $\partial\boldsymbol{D}/\partial t$,[1]

$$\nabla\times\boldsymbol{H}=\boldsymbol{J}\quad\longrightarrow\nabla\times\boldsymbol{H}=\boldsymbol{J}+\frac{\partial\boldsymbol{D}}{\partial t}, \tag{7.3}$$

thus converting Eqs. (7.1) into what we now call the Maxwell equations (4.47) (in medium, in differential form),[2]

$$\begin{aligned}
\nabla\cdot\boldsymbol{D}&=\rho && \text{(Gauss's law)}, && (7.4a)\\
\nabla\times\boldsymbol{E}+\frac{\partial\boldsymbol{B}}{\partial t}&=0 && \text{(Faraday's law)}, && (7.4b)\\
\nabla\cdot\boldsymbol{B}&=0 && \text{(No magnetic charges)}, && (7.4c)\\
\nabla\times\boldsymbol{H}-\frac{\partial\boldsymbol{D}}{\partial t}&=\boldsymbol{J} && \text{(Ampère–Maxwell law)}, && (7.4d)
\end{aligned}$$

where the last equation is now called the *Ampère–Maxwell law*. It still is the same law as before when applied to steady-state phenomena, but addition of the displacement current term means that a changing electric field can generate a magnetic field, even if there is no current. Thus, it is the converse of Faraday's law, where a changing magnetic field can produce an electric field. These (Maxwell) equations are now consistent with the continuity equation. Taking the divergence of both sides of the Ampère–Maxwell law,

$$\nabla\cdot(\nabla\times\boldsymbol{H})=\nabla\cdot\boldsymbol{j}+\frac{\partial(\nabla\cdot\boldsymbol{D})}{\partial t},$$

using $\nabla\cdot\boldsymbol{D}=\rho$ from Gauss's law and the identity $\nabla\cdot(\nabla\times\boldsymbol{B})=0$ from Eq. (A.6), this becomes

$$\frac{\partial\rho}{\partial t}+\nabla\cdot\boldsymbol{j}=0,$$

which is the continuity equation (1.3).

[1] Maxwell's original detailed physical reason for the term and corresponding name is partially bogus, but his insight that the term is required was seminal, and the name itself survives.

[2] Recall that the vacuum Maxwell equations (1.1) in differential form can be recovered by substituting $\boldsymbol{D}=\varepsilon_0\boldsymbol{E}$ and $\boldsymbol{H}=\boldsymbol{B}/\mu_0$ into these equations, remembering from Eq. (2.4) that $\mu_0\varepsilon_0=1/c^2$, and that the integral form of the Maxwell equations in medium are given by Eqs. (4.52).

Local conservation of charge: Physically the continuity equation requires that changes in electrical charge in some arbitrary volume are caused by flow of charge through the surface bounding that volume. This requires conservation of charge within a volume of space that can be arbitrarily small, thus implying that charge is *conserved locally*. A charge that disappears from one point in space and instantly reappears at another point is consistent with *global charge conservation*, but not with *local charge conservation*. The reason requires relativistic quantum field theory for its full explanation: destroying a charge at one point and simultaneously creating it at another would require instantaneous propagation of a signal between the two points, which is inconsistent with special relativity.

Maxwell's equations (7.4), supplemented by the Lorentz force law (1.4)

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}),$$

to describe electromagnetic forces, and Newton's laws of motion to translate force into particle motion are thought to describe all of classical electromagnetism.

7.1.2 Implications of the Ampère–Maxwell Law

As was introduced in Box 1.1, Maxwell's seemingly small change in the Ampère law turns out to have enormous implications for both our classical and quantum understanding of electromagnetism.

1. Maxwell's addition of $\partial \boldsymbol{D}/\partial t$ to Ampère's law (and Faraday's induction experiments) effectively brought together the previously separate subjects of electricity and magnetism. Ampère's law is only about magnetism, but both the polarized electric field $\boldsymbol{D}$ and the polarized magnetic field $\boldsymbol{H}$ appear in the Ampère–Maxwell law, while $\boldsymbol{E}$ and $\boldsymbol{B}$ both appear in Faraday's law. Changing electric fields produce magnetic fields and changing magnetic fields produce electric fields, and we may now speak of the unified subject of *electromagnetism*.
2. The fundamental equations of electromagnetism are now consistent with (local) conservation of electrical charge.
3. This modification will lead eventually to the greatest triumph of the classical Maxwell equations: the realization that the Maxwell equations have a wave solution, and that the resulting electromagnetic waves may be interpreted as light, thus unifying electricity and magnetism with optics.
4. That Maxwell's equations obey the continuity equation and thus conserve charge locally will lead to the idea of classical *electromagnetic gauge invariance*, which will underlie a quantum field theory of electromagnetism (*quantum electrodynamics* or QED).
5. Electromagnetic gauge invariance will eventually be generalized to more sophisticated local gauge invariance in the weak and strong interactions, resulting in the relativistic quantum field theory that we term the *Standard Model* of elementary particle physics.

Some of these topics are beyond the scope of classical electromagnetism, but have their historical and scientific antecedents in that discipline. A general introduction to modern relativistic quantum field theory and many of these topics specifically may be found in Ref. [9].

7.2 Vector and Scalar Potentials

In our discussions of electrostatics and magnetostatics we introduced the scalar potential Φ (Section 2.5) and the vector potential $\boldsymbol{A}$ (Section 5.5). The Maxwell equations (7.4) are a set of coupled first-order differential equations relating the components of the electric and magnetic fields. To solve those coupled differential equations it is often convenient to introduce the potentials $\boldsymbol{A}$ and Φ, which satisfy some of the Maxwell equations identically and leave a smaller number of second-order differential equations to solve.

Consider the two homogeneous equations (the ones equal to zero on the right side) in Eqs. 7.4. As we have noted in Section 5.5, since $\boldsymbol{\nabla}\cdot\boldsymbol{B}=0$, the magnetic field $\boldsymbol{B}$ can be described as the curl of a vector potential $\boldsymbol{A}$,[3]

$$\boldsymbol{B}=\boldsymbol{\nabla}\times\boldsymbol{A}. \tag{7.5}$$

Then Faraday's law (7.4b) may be written as,

$$\boldsymbol{\nabla}\times\left(\boldsymbol{E}+\frac{\partial\boldsymbol{A}}{\partial t}\right)=0. \tag{7.6}$$

Since the curl of the quantity $\boldsymbol{E}+\partial\boldsymbol{A}/\partial t$ in parentheses vanishes, it can be written as the gradient of a scalar function; let's choose it to be minus the scalar potential Φ,

$$\boldsymbol{E}+\frac{\partial\boldsymbol{A}}{\partial t}=-\boldsymbol{\nabla}\Phi,$$

which rearranges to

$$\boldsymbol{E}=-\boldsymbol{\nabla}\Phi-\frac{\partial\boldsymbol{A}}{\partial t}. \tag{7.7}$$

For simplicity, let's restrict consideration to the vacuum form of the Maxwell equations given in Eq. (1.1), which are formulated in terms of the fundamental fields $\boldsymbol{E}$ and $\boldsymbol{B}$,

$$\boldsymbol{\nabla}\cdot\boldsymbol{E}=\frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)}, \tag{7.8a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E}+\frac{\partial\boldsymbol{B}}{\partial t}=0 \qquad \text{(Faraday's law)}, \tag{7.8b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B}=0 \qquad \text{(No magnetic charges)}, \tag{7.8c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B}-\frac{1}{c^2}\frac{\partial\boldsymbol{E}}{\partial t}=\mu_0\boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}. \tag{7.8d}$$

[3] Recall the reason: if $\boldsymbol{B}=\boldsymbol{\nabla}\times\boldsymbol{A}$, then by taking the divergence of both sides $\boldsymbol{\nabla}\cdot\boldsymbol{B}=\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A})=0$, where the identity (A.6) was used. Thus the condition $\boldsymbol{\nabla}\cdot\boldsymbol{B}=0$ is guaranteed if $\boldsymbol{B}$ derives from the curl of a vector potential.

Introducing the substitutions $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$ and $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi - \partial \boldsymbol{A}/\partial t$ into the Maxwell equations (7.8), we find using the identities $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = 0$ and $\boldsymbol{\nabla} \times \boldsymbol{\nabla}\Phi = 0$, that the homogeneous equations (7.8b) and (7.8c) are *satisfied identically*, while the inhomogeneous equations (7.8a) and (7.8d) are transformed into the coupled second-order differential equations,

$$\nabla^2\Phi + \frac{\partial}{\partial t}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) = -\frac{\rho}{\varepsilon_0}, \tag{7.9a}$$

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t}\right) = -\mu_0 \boldsymbol{J}, \tag{7.9b}$$

where the identity $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}$ has been used. Thus the problem has been converted into solving the two coupled, second-order differential equations (7.9). These equations can be decoupled by a suitable gauge transformation, as will now be shown.

7.3 Exploiting Gauge Symmetry

In Section 5.5 we showed that the magnetic field is invariant under a gauge transformation on the vector potential,

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\chi,$$

where χ is an arbitrary scalar function. But in general the Maxwell equations involve both magnetic and electric fields. If the electric field is to be unaltered under this transformation of the magnetic field the scalar potential must be changed at the same time according to

$$\Phi \to \Phi' = \Phi - \frac{\partial \chi}{\partial t}.$$

Gauge Transformation: A *classical gauge transformation* on the electromagnetic field is defined by the simultaneous transformations

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \to \Phi - \frac{\partial \chi}{\partial t}, \tag{7.10}$$

on the vector potential $\boldsymbol{A}$ and the scalar potential Φ, for an arbitrary scalar function χ. A gauge transformation leaves the electric and magnetic fields, and thus Maxwell's equations, unchanged.

Notice that in the gauge transformation the same scalar function χ is used in both equations, but the choice of χ is arbitrary.

Mathematically, the invariance of electromagnetism under the gauge transformation (7.10) means that an electromagnetic field may be viewed as an *equivalence class* of potentials $\{\Phi, \boldsymbol{A}\}$, where equivalence classes are defined in Box 7.1. Thus, on the set of all possible electromagnetic potentials $\{(\Phi_1, \boldsymbol{A}_1), (\Phi_2, \boldsymbol{A}_2), \cdots\}$ we may define an equivalence relation "related by a gauge transformation (7.10)" and the equivalence class is the set of

Box 7.1 Equivalence Classes

A very useful concept for a set is that of an *equivalence class*, which is predicated on there being an *equivalence relation* defined between set members. Equivalence classes are of utility when one wishes to refer to a subset of things that are "the same" for our considerations, in that they all exhibit a specific property.

> For example, if for the set of all balls of a solid color we define an equivalence relation "having the same color", then the subset of all yellow balls would be one equivalence class of the set of all balls of a solid color, and the subset of all black balls would be another.

When a set S has an equivalence relation defined on it the set may be partitioned naturally into disjoint (non-overlapping) equivalence classes, with elements a and b belonging to the same equivalence class if and only if they are equivalent. Formally, equivalence is a binary relation $\sim$ between members of a set S exhibiting

1. *Reflexivity:* For each $a \in \mathrm{S}$, we have $a \sim a$.
2. *Symmetry:* For each $a, b \in \mathrm{S}$, if $a \sim b$, then $b \sim a$.
3. *Transitivity:* For each $a, b, c \in \mathrm{S}$, if $a \sim b$ and $b \sim c$, then $a \sim c$.

Since mathematical groups are sets with added structure, equivalence classes may be defined in a similar way for groups.

all electromagnetic potentials related to each other by a gauge transformation. The objects of the class then are "equivalent" in the sense that all members of the equivalence class give the same electric and magnetic fields, and thus lead to the same electromagnetic physics when inserted in the Maxwell equations.

7.3.1 Lorenz Gauge

Suppose that we now take advantage of the invariance of electromagnetism under gauge transformations and choose a set of potentials $\{\Phi, \boldsymbol{A}\}$ that satisfy the *Lorenz condition*

$$\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t} = 0 \qquad \text{(Lorenz gauge).} \tag{7.11}$$

A constraint like Eq. (7.11) is termed a *gauge-fixing condition* and imposing such a constraint is termed *fixing the gauge*. The gauge choice implied by Eq. (7.11) is called the *Lorenz gauge*.[4] If the Lorenz gauge condition is inserted into the coupled equations (7.9),

[4] The *Lorenz gauge* (named for Danish physicist and mathematician Ludvig Lorenz) has often mistakenly been called the *Lorentz gauge*, after the better-known Dutch physicist Hendrik Lorentz (who shared the 1902 Nobel Prize in physics for the theoretical explanation of the Zeeman effect). The present author must confess to being among the many who have made this error, referring to the "Lorentz" rather than "Lorenz" gauge in the book [9]. Ludvig Lorenz did significant work on electromagnetism contemporary with, but indepen-

the equations decouple and solving the Maxwell equations has now been reduced to solving two second-order differential equations,

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = -\frac{\rho}{\varepsilon_0}, \tag{7.12a}$$

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J}, \tag{7.12b}$$

which are *uncoupled:* solution of Eq. (7.12a) gives the scalar potential Φ independent of the vector potential $\boldsymbol{A}$, while solution of Eq. (7.12b) gives the vector potential, independent of the scalar potential. Equations (7.12) along with the gauge condition (7.11) are completely equivalent in physical content to the original Maxwell equations (7.4).

It is always possible to find potentials $\{\Phi,\boldsymbol{A}\}$ that satisfy the Lorenz condition. Suppose that we have a solution with gauge potentials that satisfies Eqs. (7.9), but does not satisfy Eqs. (7.11). Then, make a gauge transformation to new potentials $\{\Phi',\boldsymbol{A}'\}$ and require the new potentials to satisfy the Lorentz condition

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}' + \frac{1}{c^2}\frac{\partial\Phi'}{\partial t} = 0 = \boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t} + \nabla^2\chi - \frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2}.$$

Thus, if a scalar function χ can be found that satisfies

$$\nabla^2\chi - \frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2} = -\left(\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t}\right), \tag{7.13}$$

the new potentials will satisfy the Lorenz gauge conditions (7.11) and the simultaneous equations (7.12).

The Lorenz condition Eq. (7.11) does not exhaust the gauge degrees of freedom in Lorenz gauge. The *restricted gauge transformation*

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \to \Phi - \frac{\partial\chi}{\partial t}, \tag{7.14}$$

where the scalar function χ (which is arbitrary for general gauge transformations) is restricted to those that satisfy the constraint

$$\nabla^2\chi - \frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2} = 0, \tag{7.15}$$

preserves the Lorenz condition if $\{\boldsymbol{A},\Phi\}$ satisfies it to begin with. Thus the Lorenz gauge corresponds to an entire family of of gauge conditions that satisfy Eq. (7.11). The Lorenz gauge is often used for two reasons:

1. It leads to the decoupled equations (7.12) that treat Φ and $\boldsymbol{A}$ on an equal footing.

dent of, Maxwell. For example, he proposed independently that electromagnetic waves might be light waves. An overview of Lorenz's scientific work is given in Ref. [19] and the history of his work and the incorrect attribution of his famous gauge to Lorentz is discussed in Ref. [16].

2. As will be elaborated in Ch. 8, the Lorenz gauge condition is manifestly invariant under Lorentz transformations, which fits naturally into special relativity.[5]

There are infinitely many valid gauge transformations that can be made using Eq. (7.10) with different choices for the scalar function χ. However, only some prove to be useful. One of use is the transformation to Lorenz gauge described in this section. Another that we have already encountered in Section 5.5 is the transformation to Coulomb gauge.

7.3.2 Coulomb Gauge

We have already introduced the *Coulomb gauge condition* in Eq. (5.27),

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}=0 \qquad \text{(Coulomb gauge)}.$$

From Eq. (7.9a), in Coulomb gauge the scalar potential obeys a Poisson equation

$$\nabla^2\Phi=-\frac{\rho}{\varepsilon_0}, \tag{7.16}$$

which has a solution

$$\Phi(\boldsymbol{x},t)=\frac{1}{4\pi\varepsilon_0}\int\frac{\rho(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{7.17}$$

This is the instantaneous Coulomb potential caused by the charge density $\rho(\boldsymbol{x})$, which is the source of the name *Coulomb gauge*.

As Box 7.2 discusses, the Coulomb potential (7.17) suggests instantaneous transmission of information in a universe where lightspeed c is the speed limit. In Chs. 8-10 we will resolve this issue and show that classical electromagnetism is in fact causal because *no information is being transmitted at a speed $v>c$*.

From Eq. (7.9b), in Coulomb gauge the vector potential obeys

$$\nabla^2\boldsymbol{A}-\frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2}=-\mu_0\boldsymbol{J}+\frac{1}{c^2}\frac{\partial\Phi}{\partial t}. \tag{7.18}$$

As guaranteed by the Helmholtz theorem for any vector (see Box 3.1), the current density can be decomposed as a sum of two terms

$$\boldsymbol{J}=\boldsymbol{J}_{\mathrm{L}}+\boldsymbol{J}_{\mathrm{T}}, \tag{7.19}$$

where the terms have the following properties:

1. The component $\boldsymbol{J}_{\mathrm{L}}$ has vanishing curl, $\boldsymbol{\nabla}\times\boldsymbol{J}_{\mathrm{L}}=0$; it is called the *longitudinal current* or the *irrotational current*.
2. The component $\boldsymbol{J}_{\mathrm{T}}$ has vanishing divergence, $\boldsymbol{\nabla}\cdot\boldsymbol{J}_{\mathrm{T}}=0$; it is called the *transverse current* or the *solenoidal current*.

[5] Note that the equations are in (Ludvig) *Lorenz gauge*, but we shall see in Ch. 8 that they are invariant under (Hendrik) *Lorentz transformations*. This perhaps helps to explain the historical confusion in the name of the gauge.

The longitudinal and transverse currents are given explicitly by

$$\boldsymbol{J}_{\mathrm{L}} = -\frac{1}{4\pi}\boldsymbol{\nabla}\int \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \tag{7.20a}$$

$$\boldsymbol{J}_{\mathrm{T}} = \frac{1}{4\pi}\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\int \frac{\boldsymbol{J}(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{7.20b}$$

From Eq. (7.17) and the continuity equation (1.3),

$$\frac{1}{c^2}\boldsymbol{\nabla}\frac{\partial\Phi}{\partial t} = \mu_0\boldsymbol{J}_{\mathrm{T}}, \tag{7.21}$$

which gives when inserted into Eq. (7.9b), using the identity $\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) = 0$ and Eq. (7.19),

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J}_{\mathrm{T}}. \tag{7.22}$$

Thus, in Coulomb gauge the source for the equation for $\boldsymbol{A}$ can be expressed entirely in terms of the transverse current $\boldsymbol{J}_{\mathrm{T}}$. For this reason, the Coulomb gauge is sometimes termed the *transverse gauge*. The Coulomb gauge is also sometimes called the *radiation gauge*, because one finds in quantum electrodynamics that only the vector potential (which is determined by the transverse components) need be quantized.

Notice that Eq. (7.22) has the form of a wave equation with the speed of the wave equal to c, which is the behavior expected for electromagnetic waves. However, the discussion above implies that for the scalar potential the propagation speed is infinite. The resolution of this seeming paradox requires relativistic quantum field theory, but in essence the propagating classical field has only transverse components and one finds that in QED only the vector potential (and thus only the transverse components) need be quantized.

As we will see in Ch. 8, Lorentz covariance requires the 3-vector potential $\boldsymbol{A}$ and the scalar potential Φ to be combined into a spacetime 4-vector A^μ with the components of the 4-vector given by

$$A^\mu = (A^0, A^1, A^2, A^3) = (\Phi, \boldsymbol{A}) = (\Phi, A^1, A^2, A^3), \tag{7.23}$$

where Φ is the scalar potential and $\boldsymbol{A}$ is the 3-vector potential with components $A^i (i = 1,2,3)$. In the 4-vector the first component A^0 is called the *timelike component* and the other three components (A^1, A^2, A^3) are called the *spacelike components* of the 4-vector.[6]

We have asserted above that only the *two transverse components* of the vector (typically chosen to be A_1 and A_2 for $\boldsymbol{A}$ and A^2 and A^3 for A^μ) are required to describe propagating waves. But a 3-vector like $\boldsymbol{A}$ normally has three components, and a 4-vector like A^μ has four components. So how can a propagating photon have only two rather than three or four degrees of freedom. Relativistic quantum field theory applied to electrons and photons (QED) is beyond our present scope in these lectures. However, the essential answer is

[6] A small terminology problem now enters our discussion. We have been calling $\boldsymbol{A}$ the vector potential, but relativistically it will be more convenient to work with the 4-vector potential A^μ (which makes more obvious the requirement of relativity that space and time enter on an equal footing). We adopt a policy that where it is clear that we are working non-relativistically we will often call $\boldsymbol{A}$ the vector potential, while if it is clear that we are working in a relativistic context we will often call A^μ the vector potential. If there is chance for confusion we may use the explicit names "3-vector potential" for $\boldsymbol{A}$ and "4-vector potential" for A^μ to be clear.

that it is found in QED that in a covariant gauge like Lorenz gauge there are four states of polarization, but the contributions from timelike and longitudinal polarizations enter with equal magnitudes but opposite signs and exactly cancel each other, leaving only the transverse contributions for a free propagating photon. (This is called the *Gupta–Bleuler mechanism* in quantum electrodynamics.) A massive vector field would have three spatial polarization components. As we will see, the reduction to two spatial polarizations is a consequence of the photon being identically massless: massless vector fields have two rather than three states of polarization.

7.4 Retarded Green Function

From Eqs. (7.12), it is clear that the key to solving Maxwell's equations in Lorenz gauge is to be able to solve the wave equation with a source f,

$$\Box\psi = -f, \tag{7.24}$$

where we now introduce for convenience the d'Alembertian operator $\Box$ with[7]

$$\Box \equiv -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2, \tag{7.25}$$

since if we know how to solve (7.24), we can solve Eqs. (7.12). Since we are assuming Lorenz gauge, we must also satisfy Eq. (7.11), but the retarded solution of Eqs. (7.12) with sources that we will find shortly will in fact satisfy Eq. (7.11) automatically, provided that they decrease rapidly enough at infinity. Generalizing the electrostatics case, a Green function $G(t,\boldsymbol{x};t'\boldsymbol{x}')$ can be defined by

$$\Box G(t,\boldsymbol{x};t'\boldsymbol{x}') = -\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t-t'), \tag{7.26}$$

where the derivative operators in $\Box$ are understood to operate on the unprimed variables, and in contrast to Eq. (3.30) for electrostatics, the Green function depends on $(t',\boldsymbol{x}')$ and $(t,\boldsymbol{x})$ and there is a delta function in $t'-t$ as well as in $\boldsymbol{x}-\boldsymbol{x}'$. If we can obtain a Green function, a solution ψ of Eq. (7.24) is given by

$$\psi(t,\boldsymbol{x}) = \int G(t,\boldsymbol{x};t'\boldsymbol{x}')f(t',\boldsymbol{x}')\,d^3x'\,dt', \tag{7.27}$$

assuming that the integral converges.

Let us seek a solution of Eq. (7.26) using *Fourier transforms*. For an integrable function $F:\mathbb{R}\to\mathbb{R}$, define its Fourier transform $\hat{F}$ as

$$\hat{F}(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} F(x)e^{-ikx}dx. \tag{7.28}$$

[7] The d'Alembertian operator $\Box$ is a Lorentz-invariant combination of the second derivatives with respect to space and time that will play a significant role in discussing the relationship of the Maxwell equations to special relativity in Ch. 8.

Fourier transforms can be extended to distributions such as the Dirac delta function, as well as functions. For example, the Fourier transform of the delta function is,

$$\hat{\delta}_{x_0} = \frac{1}{\sqrt{2\pi}} e^{ilx_0}. \tag{7.29}$$

Assuming a function to be smooth and to fall off fast enough at infinity, the original function $F(x)$ can be recovered by the *inverse Fourier transform*,

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{F}(k) e^{+ikx} dk. \tag{7.30}$$

This formula applies also to distributions and the inverse of the Fourier transform for a delta function is

$$\delta_{x_0}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{\delta}_{x_0}(k) e^{+ikx} dk = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ikx_0} e^{+ikx} dk. \tag{7.31}$$

An important property of Fourier transforms is that differentiation in real space corresponds to multiplication by *ik* in Fourier transform space. For example, as you are asked to show in Problem 7.1,

$$\widehat{\frac{dF}{dx}}(k) = ik\hat{F}(k), \tag{7.32}$$

for the Fourier transform of $dF(x)/dt$.

Thus any partial differential equation with constant coefficients in real space can be converted to an algebraic equation in Fourier transform space.

Now let's try to solve Eq. (7.26) for G using Fourier transforms. To simplify notation we temporarily set $x' = t' = 0$ and $c = 1$, and define the 4D Fourier transform of G as[8]

$$\hat{G}(\omega, \boldsymbol{k}) = \frac{1}{(2\pi)^2} \int_{-\infty}^{+\infty} G(t, \boldsymbol{x}) e^{+i\omega t} e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} dt\, d^3x \tag{7.33}$$

Taking the Fourier transform of Eq. (7.26) with respect to t and $\boldsymbol{x}$, using Eq. (7.29) with $x_0 = 0$, and using Eq. (7.32) yields

$$(\omega^2 - k^2)\, \hat{G}(\omega, \boldsymbol{k}) = -\frac{1}{4\pi^2}, \tag{7.34}$$

where $k \equiv |k|$. Naively, this suggest the solution

$$\hat{G}(\omega, \boldsymbol{k}) = -\frac{1}{4\pi^2} \frac{1}{(\omega^2 - k^2)} = -\frac{1}{4\pi^2} \frac{1}{(\omega + k)(\omega - k)}, \tag{7.35}$$

but division by $(\omega^2 - k^2)$ is illegal since it is possible that $\omega = k$. To see the difficulty

[8] By convention the time Fourier transform is defined by integrating with $e^{+i\omega t}$ rather than with $e^{-i\omega t}$ for compatibility with the 4-momentum vector $k^\mu = (\omega/c, \boldsymbol{k})$ of special relativity that will be introduced in Ch. 8.

clearly, let's attempt to take the inverse transform of $\hat{G}$ with respect to ω (but not $\boldsymbol{k}$). This is a Fourier transform with respect to space but not time; denoting it as $\tilde{G}$,

$$\begin{aligned}\tilde{G}(t,\boldsymbol{k}) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\hat{G}(\omega,\boldsymbol{k})e^{-i\omega t}\,d\omega. \\ &= -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k)(\omega-k)}\,d\omega, \qquad (7.36)\end{aligned}$$

where Eq. (7.35) was used. This has *logarithmic divergences*[9] of the integral if $\omega \to k$, so it is ill-defined.[10] To proceed it is necessary to regularize the integration in Eq. (7.36) in such a way that

1. equation (7.34) remains valid, and
2. the right side of Eq. (7.36) becomes well defined.

One way to do this is to displace the poles at $\omega = \pm k$ into the complex ω plane and evaluate the resulting contour integral. There is more than one way to do this, each leading to a different Green function. The Green function that we seek is the *retarded Green function*, which will be appropriate for the situation where there is only outgoing radiation from a source. To get the retarded Green function the poles should be displaced into the lower half of the complex ω-plane; thus we define the retarded Green function by

$$\tilde{G}(t,\boldsymbol{k})_{\text{ret}} = -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k+i\varepsilon)(\omega-k+i\varepsilon)}\,d\omega, \qquad (7.37)$$

where ε is positive and the limit $\varepsilon \to 0$ is to be taken after the integral is evaluated. Viewing the integral in Eq. (7.37) as a contour integral in the complex ω-plane, if $t < 0$ the exponential is damped in the upper half-plane of ω and the countour integral can be closed in the upper half plane. Since the poles are in the negative half-plane, the contour does not enclose them and by the Cauchy theorem,

$$\tilde{G}(t,\boldsymbol{k})_{\text{ret}} = 0 \qquad (t < 0). \qquad (7.38)$$

This vanishing of the Green function before the time $t = 0$ is characteristic of the retarded solution, and corresponds to a solution with outgoing but no incoming radiation. This solution is relevant physically if there is no radiation present when the source is turned on.

By similar reasoning, for $t > 0$ the contour can be closed in the lower half-plane. Now the contour encloses poles at $\omega = \pm k - i\varepsilon$ and by the Cauchy theorem the integral evaluates

[9] The simplest example of a logarithmic divergence occurs in an integral of the form

$$f(x) = \int_{x_0}^{x}\frac{1}{\Lambda}\,d\Lambda.$$

Such an integral diverges as $x \to \infty$, but rather mildly, growing as $f(x) \sim \log(x)$ at large x.

[10] The basic reason for the ambiguity is that many different Green functions satisfy Eq. (7.26), so we cannot expect to solve for G without providing further information to identify the Green function that we are after.

to $2\pi i$ times a sum of residues at the two poles.[11] From Eq. (7.37),

$$\tilde{G}(t,\boldsymbol{k})_{\text{ret}} = \frac{2\pi i}{4\pi^2\sqrt{2\pi}}\left(\frac{e^{-ikt}}{2k} - \frac{e^{+ikt}}{2k}\right) = \frac{1}{2\pi\sqrt{2\pi}}\frac{\sin kt}{k} \qquad (t > 0), \tag{7.39}$$

The retarded Green function in real (position) space then results from taking the inverse transform of Eq. (7.39) with respect to k (homework problem),

$$\begin{aligned} G(t,\boldsymbol{x})_{\text{ret}} &= \frac{1}{(2\pi)^{3/2}}\frac{1}{2\pi\sqrt{2\pi}}\int \frac{\sin kt}{k}e^{i\boldsymbol{k}\cdot\boldsymbol{x}}d^3x \\ &= \frac{\delta(t-|\boldsymbol{x}|)}{4\pi|\boldsymbol{x}|} \qquad (t>0). \end{aligned} \tag{7.40}$$

Restoring t', $\boldsymbol{x}'$, and c, this result becomes

$$G(t,\boldsymbol{x};t',\boldsymbol{x}')_{\text{ret}} = \begin{cases} 0 & (t<t'), \\ \dfrac{\delta(t-t'-|\boldsymbol{x}-\boldsymbol{x}'|/c)}{4\pi|\boldsymbol{x}-\boldsymbol{x}'|} & (t>t'). \end{cases} \tag{7.41}$$

This propagator now has the desired causal behavior. Using the language of special relativity (see Ch. 8), it is nonvanishing only on the future lightcone of the source point $(t',\boldsymbol{x}')$, meaning that it is non-vanishing only if $|\boldsymbol{x}-\boldsymbol{x}'| = c(t-t')$. That is, if a field satisfying the wave equation $\Box\psi = -f$ of Eq. (7.24) vanishes at early times and a δ-function source is placed at $(t',\boldsymbol{x}')$, the resulting disturbance of the field (electromagnetic waves) will propagate away from the source at the speed of light. Thus, light waves correspond to a wave solution of the Maxwell equations.

The retarded solution of Eq. (7.24) is the solution obtained using the retarded Green function in Eq. (7.27),

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t-|\boldsymbol{x}-\boldsymbol{x}'|/c,\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'. \tag{7.42}$$

This can also be written more compactly as

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t',\boldsymbol{x}'))_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x', \tag{7.43}$$

where the notation means that $f(t',\boldsymbol{x}')$ is to be evaluated at the retarded time

$$t' = t - \frac{|\boldsymbol{x}-\boldsymbol{x}'|}{c}, \tag{7.44}$$

and where t' isn't independent but rather is a function of t, $\boldsymbol{x}$, and $\boldsymbol{x}'$.[12] Written in this form

[11] The residue $\operatorname{res} f(a)$ of a function $f(z)$ at a pole $z=a$ is given by

$$\operatorname{res} f(a) = \frac{1}{(m-1)!}\lim_{z\to 0}\left(\frac{d^{m-1}}{dz^{m-1}}(z-a)^m f(z)\right),$$

where m is the order of the pole (exponent on the denominator factor tending to zero at the pole) and a is the location of the pole (with $a \neq \infty$). In Eq. (7.37) the poles are of order $m=1$.

[12] In Eq. (7.43), the integration is not over all of space at an instant of time (as the notation might suggest). Rather is restricted to the past lightcone of the spacetime point $(t,\boldsymbol{x})$, since only points on the past lightcone can be causally connected to $(t,\boldsymbol{x})$ by a signal traveling at lightspeed (see the discussion of lightcones in Ch. 8).

the retarded solution looks like a solution of the Poisson equation, but evaluated at the retarded time (7.44).

The Maxwell equations (7.12a) and (7.12b) for the scalar and vector potentials are of the form (7.24), so we can write immediately the corresponding retarded solutions,

$$\Phi(t,\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{[\rho(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \tag{7.45a}$$

$$\boldsymbol{A}(t,\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{J}(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{7.45b}$$

Finally, substitution of Eqs. (7.45a) and (7.45b) in Eq. (7.11) and some manipulation indicates that these solutions are indeed consistent with the Lorenz gauge condition. Therefore, we may conclude that the solution of the Maxwell equations for a charge density ρ and a current density $\boldsymbol{J}$, with initial conditions of no incoming radiation, is given by Eqs. (7.45). This has been shown here in Lorenz gauge, but the Maxwell equations are invariant under gauge transformations so this solution may be taken to be valid generally.

Background and Further Reading

Much of this chapter is based on the presentations in Jackson [15], Garg [6], and Wald [25]. Further useful discussion may be found in Refs. [3, 8, 27].

Box 7.2 Causality and Coulomb and Gravitational Potentials

Causality (that actions must precede results) is a fundamental principle underlying all of modern science. The Coulomb potential of Eq. (7.17) appears to violate that principle.

Causality and the Coulomb Potential

The Coulomb potential in Eq. (7.17) is said to be "instantaneous" because the associated force acts without delay at any $\boldsymbol{x}$ corresponding a distance $|\boldsymbol{x}-\boldsymbol{x}'|$ from the source. This is inconsistent with special relativity and relativistic quantum field theory, which require that a force be transmitted by a virtual particle (the photon in this case) communicated at a speed less than or equal to the speed of light c.

Causality and the Newtonian Gravitation Potential

The Newtonianian gravitational potential has the form of a Coulomb potential with masses playing the role of charges, and has the same causality problem: it implies that the gravitational force acts instantaneously over any distance. The solution of this problem in gravitational physics is replacement of Newtonian gravity with general relativity, which generalizes special relativity and requires that the transmission speed of the gravitational force be equal to the speed of light. This prediction of general relativity has been confirmed by observations, as we now describe.

The Speed of Light and the Speed of Gravity

In 2017 the ground-based Ligo–Virgo detectors observed gravitational wave GW170817.[a] But there was more to come: 1.7 seconds later the Fermi Gamma-ray Space Telescope (Fermi) and International Gamma-Ray Astrophysics Laboratory (INTEGRAL) in orbit around Earth detected a gamma-ray burst from the same portion of the sky as the source of the gravitational wave. Detailed observations of the afterglow of the gamma-ray burst at many wavelengths, and comprehensive analysis of the gravitational and electromagnetic data sets, concluded that the gravitational wave and the gamma-ray burst were caused by a binary neutron star merger in the galaxy NGC 4993, at a distance of 40 megaparsecs (130 million lightyears).

That the gravitational and electromagnetic signals arrived within 1.7 seconds of each other after traveling 40 megaparsecs implied that the speed of gravity (for the gravitational waves) and the speed of light c (for the gamma-rays) differ by no more than 3 parts in 10^{15}. Thus *the speed of gravity is* c, just as predicted by the general theory of relativity. More extensive discussion of physical implications for GW170817 may be found in Ref. [10] and in Section 24.7 of Ref. [11].

[a] The name GW170817 indicates that the gravitational wave (GW) was observed in the year 20(17), in the (08)th month of that year, on the (17)th day of that month.

Problems

7.1 An important feature of Fourier transforms is that differentiation in real space corresponds to multiplication by ik in Fourier transform space. As an example, prove that

$$\widehat{\frac{dF}{dx}}(k) = ik\hat{F}(k),$$

for the Fourier transform of dF/dx.

7.2 Supply and explain the steps in verifying the result of Eq. (7.40).

8 Minkowski Spacetime

These lectures constitute a graduate-level course in classical electromagnetism. The distinction between classical and modern physics turns on whether the dynamics are described by quantum mechanics and quantum field theory or classical mechanics and classical field theory, and by non-relativistic mechanics or relativistic mechanics. Traditionally classical electromagnetism excludes quantum mechanics and quantum field theory (except for a few quantum concepts that must be imported for parts of the discussion involving the microscopic structure of matter), but special relativity has been considered part of the discussion of classical electromagnetism. Accordingly, this chapter will introduce Minkowski spacetime and a differential geometry and spacetime tensor formalism that is the mathematical foundation of the theory of relativity. Then in Ch. 9 we will introduce the special theory of relativity and demonstrate explicitly the Lorentz invariance of Maxwell's equations. Finally, Ch. 10 will use the gauge-invariant and Lorentz-invariant formulation of classical electromagnetism in terms of the Maxwell equations to paint electromagnetism as a relativistic gauge field theory that, when quantized, is the harbinger of the Standard Model of elementary particle physics.

It is possible to introduce special relativity in a minimal way and then use that to demonstrate the Lorentz invariance of the Maxwell equations in less space than we will use here. However, we choose to give a more thorough introduction to the differential geometry and tensor formalism underlying the theory of relativity. In particular, the tensor formalism will be developed in a way that is compatible with curved spacetime, and therefore with general relativity. Then, we will approximate by restricting to flat spacetime to recover the theory of special relativity. This approach has several advantages:

1. It does not take much longer to develop the formalism in a way compatible with either flat or curved spacetime than to describe special relativity minimally.
2. The resulting formalism gives more satisfying insight into special relativity and its relationship with Newtonian mechanics, general relativity, and the Maxwell equations.
3. Developing the formalism in this way gives the reader a solid foundation to take on more advanced topics such as general relativity.

Let us begin with an overview of Minkowski spacetime.

8.1 Minkowski Spacetime and Spacetime Tensors

The most elegant formulation of special relativity is in terms of a 4-dimensional spacetime manifold called *Minkowski space*, and in terms of *spacetime tensors* that are a consequence of the differential geometry of that manifold. Let us review Minkowski space and spacetime tensors defined for that manifold.

8.1.1 Transformations between Inertial Systems

In 1905 Einstein published the *special theory of relativity*, which was to revolutionize our conception of space and time. His motivation for the special theory was a conviction that the laws of physics must be independent of coordinate system for the observer (the *principle of relativity*), and that transformations between inertial frames (coordinate systems in which Newton's first law is valid) should be the same for particles and for light. Newtonian mechanics already implied a principle of relativity for space: the laws of mechanics were unchanged by a *Galilean transformation* between inertial frame. For motion along the x axis a Galilean transformation takes the form

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad t' = t, \tag{8.1}$$

where primed coordinates and unprimed coordinates represent the two inertial frames, the velocity is v, and a single universal time $t = t'$ is assumed for all inertial frames. But Einstein (as well as others such as Lorentz) realized that there is a problem in that these common-sense notions of relative motion between inertial frames were consistent with the motion of billiard balls and projectiles, but were inconsistent with the theory of light, which was well understood in 1905 to be an electromagnetic wave described by Maxwell's equations.

A striking aspect of the Maxwell theory was that it admitted wave solutions and these waves traveled with a speed that was a *constant of the theory* (and thus independent of inertial frame for the observer). When the constant was evaluated it was found to be equal to the speed of light, which led to Maxwell's electromagnetic waves being identified with light. The beauty of Maxwell's equations greatly impressed Einstein, but they presented a problem of interpretation for classical physics. By the Galilean transformations, the speed of light should depend upon the inertial frame of the observer; but not by Maxwell's equations because the speed of light is a constant of the theory. The interpretation that emerged to reconcile this discrepancy was that electromagnetic waves must move through some medium. (How could a wave travel through nothing?) This hypothesized medium was called the *aether*, and it possessed quite magical properties: it was required to be an invisible, rigid (because transverse light waves don't propagate through fluids) substance permeating all of space, but relevant only for light propagation, so as not to disturb other physical laws. Then the constant speed of light could be understood as an artifact of the special aether rest frame in which light propagated.

It is now understood that light waves are propagating disturbances in electric and magnetic fields that do not require a physical medium, and that the aether is a fiction. But in the latter part of the 19th century it was widely believed to exist, and various attempts were made to detect the motion of the Earth relative to the aether (an effect called the *aether drift*). Using light interferometry, Michelson and Morley showed in 1887 that there was no evidence for motion of the Earth with respect to the hypothetical aether, thus casting serious doubt on its existence.[1]

8.1.2 The Special Theory of Relativity

Without the aether fiction, Maxwell's equations for light and the Galilean invariance exhibited by material particles were clearly at odds. Others had hinted at a solution (notably Hendrik Lorentz and Henri Poincaré) but it was Einstein who concluded that the Maxwell theory was correct and that the *Galilean transformations* required modification. Thus he put forth the bold and unqualified assertion that the speed of light was constant *in all inertial frames*. This, along with the relativity assumption (physical law is the same in all coordinate systems) yielded in 1905 the special theory of relativity.

In the special theory (of relativity), requiring the speed of light c be an invariant independent of inertial frame dictated replacement of Galilean transformations (8.1) with *Lorentz transformations,*

$$\begin{aligned} x' &= \gamma(x - vt) \qquad y' = y \qquad z' = z, \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right) \qquad \gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}}, \end{aligned} \tag{8.2}$$

where γ is called the *Lorentz γ-factor.* Notice that the Galilean transformations remain quite correct in the low-velocity world since in the limit $v/c \to 0$ the factor $\gamma \to 1$ and the Lorentz transformations (8.2) become equivalent to the Galilean transformations (8.1). Notice also that time transforms non-trivially under Lorentz transformations, with the Lorentz transformations *mixing the space and time coordinates*. This is in stark contrast to the Galilean transformations, where there is a universal time shared by all observers.

The mathematician Hermann Minkowski (once Einstein's teacher) then proposed that in special relativity separate notions of space and time should be abandoned in favor of a *4-dimensional spacetime* parameterized by *spacetime coordinates*

$$(x^0, x^1, x^2, x^3) \equiv (ct, x, y, z), \tag{8.3}$$

where the superscripts are indices, not exponents. (The reason for placement of indices as superscripts will explained shortly.) In a 1908 presentation entitled *Raum und Zeit* (*Space and Time*) Minkowski introduced 4-dimensional spacetime using a now-legendary phrasing:

[1] However, efforts persisted for years to salvage the aether hypothesis. Einstein originally took the aether hypothesis seriously, but abandoned it at some point before he published the special theory of relativity. Einstein claimed that the Michelson–Morley result had no influence on his formulation of the special theory of relativity. It is likely that Einstein knew of the Michelson–Morley results, but seems to have felt that this was not as important as his own reasoning in coming to the special theory of relativity. For example, Einstein emphasized the role of a thought experiment that he first carried out as a student concerning whether one could travel fast enough to catch up with a light wave, and what the observational consequences would be.

> *The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.*
>
> Hermann Minkowski (1908)

Now time, scaled by the speed of light so that it has the same units as the other three coordinates, is just another coordinate in 4-dimensional spacetime. Minkowski's formulation showed that special relativity is simple in 4-dimensional spacetime, but becomes complicated when projected onto 3-dimensional space.[2]

8.1.3 Minkowski Space

Allowing the coordinates (x^0, x^1, x^2, x^3) to range over all their possible values traces out the manifold of 4-dimensional spacetime called *Minkowski space.* In this space the square of the infinitesimal distance ds^2 between two points (ct, x, y, z) and $(ct + cdt, x + dx, y + dy, z + dz)$ is given by

$$\begin{aligned} ds^2 = \sum_{\mu\nu} \eta_{\mu\nu} dx^\mu dx^\nu &= -c^2dt^2 + dx^2 + dy^2 + dz^2, \\ &= -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2, \end{aligned} \tag{8.4}$$

which is called the *line element* of the Minkowski space. Notice that in this equation ds^2 means $(ds)^2$, and dx^2 means $(dx)^2$, but the superscripts in (dx^0, dx^1, dx^2, dx^3) are indices and not powers. The *metric tensor* of Minkowski space $\eta_{\mu\nu}$ appearing in Eq. (8.4) may be expressed as the diagonal matrix

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.5}$$

The line element (8.4) or the metric tensor $\eta_{\nu\mu}$ determine the geometry of Minkowski space because they specify distances, distances can be used to define angles, and that is geometry.

The pattern of plus and minus signs on the right side of Eq. (8.4), or on the diagonal of the matrix in Eq. (8.5), defines the *signature of the metric.* For Minkowski space the signature that we adopt is $(- + + +)$. Some authors define the signature as the difference of the number of $+$ and $-$ signs, which conveys similar information. A metric such as (8.5) where the signs in the signature pattern are not all the same is termed an *indefinite metric*. Equally common is the choice $(+ - - -)$ for the Minkowski-space signature, which leads

[2] An English translation of the full presentation may be found at `https://en.wikisource.org/wiki/Translation:Space_and_Time`. Minkowski udoubtedly would have made further important contributions to the development of special and general relativity, but he died unexpectedly of peritonitis only months after his famous Space and Time lecture.

to the same physical results as our choice if all signs are carried through consistently. The central point for the indefinite metric of Minkowski spacetime is that the last three terms have the same sign and that the sign of the first term differs from that of the other three (assuming the usual modern convention of displaying the timelike coordinate in the first position and the spacelike coordinates in the last three positions).

> Minkowski spacetime is *not* “just like ordinary space but with more dimensions”, because the geometry of Minkowski space differs fundamentally from that of 4-dimensional Euclidean space. The difference is encoded in the metric signature, which is just the signature of the unit matrix $(+ + + +)$ for 4-dimensional euclidean space, compared with the indefinite-metric signature $(- + + +)$ for the Minkowski metric.

That change in sign between timelike and spacelike components of the metric signature for spacetime has enormous implications. Most of the surprising features of special relativity (space contraction, time dilation, relativity of simultaneity, the twin “paradox”, $\ldots$) follow directly from this difference in geometry relative to euclidean space implied by the Minkowski indefinite metric.

8.2 Symmetry under Coordinate Transformations

A physical system has a symmetry under some operation if after the operation the system is indistinguishable from the system before the operation. The theory of relativity may be viewed as a symmetry under coordinate transformations. Relativity is ultimately a statement that physics is independent of our choice of coordinate system: two observers, referencing their measurements to two different coordinate systems, should deduce from their observations the *same laws of physics*. General relativity requires invariance of the laws of physics under the most general possible coordinate transformations, while special relativity requires a symmetry under only the subset of coordinate transformations that are between inertial frames. Hence, to understand relativity it is important consider coordinate systems, transformations between coordinate systems, and the properties of those transformations.

8.3 Euclidean Coordinates and Transformations

Our ultimate goal is to describe coordinates and transformations between coordinates in a general (possibly curved) space having a Minkowski metric with three spacelike coordinates and one timelike coordinate. However, to introduce these concepts it is useful to begin with the simpler case of vector fields in euclidean space [5].

8.3.1 Parameterizing in Different Coordinate Systems

Assume a 3D euclidean space having a cartesian coordinate system (x,y,z), with a set of mutually orthogonal unit vectors $(\boldsymbol{i},\boldsymbol{j},\boldsymbol{k})$ pointing in the x, y, and z directions, respectively. Assume also an alternative coordinate system (u,v,w), perhaps not cartesian, with the (x,y,z) and (u,v,w) coordinates related by functional relationships

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w). \tag{8.6}$$

Further, we assume the transformations to be invertible so that we can solve for (u,v,w) in terms of (x,y,z).

Example 8.1 Let (u,v,w) correspond to spherical coordinates (r,θ,ϕ), so that Eq. (8.6) takes the familiar form

$$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\cos\theta, \tag{8.7}$$

with $r \geq 0$ and $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$.

It will prove useful to combine Eqs. (8.6) into an equation giving a position vector $\boldsymbol{r}$ for a point in terms of the (u,v,w) coordinates:

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k}. \tag{8.8}$$

For example, in terms of the spherical coordinates (r,θ,ϕ),

$$\boldsymbol{r} = (r\sin\theta\cos\phi)\,\boldsymbol{i} + (r\sin\theta\sin\phi)\,\boldsymbol{j} + (r\cos\theta)\,\boldsymbol{k}. \tag{8.9}$$

The second coordinate system in these examples generally may be non-cartesian but the space still is assumed to be intrinsically euclidean (not curved). The transformation from the (x,y,z) coordinates to the (r,θ,ϕ) coordinates just gives two different schemes to label points in the same flat space.

8.3.2 Basis vectors

Vectors are *geometrical objects:* they exist independent of representation in any particular coordinate system. However, it is often useful to express vectors in terms of components within a specific coordinate system by defining basis vectors that permit arbitrary vectors to be expanded in that basis. Equations (8.8) and (8.9) are familiar examples, where an arbitrary vector has been expanded in terms of the three orthogonal cartesian unit vectors $(\boldsymbol{i},\boldsymbol{j},\boldsymbol{k})$. For a flat manifold, a single basis can be chosen that applies to all points in the space. If the space is curved, or if it is represented in non-cartesian coordinates, it is often useful to define basis vectors at individual points of the space, as we shall now describe.

Parameterized curves and surfaces: At any point $P(u_0,v_0,w_0)$, three surfaces pass, which may be defined by setting $u = u_0$, $v = v_0$, or $w = w_0$, respectively. The intersections of

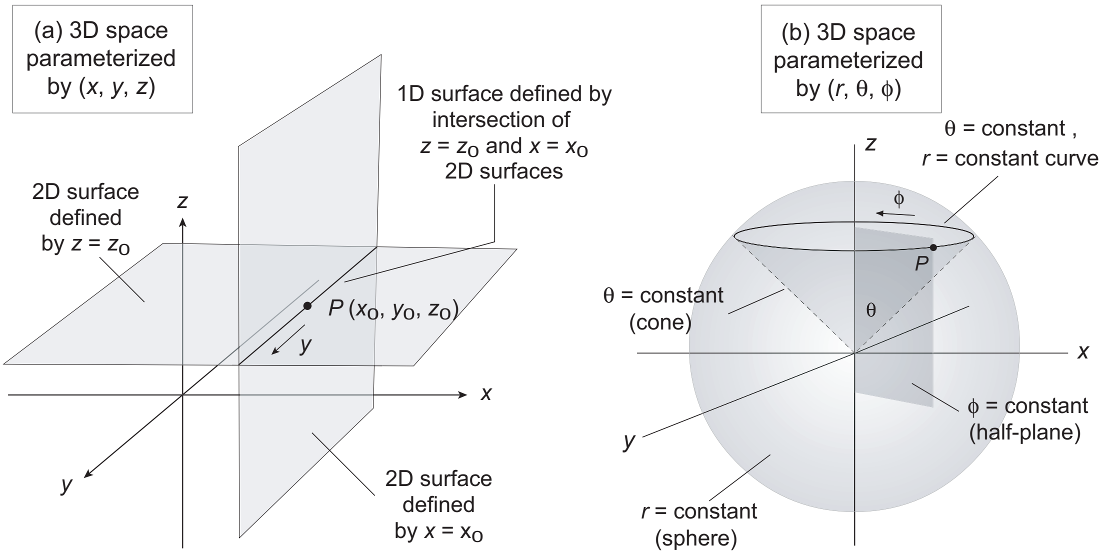

(a) 3D space parameterized by (x, y, z)
1D surface defined by intersection of z = zo and x = xo 2D surfaces
2D surface defined by z = zo
z
P (xo, yo, zo)
y
x
y
2D surface defined by x = xo
(b) 3D space parameterized by (r, θ, ϕ)
θ = constant , r = constant curve
z
ϕ
P
θ = constant (cone)
θ
x
ϕ = constant (half-plane)
y
r = constant (sphere)

Fig. 8.1 Examples of surfaces and curves arising from constraints. (a) In 3D euclidean space parameterized by cartesian coordinates (x,y,z), the constraints $x = x_0$ and $z = z_0$ define 2D planes and the intersection of these planes defines a 1D surface parameterized by the variable y. (b) In 3D space described in spherical coordinates (r,θ,ϕ), the constraint $r = \text{constant}$ defines a 2D sphere, the constraint $\theta = \text{constant}$ defines a cone, and the constraint $\phi = \text{constant}$ defines a half-plane. The intersection of any two of these surfaces defines a curve parameterized by the variable not being held constant.

these three surfaces define three curves passing through $P(u_0, v_0, w_0)$. General parametric equations for coordinate surfaces may be obtained from Eq. (8.8) by setting one of the variables (u, v, w) equal to a constant, and for curves by setting two variables to constants. For example, setting v and w to constant values, $v = v_0$ and $w = w_0$, yields an equation for a curve given by the intersection of $v = v_0$ and $w = w_0$,

$$\boldsymbol{r}(u) = x(u, v_0, w_0)\,\boldsymbol{i} + y(u, v_0, w_0)\,\boldsymbol{j} + z(u, v_0, w_0)\,\boldsymbol{k}, \tag{8.10}$$

where u acts as a coordinate along the resulting curve. Figure 8.1 illustrates for a space parameterized by cartesian and spherical coordinate systems. For example, in Fig. 8.1(b)

1. the surface corresponding to r = constant is a sphere parameterized by θ and ϕ,
2. the constraint θ = constant is a cone parameterized by the variables r and ϕ,
3. the constraint $\phi = \text{constant}$ defines a half-plane parameterized by r and θ, and
4. setting all three variables to constants defines a point P within the space.

Through any such point three curves pass that are determined by the pairwise intersections of the three surfaces just defined. (1) Setting r and θ to constants specifies a curve that is the intersection of the sphere and the cone, parameterized by the variable ϕ. (2) Setting r and ϕ to constants specifies an arc along the spherical surface parameterized by θ. (3) Setting θ and ϕ to constants specifies a ray along the conic surface parameterized by r.

The tangent basis: Partial differentiation of (8.8) with respect to u, v, and w, respectively, gives tangents to the three coordinate curves passing though a point P. These define a set of basis vectors $\boldsymbol{e}_i$ through

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w}, \tag{8.11}$$

where it is understood that all partial derivatives are to be evaluated at the point $P = (u_0, v_0, w_0)$. The basis generated by the tangents to the coordinate curves will be termed the *tangent basis.* Example 8.2 illustrates construction of the tangent basis for the example of a spherical coordinate system.

Example 8.2 Consider the spherical coordinate system defined in Eq. (8.7) and illustrated in Fig. 8.1(b). The position vector $\boldsymbol{r}$ is

$$\boldsymbol{r} = (r\sin\theta\cos\phi)\,\boldsymbol{i} + (r\sin\theta\sin\phi)\,\boldsymbol{j} + (r\cos\theta)\,\boldsymbol{k}$$

and the tangent basis is obtained from Eq. (8.11) as

$$\begin{aligned}
\boldsymbol{e}_1 \equiv \boldsymbol{e}_r &= \frac{\partial \boldsymbol{r}}{\partial r} = (\sin\theta\cos\phi)\,\boldsymbol{i} + (\sin\theta\sin\phi)\,\boldsymbol{j} + (\cos\theta)\,\boldsymbol{k},\\
\boldsymbol{e}_2 \equiv \boldsymbol{e}_\theta &= \frac{\partial \boldsymbol{r}}{\partial \theta} = (r\cos\theta\cos\phi)\,\boldsymbol{i} + (r\cos\theta\sin\phi)\boldsymbol{j} - (r\sin\theta)\,\boldsymbol{k},\\
\boldsymbol{e}_3 \equiv \boldsymbol{e}_\phi &= \frac{\partial \boldsymbol{r}}{\partial \phi} = -(r\sin\theta\sin\phi)\,\boldsymbol{i} + (r\sin\theta\cos\phi)\,\boldsymbol{j}.
\end{aligned}$$

These basis vectors are mutually orthogonal because $\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = \boldsymbol{e}_2 \cdot \boldsymbol{e}_3 = \boldsymbol{e}_3 \cdot \boldsymbol{e}_1 = 0$. For example,

$$\begin{aligned}
\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 &= r\sin\theta\cos\theta\cos^2\phi + r\sin\theta\cos\theta\sin^2\phi - r\cos\theta\sin\theta\\
&= r\sin\theta\cos\theta(\cos^2\phi + \sin^2\phi) - r\cos\theta\sin\theta = 0.
\end{aligned}$$

From the scalar products of the basis vectors with themselves, their lengths are

$$|\boldsymbol{e}_1| = 1 \qquad |\boldsymbol{e}_2| = r \qquad |\boldsymbol{e}_3| = r\sin\theta,$$

and these can be used to define a normalized basis,

$$\begin{aligned}
\hat{\boldsymbol{e}}_1 &\equiv \frac{\boldsymbol{e}_1}{|\boldsymbol{e}_1|} = (\sin\theta\cos\phi)\,\boldsymbol{i} + (\sin\theta\sin\phi)\boldsymbol{j} + (\cos\theta)\,\boldsymbol{k},\\
\hat{\boldsymbol{e}}_2 &\equiv \frac{\boldsymbol{e}_2}{|\boldsymbol{e}_2|} = (\cos\theta\cos\phi)\,\boldsymbol{i} + (\cos\theta\sin\phi)\boldsymbol{j} - (\sin\theta)\,\boldsymbol{k},\\
\hat{\boldsymbol{e}}_3 &\equiv \frac{\boldsymbol{e}_3}{|\boldsymbol{e}_3|} = -(\sin\phi)\,\boldsymbol{i} + (\cos\phi)\boldsymbol{j}.
\end{aligned}$$

These basis vectors are now mutually orthogonal and of unit length, with a geometry that is illustrated in Fig. 8.2.

In elementary physics it is common to use an orthogonal coordinate system so that the basis

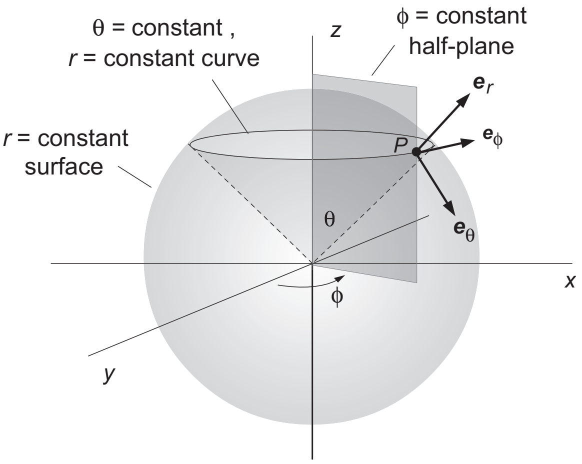


Fig. 8.2 Unit vectors in the tangent basis at point P for Example 8.2. The three basis vectors are tangents to the curves passing through P that are defined by setting any two of the three variables to a constant.

vectors are mutually orthogonal, and to normalize them to unit length, as in this example. However, for more general applications the tangent basis defined by the partial derivatives as in Eq. (8.11) need not be orthogonal or normalized to unit length.

The dual basis: We may also construct a basis at a point P by using the normals to coordinate surfaces to define the basis vectors. Solving for

$$u = u(x,y,z) \qquad v = v(x,y,z) \qquad w = w(x,y,z),$$

an alternative set of basis vectors $(\boldsymbol{e}^u, \boldsymbol{e}^v, \boldsymbol{e}^w)$ may be defined using the gradients

$$\begin{aligned} \boldsymbol{e}^u &\equiv \boldsymbol{\nabla} u = \frac{\partial u}{\partial x}\boldsymbol{i} + \frac{\partial u}{\partial y}\boldsymbol{j} + \frac{\partial u}{\partial z}\boldsymbol{k}, \\ \boldsymbol{e}^v &\equiv \boldsymbol{\nabla} v = \frac{\partial v}{\partial x}\boldsymbol{i} + \frac{\partial v}{\partial y}\boldsymbol{j} + \frac{\partial v}{\partial z}\boldsymbol{k}, \\ \boldsymbol{e}^w &\equiv \boldsymbol{\nabla} w = \frac{\partial w}{\partial x}\boldsymbol{i} + \frac{\partial w}{\partial y}\boldsymbol{j} + \frac{\partial w}{\partial z}\boldsymbol{k}, \end{aligned} \tag{8.12}$$

which are normal to the three coordinate surfaces through P defined by $u = u_0$, $v = v_0$, and $w = w_0$, respectively. This basis $(\boldsymbol{e}^u, \boldsymbol{e}^v, \boldsymbol{e}^w)$ defined in terms of normals to surfaces is said to be the *dual* of the basis (8.11), which is defined in terms of tangents to curves. Notice that the two basis sets have been distinguished by the use of *superscript indices* on the basis vectors (8.12) and *subscript indices* on the basis vectors (8.11).

Orthogonal and non-orthogonal coordinate systems: The tangent basis and dual basis are equally valid. For orthogonal coordinate systems the set of normals to the planes corresponds to the set of tangents to the curves in orientation, differing possibly only in the lengths of basis components. Thus, if the basis vectors are normalized the tangent basis and the dual basis are equivalent for orthogonal coordinates and the preceding distinctions have little practical significance. However, for non-orthogonal coordinate systems the two bases generally are not equivalent and the distinction between upper and lower indices is relevant. Example 8.3 illustrates.

Example 8.3 Define a coordinate system (u, v, w) in terms of cartesian coordinates (x, y, z) through [5]

$$x = u + v \qquad y = u - v \qquad z = 2uv + w.$$

The position vector for a point $\boldsymbol{r}$ is then

$$\boldsymbol{r} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} = (u+v)\,\boldsymbol{i} + (u-v)\,\boldsymbol{j} + (2uv+w)\,\boldsymbol{k}$$

and from Eq. (8.11) the tangent basis is

$$\boldsymbol{e}_1 \equiv \boldsymbol{e}_u = \frac{\partial \boldsymbol{r}}{\partial u} = \boldsymbol{i} + \boldsymbol{j} + 2v\boldsymbol{k} \qquad \boldsymbol{e}_2 \equiv \boldsymbol{e}_v = \frac{\partial \boldsymbol{r}}{\partial v} = \boldsymbol{i} - \boldsymbol{j} + 2u\boldsymbol{k} \qquad \boldsymbol{e}_3 \equiv \boldsymbol{e}_w = \frac{\partial \boldsymbol{r}}{\partial w} = \boldsymbol{k}.$$

Solving the original equations for (u, v, w),

$$u = \tfrac{1}{2}(x+y) \qquad v = \tfrac{1}{2}(x-y) \qquad w = z - \tfrac{1}{2}(x^2 - y^2),$$

and Eq. (8.12) gives for the dual basis

$$\boldsymbol{e}^1 \equiv \boldsymbol{e}^u = \frac{\partial u}{\partial x}\boldsymbol{i} + \frac{\partial u}{\partial y}\boldsymbol{j} + \frac{\partial u}{\partial z}\boldsymbol{k} = \tfrac{1}{2}(\boldsymbol{i} + \boldsymbol{j}) \qquad \boldsymbol{e}^2 \equiv \boldsymbol{e}^v = \frac{\partial v}{\partial x}\boldsymbol{i} + \frac{\partial v}{\partial y}\boldsymbol{j} + \frac{\partial v}{\partial z}\boldsymbol{k} = \tfrac{1}{2}(\boldsymbol{i} - \boldsymbol{j})$$

$$\boldsymbol{e}^3 \equiv \boldsymbol{e}^w = \frac{\partial w}{\partial x}\boldsymbol{i} + \frac{\partial w}{\partial y}\boldsymbol{j} + \frac{\partial w}{\partial z}\boldsymbol{k} = -(u+v)\,\boldsymbol{i} + (u-v)\boldsymbol{j} + \boldsymbol{k}.$$

For the tangent basis the preceding expressions give

$$\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = 4uv \qquad \boldsymbol{e}_2 \cdot \boldsymbol{e}_3 = 2u \qquad \boldsymbol{e}_3 \cdot \boldsymbol{e}_1 = 2v,$$

where the orthonormality of the basis $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$ has been used. Thus the tangent basis is non-orthogonal. Taking scalar products of tangent basis vectors with themselves gives

$$\boldsymbol{e}_1 \cdot \boldsymbol{e}_1 = 2 + 4v^2 \qquad \boldsymbol{e}_2 \cdot \boldsymbol{e}_2 = 2 + 4u^2 \qquad \boldsymbol{e}_3 \cdot \boldsymbol{e}_3 = 1,$$

so the tangent basis in this example is also not normalized to unit length. In this non-orthogonal case the normal basis and the dual basis are distinct.

The preceding example illustrates that Eqs. (8.11) and (8.12) define *different but equally-valid bases*, and that they are physically distinguishable in the general case of non-cartesian coordinate systems, and hence that the placement of indices in upper or lower positions matters. It should be assumed going forward that the vertical placement of indices in equations (upper or lower positions) is significant.

8.3.3 Expansion of Vectors and Dual Vectors

Any vector $\boldsymbol{V}$ may be expanded in terms of the tangent basis $\{\boldsymbol{e}_i\}$ and any dual vector $\boldsymbol{\omega}$ may be expanded in terms of the dual basis $\{\boldsymbol{e}^i\}$:

$$\boldsymbol{V} = V^1\boldsymbol{e}_1 + V^2\boldsymbol{e}_2 + V^3\boldsymbol{e}_3 = \sum_i V^i\boldsymbol{e}_i \equiv V^i\boldsymbol{e}_i, \tag{8.13}$$

$$\boldsymbol{\omega} = \omega_1\boldsymbol{e}^1 + \omega_2\boldsymbol{e}^2 + \omega_3\boldsymbol{e}^3 = \sum_i \omega_i\boldsymbol{e}^i \equiv \omega_i\boldsymbol{e}^i, \tag{8.14}$$

where in the last step of each equation the *Einstein summation convention* has been introduced:

> ***Einstein Summation Convention:*** An index appearing twice on one side of an equation, once in a lower position and once in an upper position, implies a summation on that repeated index. The index that is summed over is termed a *dummy index;* summation on a dummy index on one side of an equation implies that it does not appear on the other side. If the same index appears more than twice on the same side of an equation, or appears more than once in an upper position or more than once in a lower position, you have likely made a mistake. Since the dummy (repeated) index is summed over, it does not matter what the repeated index is, as long as it is not equivalent to another index in the equation.

From this point onward the Einstein summation convention usually will be assumed because it leads to more compact, easier to read equations.

The upper-index coefficients V^i appearing in Eq. (8.13) are the *components of the vector* in the basis $\boldsymbol{e}_i = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$, while the lower-index coefficients ω_i appearing in Eq. (8.14) are the *components of the dual vector* in the basis $\boldsymbol{e}^i = \{\boldsymbol{e}^1, \boldsymbol{e}^2, \boldsymbol{e}^3\}$. The vector and dual vector components generally are distinct because they are components in two different bases. However, as will now be discussed the vector and dual vector spaces are related fundamentally way such that vector components V^i and dual vector components ω_i may be treated operationally as if they were different components of the same vector.

8.3.4 Vector Scalar Product and the Metric Tensor

Utilizing Eq. (8.13), the scalar product of two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ may be expressed as

$$\boldsymbol{A}\cdot\boldsymbol{B} = (A^i\boldsymbol{e}_i)\cdot(B^j\boldsymbol{e}_j) = \boldsymbol{e}_i\cdot\boldsymbol{e}_j A^iB^j = g_{ij}A^iB^j, \tag{8.15}$$

where the components of the *metric tensor* g_{ij} in this basis are given by

$$g_{ij} \equiv \boldsymbol{e}_i\cdot\boldsymbol{e}_j. \tag{8.16}$$

Two vectors alone cannot form a scalar product, but the scalar product of two vectors can be computed with the aid of the metric tensor, as Eq. (8.15) illustrates. Equivalently, the scalar product of dual vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may be expressed as

$$\boldsymbol{\alpha}\cdot\boldsymbol{\beta} = \alpha_i\boldsymbol{e}^i\cdot\beta_j\boldsymbol{e}^j = g^{ij}\alpha_i\beta_j, \tag{8.17}$$

where the metric tensor components g^{ij} with two upper indices are defined by

$$g^{ij} \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}^j, \tag{8.18}$$

and the scalar product of dual vectors and vectors as

$$\boldsymbol{\alpha} \cdot \boldsymbol{B} = \alpha_i \boldsymbol{e}^i \cdot B^j \boldsymbol{e}_j = g^i_{\,j} \alpha_i B^j, \tag{8.19}$$

where the metric tensor components $g^i_{\,j}$ with mixed upper and lower indices are defined by

$$g^i_{\,j} \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}_j. \tag{8.20}$$

Properties of the metric tensor will be discussed below but first we shall use the metric tensor to establish a relationship called *duality* between the vector and dual vector spaces.

8.3.5 Duality of Vectors and Dual Vectors

There is little practical distinction between vectors and dual vectors in euclidean space with cartesian coordinates. However, in a curved space and/or with non-cartesian coordinates the situation is more complex. Although the examples in this chapter are primarily from non-curved spaces where it is possible to finesse the issue, it is important not to build into the discussion at this stage methods and terminologies that will not serve us well in the more general case . The essential mathematics will be discussed in more depth later, primarily in Section 8.4.6, but the salient points are that

1. Vectors are not defined directly in the manifold, but instead are defined in a euclidean vector space (see Box 8.3) attached to the (possibly curved) manifold at each spacetime point that is called the *tangent space.*
2. Dual vectors are not defined directly in the manifold, but instead are defined in a euclidean vector space attached to the (possibly curved) manifold at each spacetime point that is called the *cotangent space.*
3. The tangent space of vectors and the cotangent space of dual vectors at a point P of the manifold are different but *dual to each other* in a manner that will be made precise below.
4. This duality allows objects in the two different spaces to be treated as effectively the same kinds of objects.

As will be discussed further below, vectors and dual vectors are examples of more general objects called *tensors,* which permits an abstract definition in terms of mappings from vectors and dual vectors to the real numbers.[3] To be specific,

1. Dual vectors $\boldsymbol{\omega}$ are linear maps of vectors $\boldsymbol{V}$ to the real numbers: $\boldsymbol{\omega}(\boldsymbol{V}) = \omega_i V^i \in \mathbb{R}$.
2. Vectors $\boldsymbol{V}$ are linear maps of dual vectors $\boldsymbol{\omega}$ to the real numbers: $\boldsymbol{V}(\boldsymbol{\omega}) = V^i \omega_i \in \mathbb{R}$.

[3] A *mapping* generalizes a *function.* For example, $y = f(x)$ is a map that associates the real number y with the real number x. In this example the map is from a space to the same space (real numbers to real numbers). More generally the mapping can be between different spaces, such as from vectors to real numbers.

In these definitions expressions like $\boldsymbol{\omega}(\boldsymbol{V}) = \omega_i V^i \in \mathbb{R}$ mean "the dual vectors $\boldsymbol{\omega}$ act linearly on the vectors $\boldsymbol{V}$ to produce $\omega_i V^i \equiv \sum_i \omega_i V^i$, which are elements of the real numbers," or "dual vectors $\boldsymbol{\omega}$ are functions (maps) that take vectors $\boldsymbol{V}$ as arguments and yield $\omega_i V^i$, which are real numbers". Linearity of the mapping means, for example,

$$\boldsymbol{\omega}(\alpha \boldsymbol{A} + \beta \boldsymbol{B}) = \alpha \boldsymbol{\omega}(\boldsymbol{A}) + \beta \boldsymbol{\omega}(\boldsymbol{B}),$$

where $\boldsymbol{\omega}$ is a dual vector, α and β are arbitrary real numbers, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are arbitrary vectors.

It is easy to show (see Box 8.3) that, just as the space of vectors satisfies the conditions required of a vector space, the space of dual vectors as defined by the map above satisfies the same conditions and also is a vector space. The vector space of vectors and corresponding vector space of dual vectors are said to be *dual to each other* because they are related by

$$\boldsymbol{\omega}(\boldsymbol{V}) = \boldsymbol{V}(\boldsymbol{\omega}) = V^i \omega_i \in \mathbb{R}. \tag{8.21}$$

Notice further that the expression $\boldsymbol{A} \cdot \boldsymbol{B} = g_{ij} A^i B^j$ from Eq. (8.15) defines a linear map from the vectors to the real numbers, since it takes two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as arguments and returns the scalar product, which is a real number. Thus one may write

$$\boldsymbol{A}(\boldsymbol{B}) = \boldsymbol{A} \cdot \boldsymbol{B} \equiv A_i B^i = g_{ij} A^i B^j. \tag{8.22}$$

But since in $A_i B^i = g_{ij} A^i B^j$ the vector B is arbitrary, in general

$$A_i = g_{ij} A^j, \tag{8.23}$$

which specifies a correspondence between a vector with components A^i in the tangent space of vectors and a dual vector with components A_i in the cotangent space of dual vectors. Likewise, Eq. (8.23) can be inverted using that the inverse of g_{ij} is g^{ij} (see Section 8.3.6) to give

$$A^i = g^{ij} A_j. \tag{8.24}$$

Hence, using the metric tensor to raise and lower indices by summing over a repeated index (an operation called *contraction*) as in Eqs. (8.23) and (8.24), we see that the vector and dual vector components are related through contraction with the metric tensor.

> This is the precise sense in which the tangent and cotangent spaces are dual: they are different, but closely related through the metric tensor.

The duality of the vector and dual vector spaces may be incorporated concisely by requiring that for the basis vectors $\{\boldsymbol{e}_i\}$ and basis dual vectors $\{\boldsymbol{e}^i\}$ in Eqs. (8.13) and (8.14)

$$\boldsymbol{e}^i(\boldsymbol{e}_j) = \boldsymbol{e}^i \cdot \boldsymbol{e}_j = \delta^i_j, \tag{8.25}$$

where the Kronecker delta is defined by

$$\delta^i_j = \left\{ \begin{array}{ll} 1 & i = j \\ 0 & i \neq j \end{array} \right. . \tag{8.26}$$

This implies that the basis vectors can be used to project out the components of a vector $\boldsymbol{V}$ by taking the scalar product with the vector,

$$V^i = \boldsymbol{e}^i \cdot \boldsymbol{V} \qquad V_i = \boldsymbol{e}_i \cdot \boldsymbol{V}. \tag{8.27}$$

A lot of important mathematics has transpired in the last few equations, so let's pause for a moment and take stock. For a space with a metric tensor defined,

1. Eqs. (8.21)–(8.27) imply that vectors and dual vectors are in a one-to-one relationship
2. that permits them to be manipulated effectively as if a dual vector component were just a vector component with a lower index, and
3. component indices can be raised or lowered as desired by contraction with the metric tensor.

All spaces of interest here will have metrics, so this reduces the practical implications of the distinction between vectors and dual vectors to a simple matter of keeping proper track of upper and lower positions for indices.

8.3.6 Properties of the Metric Tensor

The metric tensor will play a fundamental role in our discussion. Accordingly, let us summarize some of its properties in simple euclidean spaces, since most will carry over (suitably generalized) to 4D (possibly curved) spacetime. The metric tensor must be symmetric in its indices:

$$g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}. \tag{8.28}$$

From Eqs. (8.23) and (8.24)

$$g_{ij}A^j = A_i \qquad g^{ij}A_j = A^i. \tag{8.29}$$

That is, *contraction with the metric tensor may be used to raise or lower an index.* The scalar product of vectors may be written in any of the equivalent ways

$$\boldsymbol{A} \cdot \boldsymbol{B} = g_{ij}A^iB^j = g^{ij}A_iB_j = g^i_jA_iB^j = A^iB_i = A_iB^i. \tag{8.30}$$

From the two expressions in (8.29), $A^i = g^{ij}A_j = g^{ij}g_{jk}A^k$, and since this is valid for arbitrary components A^i it follows that the metric tensor obeys

$$g^{ij}g_{jk} = g_{kj}g^{ji} = \delta^i_k. \tag{8.31}$$

Viewing g^{ij} as the elements of a matrix G and g_{ij} as the elements of a matrix $\tilde{G}$, Eqs. (8.28) are equivalent to the matrix equations

$$G = G^{\mathrm{T}} \qquad \tilde{G} = \tilde{G}^{\mathrm{T}}, \tag{8.32}$$

where T denotes the transpose of the matrix. The Kronecker delta is just the unit matrix I, implying that Eq. (8.31) may be written as the matrix equations

$$\tilde{G}G = G\tilde{G} = I. \tag{8.33}$$

Therefore, we note the useful property that

Box 8.1 The metric tensor for 3-dimensional euclidean space

Components: $g_{ij} \equiv \boldsymbol{e}_i \cdot \boldsymbol{e}_j \qquad g^{ij} \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}^j \qquad g^i_j \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}_j = \delta^i_j$

Scalar product: $\boldsymbol{A} \cdot \boldsymbol{B} = g_{ij}A^iB^j = g^{ij}A_iB_j = g^i_jA_iB^j = A^iB_i = A_iB^i$

Symmetry: $g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}$

Contractions: $g_{ij}A^j = A_i \qquad g^{ij}A_j = A^i$

Orthogonality: $g^{ij}g_{jk} = g_{kj}g^{ji} = \delta^i_k$

Matrix properties : $\tilde{G}G = G\tilde{G} = I \qquad G \equiv [g^{ij}] \qquad \tilde{G} \equiv [g_{ij}]$

> The matrix corresponding to the metric tensor with two lower indices is the *inverse* of the matrix corresponding to the metric tensor with two upper indices, and one may be obtained from the other by matrix inversion.

Some fundamental properties of the metric tensor for three-dimensional euclidean space are summarized in Box 8.1.

8.3.7 Line Elements

For coordinates $u^1(t)$, $u^2(t)$, and $u^3(t)$ that are parameterized by the variable t, as t varies the points corresponding to specific values of the coordinates

$$u^1 = u^1(t) \qquad u^2 = u^2(t) \qquad u^3 = u^3(t)$$

will trace out a curve in the 3D euclidean space. From Eq. (8.8), the position vector for points on the curve as a function of t is

$$\boldsymbol{r}(t) = x\big(u^1(t), u^2(t), u^3(t)\big)\,\boldsymbol{i} + y\big(u^1(t), u^2(t), u^3(t)\big)\,\boldsymbol{j} + z\big(u^1(t), u^2(t), u^3(t)\big)\,\boldsymbol{k},$$

and by the chain rule

$$\frac{d\boldsymbol{r}}{dt} = \frac{\partial \boldsymbol{r}}{\partial u^1}\frac{du^1}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^2}\frac{du^2}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^3}\frac{du^3}{dt} = \dot{u}^1\boldsymbol{e}_1 + \dot{u}^2\boldsymbol{e}_2 + \dot{u}^3\boldsymbol{e}_3, \tag{8.34}$$

where (8.11) was used and $du^i/dt \equiv \dot{u}^i$. The squared infinitesimal distance along the curve is then given by

$$\begin{aligned} ds^2 = d\boldsymbol{r} \cdot d\boldsymbol{r} &= du^i\boldsymbol{e}_i \cdot du^j\boldsymbol{e}_j \\ &= \boldsymbol{e}_i \cdot \boldsymbol{e}_j\, du^i du^j \\ &= g_{ij}\, du^i du^j, \end{aligned} \tag{8.35}$$

where Eq. (8.16) was used. [Notice the notational convention $d\alpha^2 \equiv (d\alpha)^2$.] Thus $ds^2 = g_{ij}du^idu^j$ is the infinitesimal line element implied by the metric g_{ij}, and the length d of a

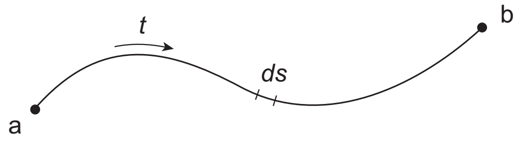


Fig. 8.3 Distance ds between points a and b along a curve parameterized by t.

finite segment of the curve between points a and b is obtained integration,

$$d = \int_a^b \left(g_{ij} \frac{du^i}{dt} \frac{du^j}{dt} \right)^{1/2} dt, \tag{8.36}$$

where t parameterizes the position along the curve, as illustrated in Fig. 8.3.

8.3.8 Euclidean Line Element

The line element for 2D euclidean space in cartesian coordinates (x, y) is

$$ds^2 = dx^2 + dy^2, \tag{8.37}$$

which is just the Pythagorean theorem in differential form. The corresponding line element in plane polar coordinates (r, ϕ) is then

$$ds^2 = dr^2 + r^2 d\phi^2, \tag{8.38}$$

as worked out in Example 8.4.

Example 8.4 For plane polar coordinates (r, ϕ)

$$x = r\cos\phi \qquad y = r\sin\phi,$$

so the position vector (8.8) is

$$\boldsymbol{r} = (r\cos\phi)\,\boldsymbol{i} + (r\sin\phi)\,\boldsymbol{j}.$$

Then from Eq. (8.11) the basis vectors of the tangent basis are

$$\boldsymbol{e}_1 = \frac{\partial \boldsymbol{r}}{\partial r} = (\cos\phi)\boldsymbol{i} + (\sin\phi)\,\boldsymbol{j} \qquad \boldsymbol{e}_2 = \frac{\partial \boldsymbol{r}}{\partial \phi} = -r(\sin\phi)\,\boldsymbol{i} + r(\cos\phi)\,\boldsymbol{j}.$$

The elements of the metric tensor are then given by Eq. (8.16),

$$g_{11} = \cos^2\phi + \sin^2\phi = 1 \qquad g_{22} = r^2(\cos^2\phi + \sin^2\phi) = r^2$$

and $g_{12} = g_{21} = 0$, or in matrix form,

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}.$$

Then the line element is given by Eq. (8.35),

$$ds^2 = g_{11}(du^1)^2 + g_{22}(du^2)^2 = dr^2 + r^2 d\phi^2,$$

where $u^1 = r$ and $u^2 = \phi$. Equivalently, the line element can be expressed as

$$ds^2 = (dr \;\; d\phi) \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \begin{pmatrix} dr \\ d\phi \end{pmatrix} = dr^2 + r^2 d\phi^2,$$

in matrix form.

The form of the line element differs between cartesian and polar coordinates, but for any two nearby points the distance between them is given by ds, for either coordinate system. Thus, the line element ds is invariant under coordinate transformations. The distance between two points that are not nearby can be obtained by integrating ds along the curve, so *the distance interval is invariant under coordinate transformations for metric spaces.*[4] The line element, which is specified in terms of the metric tensor, characterizes the geometry of the space because integrals of the line element define distances and angles can be defined in terms of ratios of distances. Indeed, all the axioms of euclidean geometry could be verified starting from the line elements (8.37) or (8.38).

8.3.9 Integration and Differentiation

It is important to know how the volume element for integrals behaves under change of coordinates. This is trivial in euclidean space with orthonormal coordinates, but becomes non-trivial in curved spaces, or in flat spaces parameterized in non-cartesian coordinates. We may illustrate in flat 2D space with coordinates (x^1, x^2) and basis vectors $(\boldsymbol{e}_1, \boldsymbol{e}_2)$, assuming an angle θ between the basis vectors. The 2D volume (area) element in this case is

$$dA = \sqrt{\det g}\, dx^1 dx^2, \tag{8.39}$$

where $\det g$ is the determinant of the metric tensor matrix g_{ij}. For orthonormal coordinates g_{ij} is the unit matrix and $(\det g)^{1/2} = 1$, but in the general case the $(\det g)^{1/2}$ factor is not unity and its presence is essential to making integration invariant under change of coordinates.

Derivatives of vectors in spaces defined by position-dependent metrics are crucial in the formulation of general relativity. Let us introduce the issue with the simpler case of the derivative of a vector in a flat euclidean space, but parameterized with a vector basis that depends on the coordinates. A vector $\boldsymbol{V}$ may be expanded in a basis $\boldsymbol{e}_i$,

$$\boldsymbol{V} = V^i \boldsymbol{e}_i. \tag{8.40}$$

Applying the usual (Leibniz) rule for the derivative of a product, the partial derivative is

$$\frac{\partial \boldsymbol{V}}{\partial x^j} = \frac{\partial V^i}{\partial x^j} \boldsymbol{e}_i + V^i \frac{\partial \boldsymbol{e}_i}{\partial x^j}, \tag{8.41}$$

[4] Mathematically, spaces are equipped with a hierarchy of characteristics and a distance-measuring prescription (a metric) need not be one of them. If such a prescription is defined, the space is termed a *metric space*. This distinction may seem pedantic to a physicist since almost all spaces employed in physics are metric spaces, but it is important from a fundamental mathematical perspective.

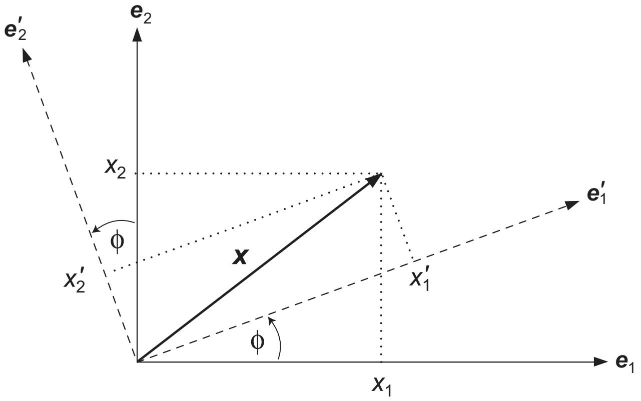


Fig. 8.4 Rotation of the coordinate system for a vector $\boldsymbol{x}$. The vector is invariant under the rotation but its components in the original and rotated coordinate systems are different.

where the first term represents the *change in the component* V^i and the second term represents the *change in the basis vectors* $\boldsymbol{e}_i$. In the second term the factor $\partial \boldsymbol{e}_i / \partial x^j$ is itself a vector and can be expanded in the vector basis (8.40),

$$\frac{\partial \boldsymbol{e}_i}{\partial x^j} = \Gamma^k_{ij} \boldsymbol{e}_k. \tag{8.42}$$

The Γ^k_{ij} appearing in Eq. (8.42) are called *connection coefficients*. Later we will see that the connection coefficients can be evaluated from the metric tensor and and that they may be used in either curved or flat spacetime to define derivatives and to specify a prescription for parallel transport of vectors.

8.3.10 Transformations

Often it is essential to be able to express physical quantities in more than one coordinate system, so we need to understand how to transform between coordinate systems. This issue is particularly important in both general and special relativity, where it is essential to ensure that the laws of physics are not altered by transformation between coordinate systems. We may illustate the principles involved by consider two familiar examples: spatial rotations and Galilean boosts.

Rotational Transformations

Consider the description of a vector under rotation of a coordinate system about the z axis by an angle ϕ, as in Fig. 8.4. The vector $\boldsymbol{x}$ has the components x_1 and x_2 with respect to basis vectors $\{\boldsymbol{e}_i\}$ for the original coordinate system before rotation. After rotation of the coordinate system to give the new basis vectors $\{\boldsymbol{e}'_i\}$, the vector $\boldsymbol{x}$ has the components x'_1 and x'_2 in the new coordinate system. The vector $\boldsymbol{x}$ can be expanded in terms of the components for either basis:

$$\boldsymbol{x} = x^i \boldsymbol{e}_i = x'^i \boldsymbol{e}'_i, \tag{8.43}$$

and from the geometry of Fig. 8.4 the components in the two bases are related by (assuming a clockwise rotation)

$$\begin{pmatrix} x'^1 \\ x'^2 \\ x'^3 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix}, \tag{8.44}$$

which may be written compactly as

$$x'^i = R^i_j x^j, \tag{8.45}$$

where the R^i_j are the elements of the matrix in Eq. (8.44). This transformation law holds for any vector; indeed, a vector in the x–y plane may be *defined* by this transformation law.

Galilean Transformations

Another simple example is Galilean transformations between inertial frames in classical mechanics. Transformations with the same orientation but different relative velocities are called *boosts*. In Newtonian physics time t is considered to be the same for all observers and Galilean boosts for motion along the x axis take the form

$$\boldsymbol{x}' = \boldsymbol{x}'(\boldsymbol{x},t) = \boldsymbol{x} - \boldsymbol{v}t \qquad t' = t'(\boldsymbol{x},t) = t. \tag{8.46}$$

"Newtonian relativity" asserts that the laws of physics are invariant under Galilean transformations. The laws of mechanics at low velocity are approximately invariant under (8.46), but the laws of electromagnetism (Maxwell's equations) and the laws of mechanics for velocities near that of light are not. Indeed, the failure of Galilean invariance for the Maxwell equations was a large motivation for Einstein's belief that the existing laws of mechanics required modification, leading to the special theory of relativity.

In the absence of gravity, the laws of both special-relativistic mechanics and of electromagnetism are invariant under Lorentz transformations but not under Galilean transformations. In the presence of a gravity, neither Galilean nor Lorentz invariance holds globally for either electromagnetism or mechanics, and it is necessary to seek a more comprehensive invariance to describe systems subject to gravitational forces. This quest eventually leads to the theory of general relativity, but that is beyond the scope of these lectures.

Having introduced most of the important concepts in a "toy model" of euclidean space, let us now turn our attention to application of these ideas to the actual arena of special and general relativity, 4D spacetime.

8.4 Spacetime Tensors and Covariance

The *principle of relativity* implies that coordinates should be viewed as *arbitrary labels,* so that the laws of physics are independent of the coordinate system in which they are formulated. We may implement this coordinate independence by formulating the laws of physics in terms of equations that are *covariant* with respect to the coordinate transformations.[5] The term *covariance* implies that a set of equations maintains the same mathematical form under a specified set of transformations. Covariance can be stated most concisely in terms of *tensors,* which we may think of as generalizing the idea of vectors. We now give an introduction to tensors, tensor notation, and tensor properties for 4D spacetime, and illustrate with an application of the tensor formalism to Lorentz covariance, which will lead to the theory of special relativity.

8.4.1 Spacetime Coordinates

Relativity implies that space and time enter physical descriptions on comparable footings, so it is useful to unify them into a 4-dimensional continuum termed *spacetime*. Spacetime is an example of a *differentiable manifold,* as described in Box 8.2 (adapted from Ref. [11]). Spacetime points are defined by coordinates having four components, the first labeling the time t multiplied by the speed of light c, the other three labeling the spatial coordinates:

$$x \equiv x^\mu = (x^0, x^1, x^2, x^3) = (ct, \boldsymbol{x}), \tag{8.47}$$

where $\boldsymbol{x}$ denotes a 3-vector with components (x^1, x^2, x^3) labeling the spatial position. The first component x^0 is termed *timelike* and the last three components (x^1, x^2, x^3) are termed *spacelike.* As for earlier discussion, the upper or lower placement of indices is meaningful. Bold symbols will denote (ordinary) vectors defined in the three spatial degrees of freedom, and 4-component vectors in spacetime will be denoted in non-bold symbols. The modern convention is to number the indices beginning with zero rather than one. The coordinate systems of interest will be subject only to the requirements that they assign a coordinate uniquely and be differentiable to sufficient order at each point of spacetime.

A point in an n-dimensional manifold can be labeled using a coordinate system of n parameters, but the choice of coordinate system is arbitrary. Points may be relabeled by a *passive coordinate transformation* that switches the coordinate labels of the points, $x^\mu \rightarrow x'^\mu$, with x^μ and x'^μ labeling the *same point* but in two different coordinate systems. Our generic concern is with a transformation between one set of spacetime coordinates denoted by (x^0, x^1, x^2, x^3), and a new set

$$x'^\mu = x'^\mu(x) \qquad (\mu = 0, 1, 2, 3), \tag{8.48}$$

where $x = x^\mu$ denotes the original (untransformed) coordinates. This notation is an eco-

[5] Covariance is defined relative to a particular set of transformations. There is a subtle difference in meaning between *invariance* and *covariance*, with respect to a set of transformations. Invariance means that the physical observables of the system are not changed by the transformations. Covariance means that the system is formulated mathematically so that the form of the equations does not change under the transformations. *Manifest covariance* means that the invariance is *manifest* in the formulation of the theory (can be "seen at a glance"). Thus, a system could be invariant under some set of transformations, but not manifestly covariant because the invariance is obscured by the way the equations are written. This is the case with the Maxwell equations. As we shall see in Section 9.8, the Maxwell equations written in their usual form (1.1) do not look to be invariant under Lorentz transformation but they are, and they can be transformed so that the Lorentz invariance is manifest.

Box 8.2 Manifolds

Loosely, a manifold is a space. More formally, an *n-dimensional manifold* is a set that can be parameterized continuously by n independent real coordinates for each point (member of the set). Physics is usually concerned with *differentiable manifolds,* which are continuous and differentiable (to a suitable order). Again, more formally: a manifold is continuous if for every point there are neighboring points having infinitesimally different coordinates, and differentiable if a scalar field can be defined on the manifold that is everywhere differentiable. Special and general relativity assume spacetime to be a *Riemannian manifold:* a continuous, differentiable manifold with geometry described by a metric tensor of quadratic form that may depend on spacetime coordinates.

Coordinates, charts, and atlases

A *coordinate system* or *chart* associates n real parameter values (labels) uniquely with each point of an n-dimensional manifold M through a one-to-one mapping from $\mathbb{R}^n$ (cartesian product of n copies of the real numbers $\mathbb{R}$) to M. For example, the set of continuous rotations about a single axis defines a one-dimensional manifold parameterized by an angle $\phi \in \mathbb{R}$. The one-to-one association of points in the n-dimensional manifold with the values of their parameter labels is analogous to mapping points of the manifold to points of an n-dimensional euclidean space. Thus, *locally* a manifold looks like euclidean space in its most general properties, such as dimensionality and differentiability (but not necessarily in geometry, since this depends on whether the manifold has a metric and its nature).

A single coordinate system is usually insufficient to give a unique correspondence between points and coordinate labels for all but the simplest manifolds. For example, the latitude–longitude system for the Earth (viewed as a 2-sphere, S^2) is degenerate at the poles where all values of longitude correspond to a single point. In such cases the manifold must be parameterized by overlapping *coordinate patches* (*charts*), with *transition functions* between the different sets of coordinates for points in each overlap region. An *atlas* is a collection of charts sufficient to parameterize an entire manifold. For the latitude–longitude example it may be shown that the atlas must contain at least two overlapping charts to parameterize the full manifold uniquely.

Curves and surfaces

Subsets of points within a manifold can be used to define *curves* and *surfaces*, which represent *submanifolds* of the full manifold. Often it is convenient to represent these parametrically, with an m-dimensional submanifold parameterized by m parameters. A curve is a one-dimensional submanifold parameterized by a single parameter, while a *hypersurface* is a surface of one less dimension than the full manifold, parameterized by $n-1$ real numbers.

nomical form of

$$x'^{\mu} = f^{\mu}(x^0, x^1, x^2, x^3) \qquad (\mu = 0, 1, \ldots), \tag{8.49}$$

where the single-valued, continuously-differentiable function f^{μ} assigns new (primed) coordinates (x'^0, x'^1, x'^2, x'^3) to a point of the manifold with old coordinates (x^0, x^1, x^2, x^3), which may be abbreviated to $x'^{\mu} = f^{\mu}(x)$ and, even more tersely, to Eq. (8.48). The coordinates in Eq. (8.48) are just labels so the laws of physics cannot depend on them. Hence the system x'^{μ} is not privileged and Eq. (8.48) should be invertible.

The generalized form of the Einstein summation convention that we introduced earlier will be assumed for all subsequent equations: for any term on one side of an equation any index that is repeated, once as a superscript and once as a subscript, implies a summation over that index. A superscript (subscript) in a denominator counts as a subscript (superscript) in a numerator. We will use Greek indices $(\alpha, \beta, \ldots)$ to denote the full set of spacetime indices running over 0, 1, 2, 3, while roman indices $(i, j, \ldots)$ will denote the indices 1, 2, 3 running only over the spatial coordinates. Thus x^{μ} means any of the components x^0, x^1, x^2, x^3, but x^i means any of the components x^1, x^2, x^3.

8.4.2 Vectors in Non-Euclidean Space

Spacetime is characterized by a non-euclidean manifold. In euclidean space we are used to representing vectors as directed line segments of finite length. This picture will not do in curved spacetime, which is locally but not globally, euclidean so extended straight lines have no clear meaning. Thus we need a more general way to define vectors that works in both euclidean and (possibly curved) non-euclidean manifolds. The standard solution is to define vectors for an n-dimensional manifold, not in manifold itself, but in n-dimensional euclidean *tangent spaces*, with an independent tangent space T_P attached to each point P of the manifold. This is illustrated in Fig. 8.5 for 1D and 2D spheres.[6] Just as vectors may be defined in tangent spaces attached to each point of a manifold, dual vectors my be defined in *cotangent spaces* T^*_P attached to each point of a manifold.

A *tangent bundle* TM for a manifold M is a manifold constructed from the disjoint union of all the tangent spaces T_P defined on the manifold M. Likewise, a contangent bundle for a manifold M is the disjoint union of all the cotangent spaces defined on the manifold. Such bundles and their generalization form the basis of the theory of *fiber bundles*. We will sometimes use the bundle terminology but will not use the theory of fiber bundles directly in our discussion.

8.4.3 Coordinates in Spacetime

A universal coordinate system can be chosen in euclidean space, with basis vectors that are mutually orthogonal and constant, and these constant basis vectors can be normalized to unit length for convenience. Much of ordinary physics may be described using such an

[6] The idea conveyed by Fig. 8.5 in which planes tangent to a 2D surface are shown embedded in a 3D space is useful conceptually, but defining the tangent space at each point is an *intrinsic process* with respect to a manifold and does not require embedding it in a higher-dimensional manifold, as will be shown below.

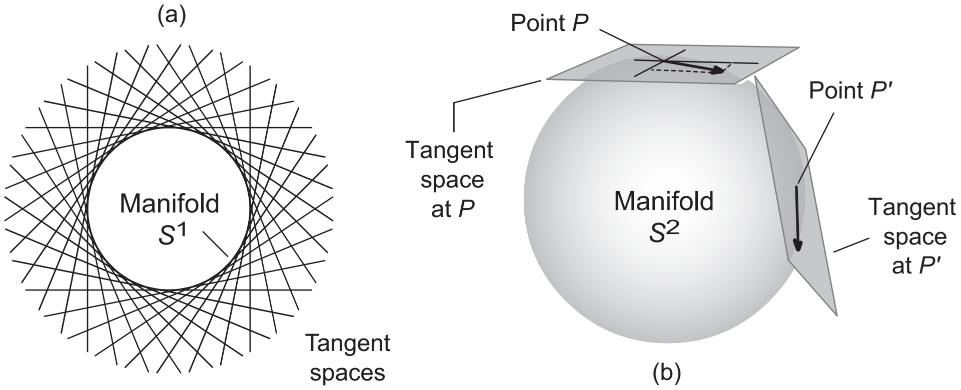


Fig. 8.5 Tangent spaces in curved manifolds, illustrated (a) for the manifold S^1 and (b) for the manifold S^2. As illustrated for S^2, vectors (indicated by arrows) are defined in the tangent spaces at each point, not in the curved manifold. Embedding the 1D tangent-space manifold in 2D euclidean space and the 2D tangent space in 3D euclidean space is for visualization purposes only; the tangent space has a specification that is intrinsic to the manifold.

orthonormal basis. The situation is more complicated in non-euclidean (possibly curved) manifolds. Because of the position-dependent metric of curved spacetime, it is most convenient to choose basis vectors that depend on position and that need not be orthogonal, and it usually is not useful to normalize them since they are position dependent.

The key to specifying vectors in curved space (which will also work in flat space) is to separate the "directed" part from the "line segment" part of the usual conception of a vector as a directed line segment, because the direction for vectors of infinitesimal length can be defined consistently in curved or flat spaces using *directional derivatives.*

8.4.4 Coordinate and Non-Coordinate Bases

Consider a curve in a differentiable manifold along which one coordinate x^μ varies while all others $x^\nu(\nu \neq \mu)$ are held constant. This curve will be termed *the coordinate curve* x^μ. Four such coordinate curves will pass through any point P in a spacetime manifold, corresponding to the coordinate curves x^μ with $\mu = (0,1,2,3)$. A set of position-dependent basis vectors $e_\mu(\mu = 0,1,2,3)$ can be defined at an arbitrary point P in the manifold by

$$e_\mu = \lim_{\delta x^\mu \to 0} \frac{\delta s}{\delta x^\mu}, \tag{8.50}$$

where δs is the infinitesimal distance along the coordinate curve x^μ between the point P with coordinate x^μ and a nearby point Q with coordinate $x^\mu + \delta x^\mu$. For a parameterized

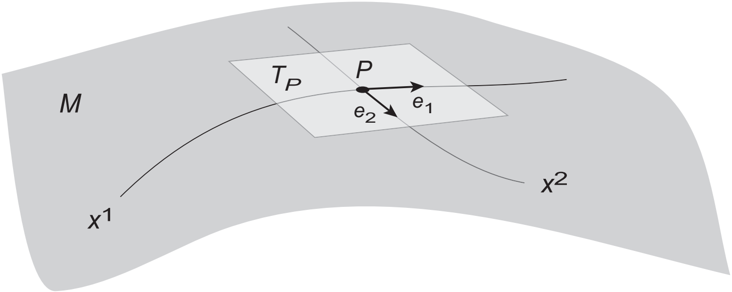


Fig. 8.6 Tangent space T_P at a point P for a curved 2D manifold M. The vectors tangent to the coordinate curves at each point define a coordinate or holonomic basis. This figure is a generalization of Fig. 8.5 to an arbitrary curved 2D manifold with a position-dependent, non-orthogonal (coordinate) basis. This embedding of M in 3D euclidean space is for visualization purposes only; the basis vectors e_1 and e_2 of the tangent space are specified by directional derivatives of the coordinate curves evaluated entirely in M at the point P, as described in Eqs. (8.50)–(8.53).

curve $x^\mu(\lambda)$ having a tangent vector t with components $t^\mu = dx^\mu/d\lambda$,

$$t = t^\mu e_\mu = \frac{dx^\mu}{d\lambda} e_\mu,$$

the directional derivative of an arbitrary scalar function $f(x^\mu)$ defined in the neighborhood of the curve is

$$\frac{df}{d\lambda} \equiv \operatorname*{Lim}_{\varepsilon\to 0} \left[\frac{f(x^\mu(\lambda+\varepsilon)) - f(x^\mu(\lambda))}{\varepsilon} \right] = \frac{dx^\mu}{d\lambda}\frac{\partial f}{\partial x^\mu} = t^\mu \frac{\partial f}{\partial x^\mu},$$

and since $f(x)$ is arbitrary this implies the operator relation

$$\frac{d}{d\lambda} = \frac{dx^\mu}{d\lambda}\frac{\partial}{\partial x^\mu} = t^\mu \frac{\partial}{\partial x^\mu}. \tag{8.51}$$

Hence the components t^μ are associated with a unique directional derivative and the partial derivative operators $\partial/\partial x^\mu$ may be identified with the basis vectors e_μ,

$$e_\mu = \frac{\partial}{\partial x^\mu} \equiv \partial_\mu, \tag{8.52}$$

which permits an arbitrary vector to be expanded as

$$V = V^\mu e_\mu = V^\mu \frac{\partial}{\partial x^\mu} = V^\mu \partial_\mu. \tag{8.53}$$

Position-dependent basis vectors specified in this way define a *coordinate basis* or *holonomic basis;* a basis using orthonormal coordinates is then termed a *non-coordinate basis* or an *anholonomic basis.* A coordinate basis is illustrated schematically in Fig. 8.6 for a generic curved 2D manifold.

The definition of a vector in terms of directional derivatives evaluated at a point of the manifold is valid in any curved or flat differentiable manifold. It replaces the standard idea of a vector as the analog of a displacement vector between two points, which does not generalize to curved manifolds.

From Eq. (8.50) the separation between nearby points is $ds = e_\mu(x)dx^\mu$, from which

$$ds^2 = ds \cdot ds = (e_\mu \cdot e_\nu)dx^\mu dx^\nu = g_{\mu\nu}dx^\mu dx^\nu$$

with the metric tensor $g_{\mu\nu}$ defined by,

$$e_\mu(x) \cdot e_\nu(x) \equiv g_{\mu\nu}(x). \tag{8.54}$$

The scalar product of vectors A and B in a coordinate basis is given by

$$A \cdot B = (A^\mu e_\mu) \cdot (B^\nu e_\nu) = g_{\mu\nu}A^\mu B^\nu. \tag{8.55}$$

Equation (8.54) may be taken as a definition of a vector coordinate basis $\{e_\mu\}$.

The preceding discussion has been specifically for vectors and involves defining a basis for the tangent space T_P at each point P using the tangents $\partial/\partial x^\mu$ to coordinate curves passing through P. A similar intrinsic procedure can be invoked to construct a basis for dual vectors in the cotangent space T_P^* at a point P using gradients to define basis vectors. This leads to equations analogous to (8.54)–(8.55), but with the indices of the basis vectors in the upper position. A set of dual basis vectors e^μ may be used to expand dual vectors ω as[7]

$$\omega = \omega_\mu e^\mu, \tag{8.56}$$

allowing the metric tensor with upper indices to be defined through

$$e^\mu(x) \cdot e^\nu(x) \equiv g^{\mu\nu}(x), \tag{8.57}$$

with the scalar product of arbitrary dual vectors α and β given by

$$\alpha \cdot \beta = g^{\mu\nu}\alpha_\mu \beta_\nu. \tag{8.58}$$

Equation (8.57) may be taken as a definition of a dual-vector coordinate basis $\{e^\mu\}$.

Just as Eqs. (8.54) or (8.57) are characteristic of a coordinate basis, an orthonormalized non-coordinate basis is specified by the requirement

$$e_{\hat\mu}(x) \cdot e_{\hat\nu}(x) = \eta_{\hat\mu\hat\nu}, \tag{8.59}$$

where $\eta = \text{diag}\{-1, 1, 1, 1\}$. In this expression, hats on indices indicate explicitly that this is an orthonormal and not coordinate basis. Our discussion will seldom require display of explicit basis vectors but usually we will assume implicitly the use of a coordinate basis such that Eqs. (8.50)–(8.58) are valid.

[7] The same symbol e will be used for vector and dual vector basis vectors, with the index in the lower position for a vector basis and upper position for a dual vector basis. The justification for this notation is given in Section 8.3.5.

8.4.5 Tensors and Coordinate Transformations

In formulating special and general relativity we are interested in how quantities that enter physical descriptions change when the spacetime coordinates are transformed as in Eq. (8.48). This requires understanding the transformations of fields, their derivatives, and their integrals. To this end, it is useful to introduce spacetime *tensors.* These have a fundamental definition without reference to specific coordinate systems. but it often proves convenient to view tensors as objects expressed in a basis with components that carry some number of upper and lower indices, and that change in a specific way under coordinate transformations. This more practical interpretation of tensors will be developed below.[8]

The *rank* of a tensor will be given a more fundamental definition below but practically it is equal to the total number of indices required to label its components when evaluated in a basis. Thus scalars are tensors of rank zero and vectors or dual vectors are tensors of rank one. This may be generalized to tensors carrying more than one index. Tensors carrying only lower indices are termed *covariant tensors,* tensors carrying only upper indices are termed *contravariant tensors,* and tensors carrying both lower and upper indices are termed *mixed tensors.*

> ***Tensor Types:*** It is convenient to indicate the *type* of a tensor by the ordered pair (p,q), where p is the number of contravariant (upper) indices for components, q is the number of covariant (lower) indices for components, and the rank of the tensor is $p+q$. In principle p and q can take any non-negative value, but practically most physical applications of tensors involve ranks of four or less.

Thus a dual vector is a tensor of type $(0,1)$ with a rank of one, while the Kronecker delta δ_μ^ν is a mixed tensor of type $(1,1)$ and rank 2.

8.4.6 Tensors as Linear Maps to Real Numbers

The characterization of tensors in terms of their transformation properties in a particular representtion that will be discussed in Section 8.5 below is the most pragmatic approach to the mathematics required to solve realistic problems. However, mathematicians prefer to define tensors in a more abstract manner that makes manifest that they are independent of representation in a particular coordinate system. This approach, which frequently goes by the name *index-free formalism*, is described in this section.

> Fundamentally, a tensor of type (n,m) has input slots for n vectors and m dual vectors, and acts linearly on these inputs to produce a real number.

[8] Hermann Minkowski introduced the use of tensors for the theory of special relativity. In his original special relativity paper, Einstein had not used tensors and at first he dismissed Minkowski's tensor formulation of special relativity as needlessly pedantic. However, Einstein soon realized the power of this new (for physicists, not for mathematicians) approach and adopted the framework of tensors and differential geometry in his later formulation of general relativity.

For example, if ω is a $(0,1)$ tensor (that is, a dual vector) and A and B are $(1,0)$ tensors (that is, vectors), linearity implies

$$\omega(aA+bB) = a\omega(A)+b\omega(B) \in \mathbb{R}, \tag{8.60}$$

where a and b are arbitrary scalars and $\mathbb{R}$ denotes the set of real numbers. This definition makes no reference to components of the vectors or dual vectors, so the tensor map must give the same real number, irrespective of any choice of coordinate system.

Thus, a tensor may be viewed as a *function of the vectors and dual vectors themselves,* rather than as a function of their components, or as an *operator* that accepts vectors and dual vectors as input and outputs a real number.

Example 8.5 As a warmup exercise, consider a real-valued function of the coordinates $f(x)$. Since this function takes no vectors or dual vectors as input and yields a real number (the value of the function at x) as output, it is a tensor of rank zero (a scalar).

Let's now give a few less-trivial examples of how this approach works, beginning with vectors and dual vectors.

Vectors and Dual Vectors

Vector spaces are discussed in Box 8.3. Suppose a vector field to be defined on a manifold such that each point P has associated with it a vector V that may be expanded in a vector basis e_μ,

$$V = V^\mu e_\mu, \tag{8.61}$$

and that there is a corresponding dual vector field ω defined at each point P that may be expanded in a dual-vector basis e^μ,

$$\omega = \omega_\mu e^\mu, \tag{8.62}$$

where the basis vectors e_μ are defined in the tangent space T_P and the basis dual vectors e^μ are defined in the cotangent space T_P^* at each point P of the manifold, as described in Section 8.4.2. Hence the e_μ are basis vectors in the tangent bundle and the e^μ are basis vectors in the cotangent bundle. As discussed in Section 8.3.5, the vector spaces for V and ω are said to be *dual* in the following sense.

Duality of Vectors and Dual Vectors: For a manifold, the space of vectors (tangent bundle) consists of all linear maps of dual vectors to the real numbers; conversely, the space of dual vectors (cotangent bundle) consists of all linear maps of vectors to the real numbers.

Box 8.3 Vector spaces

A *vector space* has a precise axiomatic definition in mathematics, but for our purposes it will be sufficient to view it more loosely as a set of objects (the vectors) that can be multiplied by real numbers and added together in a linear way while exhibiting *closure:* any such operations on elements of the set give back a linear combination of elements. For arbitrary vectors A and B, and arbitrary scalars a and b, one expects then that expressions like

$$(a+b)(A+B) = aA + aB + bA + bB$$

should be satisfied. A few other things are necessary: a zero vector that serves as an identity under vector addition, an inverse for every vector, and the usual assortment of associativity, distributivity, and commutativity rules, for example; but it is clear that it isn't very hard to be a vector space. Vector spaces of use in physics often have additional structure like a norm and inner product, but that is over and above the minimal requirements for being a bonafide vector space.

A *basis* for a vector space is a set of vectors that *span the space* (any vector is a linear combination of basis vectors) and that are *linearly independent* (no basis vector is a linear combination of other basis vectors). The number of basis vectors is the *dimension* of the space. For spacetime the vector spaces of interest will be defined at each point of the manifold and will be of dimension four.

This *duality* of vector and dual vector spaces can be implemented systematically by requiring the basis vectors to satisfy

$$e^\mu(e_\nu) = e^\mu \cdot e_\nu = \delta^\mu_\nu, \tag{8.63}$$

where the Kronecker delta is given by

$$\delta^\mu_\nu = \begin{cases} 1 & \mu = \nu \\ 0 & \mu \neq \nu \end{cases} .$$

Note that $A(B)$, which indicates the action of A on B, can be expressed in the alternative form $\langle A, B\rangle$, so Eq. (8.63) is also commonly written as $\langle e^\mu, e_\nu\rangle = \delta^\mu_\nu$. From Eqs. (8.60)–(8.63) it follows that a dual vector ω acts on a vector V in the manner

$$\begin{aligned} \omega(V) = \langle \omega, V\rangle &= \omega_\mu e^\mu(V^\nu e_\nu) \\ &= \omega_\mu V^\nu e^\mu(e_\nu) \\ &= \omega_\mu V^\nu \delta^\mu_\nu \\ &= \omega_\mu V^\mu \in \mathbb{R}, \end{aligned} \tag{8.64}$$

where $\mathbb{R}$ denotes the real numbers.[9] This illustrates clearly that a dual vector is an operator

[9] As has been noted previously, basis vectors are defined in the tangent bundle and basis dual vectors are defined in the cotangent bundle of the manifold. Thus for applications in spacetime our concern is really with the action

that accepts a vector as an argument and produces a real number (the scalar product $\omega_\mu V^\mu$, which is unique and independent of basis) as output. By the same token, a vector is an operator that accepts a dual vector as an argument and produces a real number equal to the scalar product, $V(\omega) = \omega_\mu V^\mu$. These definitions involve no uncontracted indices, so the results are independent of any basis choice. This suggests that vectors and dual vectors may be defined fundamentally in terms of linear maps to the real numbers:

1. A *dual vector* is an operator that acts linearly on a vector to return a real number.
2. A *vector* is as an operator that acts linearly on a dual vector to return a real number.

For those having a knowledge of linear algebra or quantum mechanics this may sound vaguely familiar, as suggested by the following example.

Example 8.6 In the language of linear algebra, vectors may be represented as column vectors and dual vectors as row vectors, and their matrix product is a number. For example,

$$A \equiv (a \;\; b) \qquad B \equiv \begin{pmatrix} c \\ d \end{pmatrix} \qquad AB = (a \;\; b) \begin{pmatrix} c \\ d \end{pmatrix} = ac + bd \in \mathbb{R}$$

may be regarded as the dual vector A acting linearly on the vector B to produce the real number $ac+bd$: $A(B) \in \mathbb{R}$. For Dirac notation in quantum mechanics, $|a\rangle$ may be viewed as representing a vector and $\langle a|$ as representing a dual vector *in Hilbert space*, and mathematically the vector space of $\langle a|$ is the dual of the vector space of $|a\rangle$ (see Ref. [24]). Thus the overlap $\langle f|\, i\rangle$ is a number, and a matrix element $\langle f|M|i\rangle$ is a map from vectors and dual vectors of Hilbert space to the real numbers.

8.4.7 Evaluating Components in a Basis

The preceding definitions of vectors and dual vectors are independent of any choice of basis, but practically it often is convenient to work in a basis. The components of a vector or dual vector in a specific basis are obtained by evaluating them with respect to the corresponding basis vectors and dual basis vectors; for example,

$$V^\mu = V(e^\mu) = e^\mu \cdot V \qquad \omega_\mu = \omega(e_\mu) = e_\mu \cdot \omega, \tag{8.65}$$

which follows from Eq. (8.63). Equations such as (8.65) may be interpreted (for example) as a vector accepting a basis vector e^μ as input and acting linearly on it to return a real number that is the component of the vector evaluated in that basis.

of vector fields on dual vector fields and vice versa, in which case what is returned is not a real number but rather a scalar field of real numbers defined over the manifold. We trust that the reader is sophisticated enough at this point to realize when "thing" really means "field of things".

Example 8.7 The validity of Eqs. (8.65) may be checked easily:

$$e^\mu \cdot V = e^\mu \cdot (V^\alpha e_\alpha) = V^\alpha e^\mu (e_\alpha) = V^\alpha \delta^\mu_\alpha = V^\mu,$$
$$e_\mu \cdot \omega = e_\mu \cdot (\omega_\alpha e^\alpha) = \omega_\alpha e_\mu (e^\alpha) = \omega_\alpha \delta^\alpha_\mu = \omega_\mu,$$

where the expansions (8.61)–(8.62), the linearity requirement (8.60), and the orthogonality condition (8.63) were employed.

Vector and dual vector components as in Eq. (8.65), and more generally tensor components, are then found to obey the same transformation laws and tensor calculus that will be presented in following sections as alternative defining characteristics of tensors. Example 8.8 illustrates for dual vectors and vectors.

Example 8.8 Consider a coordinate transformation $x^\mu \to x'^\mu$ on a dual vector $\omega = \omega_\mu e^\mu$ and on a vector $V = V^\mu e_\mu$. The basis dual vectors e^μ and basis vectors e_μ transform as

$$e^\mu \to e'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} e^\nu \qquad e_\mu \to e'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} e_\nu .$$

How do the components ω_μ transform? This may be determined by noting that the dual vector ω is a geometrical object having an existence independent of representation in a specific coordinate system, so it must be invariant under coordinate transformations. This will be ensured only if the components of ω transform as

$$\omega_\nu \to \omega'_\nu = \frac{\partial x^\alpha}{\partial x'^\nu} \omega_\alpha ,$$

since then the dual vector ω is invariant under $x^\mu \to x'^\mu$:

$$\begin{aligned} \omega' &= \omega'_\mu e'^\mu \\ &= \frac{\partial x^\alpha}{\partial x'^\mu} \omega_\alpha \frac{\partial x'^\mu}{\partial x^\nu} e^\nu \\ &= \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x'^\mu}{\partial x^\nu} \omega_\alpha e^\nu \\ &= \omega_\alpha e^\nu \delta^\alpha_\nu \\ &= \omega_\alpha e^\alpha = \omega . \end{aligned}$$

This transformation law for the components ω_ν is the same one that will be used to *define* a dual vector in Eq. (8.71). By a similar proof, vector components V^μ may be shown to have the transformation law (8.73).

In later sections, such transformation laws will be offered as a *definition* of tensors. In the index-free picture currently under discussion tensors are defined instead as linear maps of some number of vectors and dual vectors to the real numbers, with the transformation laws and associated tensor calculus that will be discussed in Sections 8.5–8.8 following as a *consequence* of that definition.

Let us now greatly simplify keeping track of the distinction between upper indices and lower indices on tensors by demonstrating that the metric tensor map may be used to establish a one-to-one relationship between a vector in the tangent space and a corresponding dual vector in the cotangent space.

8.4.8 Identification of Vectors and Dual Vectors

Consider the metric tensor, viewed as a rank-2 covariant tensor that accepts two vector inputs and acts on them (multi-)linearly to give a real number. Schematically, this may be written as the operator $g(\cdot\,,\cdot)$, where the dots indicate the input slots for the two vectors. Suppose that a vector V is inserted into only *one* of the slots, giving $g(V,\cdot)$. What is this object? It has one open slot that can accept a vector, on which it will act linearly to return a real number. But that should sound familiar: it is the definition of a dual vector! Because it is associated directly with the vector V, let us call this dual vector $\tilde{V} \equiv g(V\,,\cdot)$. The components of this dual vector may be evaluated by inserting a basis vector as argument in the usual way,

$$\begin{aligned} V_\mu &\equiv \tilde{V}(e_\mu) = g(V, e_\mu) \\ &= g(V^\nu e_\nu, e_\mu) \\ &= V^\nu g(e_\nu, e_\mu) \\ &= g_{\mu\nu} V^\nu . \end{aligned} \tag{8.66}$$

Likewise, by using that $g_{\mu\nu}$ and $g^{\mu\nu}$ are matrix inverses, $V^\mu = g^{\mu\nu} V_\nu$. Summing over repeated indices in tensor products is called *contraction*. Thus, the properties of the metric tensor allow vectors and dual vectors to be treated effectively as if they were both vectors, one with an upper index and one with a lower index, with the two related by contraction with the metric tensor,

$$V_\mu = g_{\mu\nu} V^\nu \qquad V^\mu = g^{\mu\nu} V_\nu . \tag{8.67}$$

This is of great practical importance since it allows the same symbol to be used for a vector and its corresponding dual vector, and it reduces the handling of vectors and dual vectors to keeping proper track of the vertical position of indices in the Einstein summation convention. This identification works only for manifolds with metric tensors but that is no limitation for special or general relativity, which deal *only* with metric spaces.

Once the operations of raising and lowering indices by contraction with the metric tensor are established through Eq. (8.67), the scalar product between two vectors U and V can be calculated as the *complete contraction* $U_\alpha V^\alpha$ of one of the vectors with the dual vector associated with the other vector:

$$g(U,V) = g_{\mu\nu} U^\mu V^\nu = U_\nu V^\nu . \tag{8.68}$$

The scalar product has no indices left after contraction and is said to be *fully contracted*. Because tensors of higher rank are products of vectors and dual vectors, the preceding discussion is easily generalized and contraction with the metric tensor can be used to raise or lower any index for a tensor of any rank. For example,

$$A^{\mu\nu} = g^{\mu\alpha} g^{\nu\beta} A_{\alpha\beta} \qquad A_{\mu\nu\lambda\sigma} = g_{\mu\rho} A^{\rho}{}_{\nu\lambda\sigma}.$$

Since indices can be raised or lowered at will by a metric, tensors may be thought of as objects of a particular tensorial rank, irrespective of their particular vertical arrangement of indices (for example, the identification of vectors and dual vectors discussed above). Of course this is true only in the abstract; index placement matters when tensors are evaluated in a basis.

8.4.9 Index-Free versus Component Transformations

The material in this section has been a brief introduction to the index-free formulation of tensors favored by mathematicians. The discussion above and that in Section 8.5 indicates that the index-free formalism leads to the same transformation laws for tensors. Thus the practical outcome will be the same with application of either approach to the issues addressed in these lectures, but index-free concepts provide a more solid mathematical foundation for physical results while often the component transformation approach affords a more direct path to obtaining them.

8.5 Tensors Specified by Transformation Laws

In the preceding discussion tensors have been introduced at a fundamental level through linear maps from vectors and dual vectors to the real numbers, but it was shown also that these linear maps imply that when tensors of a given type are expressed in an arbitrary basis (see Section 8.4.7) their components obey well-defined transformation laws under change of coordinates. This view of tensors as groups of quantities obeying particular transformation laws is often the most practical for physical applications because (1) it is less abstract and requires less new mathematics for the novice, (2) a physical interpretation often requires expression of the problem in a well-chosen basis anyway, and (3) the component index formalism has a built-in error checking mechanism of great practical utility in solving problems: failure of indices to balance on the two sides of an equation is a sure sign of an error.

This section summarizes the use of tensors to formulate invariant equations by exploiting the transformation properties of their components. Tensors may be viewed as generalizing the idea of scalars and vectors, so let's begin with these more familiar quantities. The following discussion is an adaptation to 4-dimensional (possibly curved) spacetime of many concepts discussed earlier in Section 8.3 for simpler euclidean spaces; you are urged to review that material if any conceptual difficulties are encountered in the following material.

8.5.1 Scalar Transformation Law

Consider the behavior of fields under the coordinate transformations introduced in Section 8.4.1. The simplest possibility is that the field has a single component at each point of the manifold, with a value that is unchanged by the transformation (8.48),

$$\phi'(x') = \phi(x). \tag{8.69}$$

Quantities such as $\phi(x)$ that are unchanged under the coordinate transformation are called *scalars*. As the notation indicates, scalar quantities are generally functions of the coordinates but their value at a given point does not change if the coordinate system changes.[10] Scalar quantities may be expected to play a central role in physical theories because measurable observables must be scalars if the goal of formulating physical laws such that they do not depend on the coordinate labels is to be fulfilled.

8.5.2 Dual Vector Transformation Law

In parallel with the discussion of vectors defined in the tangent and dual bases for euclidean space in Section 8.6, it is useful to define two kinds of spacetime vectors having distinct transformation laws. Both will be termed vectors because sets of each type separately obey the axioms to form a *vector space,* as discussed in Box 8.3, and because of the duality discussed in Section 8.3.5. By the argument in Section 8.4.8, the same symbol will be used for both but they will be distinguished by vertical placement of indices. By the rules of ordinary partial differentiation the gradient of a scalar field $\phi(x)$ obeys

$$\frac{\partial \phi(x)}{\partial x'^{\mu}} = \frac{\partial \phi(x)}{\partial x^{\nu}} \frac{\partial x^{\nu}}{\partial x'^{\mu}}. \tag{8.70}$$

Now consider a vector having a transformation law mimicking that of the scalar field gradient (8.70),

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)}. \tag{8.71}$$

A quantity transforming in this way is termed a *dual vector* (it also goes by the names *one-form*, *covariant vector,* or *covector*). Understand clearly that in (8.71) the two sides of the equation refer to the *same point* in spacetime. Thus, the argument is x' on the left side and x on the right side, and these label the same spacetime point in two different coordinate systems. Before continuing we stop to note that even the relatively simple expressions that have been introduced so far emphasize the importance of the compact notation that we are

[10] This may be contrasted with the behavior of the components of a vector under coordinate transformation, which will be discussed shortly. Geometrically the components of a vector are projections of the vector on particular coordinate axes. Thus, if the coordinate system is changed the components of a vector at a given point typically change their values, but the length of the vector, which is a scalar, is invariant.

using. For example, Eq. (8.71) is a concise way to write the matrix equation

$$\begin{pmatrix} A'_0 \\ A'_1 \\ A'_2 \\ A'_3 \end{pmatrix} = \begin{pmatrix} \partial x^0/\partial x'^0 & \partial x^1/\partial x'^0 & \partial x^2/\partial x'^0 & \partial x^3/\partial x'^0 \\ \partial x^0/\partial x'^1 & \partial x^1/\partial x'^1 & \partial x^2/\partial x'^1 & \partial x^3/\partial x'^1 \\ \partial x^0/\partial x'^2 & \partial x^1/\partial x'^2 & \partial x^2/\partial x'^2 & \partial x^3/\partial x'^2 \\ \partial x^0/\partial x'^3 & \partial x^1/\partial x'^3 & \partial x^2/\partial x'^3 & \partial x^3/\partial x'^3 \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix}.$$

Furthermore, although we usually suppress it in the notation, the partial derivatives appearing in these transformation equations *depend on spacetime coordinates* in the general case and all partial derivatives are understood implicitly to be evaluated at a specific point P labeled by x in one coordinate system and x' in the other.

8.5.3 Vector Transformation Law

Now consider application of the rules of partial differentiation to transformation of the differential,

$$dx'^{\mu} = \frac{\partial x'^{\mu}}{\partial x^{\nu}} dx^{\nu}. \tag{8.72}$$

This suggests a second vector transformation rule,

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}. \tag{8.73}$$

A quantity behaving in this way is termed a *vector*.[11] Most physical quantities that are thought of loosely as "vectors" (displacement or velocity, for example) are vectors in the restricted sense defined by the transformation law (8.73).

Example 8.9 Equations (8.71) and (8.73) may be viewed as matrix equations,

$$A'_{\mu}(x') = \hat{U}^{\nu}_{\mu} A_{\nu}(x) \qquad A'^{\mu}(x') = U^{\mu}_{\nu} A^{\nu}(x), \tag{8.74}$$

with the matrices $U = \partial x'/\partial x$ and $\hat{U} = \partial x/\partial x'$ obeying $\hat{U}U = I$, where I is the unit matrix. In these transformations the matrix U is called the *Jacobian matrix* and the matrix $\hat{U}$ is called the *inverse Jacobian matrix.*

8.5.4 Duality of Vectors and Dual Vectors

The preceding discussion distinguishes two kinds of "vectors": dual vectors, which carry a lower index and transform like (8.71), and vectors, which carry an upper index and transform like (8.73). This distinction is analogous to that introduced in Section 8.6 for vectors and dual vectors in euclidean spaces. In particular instances vectors and dual vectors may

[11] Some authors call vectors *contravariant vectors,* indicating explicitly that when expanded in a basis the component index is in the upper position, and call dual vectors *covariant vectors,* indicating explicitly that the component index is in the lower position.

be considered equivalent as a practical matter (albeit with some loss of mathematical rigor), but generally they are not. However, the duality discussed in Section 8.4.8 allows us to use the same symbol for vectors and dual vectors, with the distinction between them residing in upper and lower positioning of indices in the summation convention.

8.6 Scalar Product of Vectors

Just as was found in Section 8.3.4 for euclidean spaces, the introduction of vectors and dual vectors, and their relationship through the metric tensor, allows a natural definition of a scalar product

$$A \cdot B \equiv A_\mu B^\mu, \tag{8.75}$$

where the dual vector A_μ and the corresponding vector A^μ are related through the metric tensor according to $A_\mu = g_{\mu\nu} A^\nu$. This product transforms as a scalar because from Eqs. (8.71) and (8.73),

$$\begin{aligned} A' \cdot B' = A'{}_\mu B'^\mu &= \frac{\partial x^\nu}{\partial x'^\mu} A_\nu \frac{\partial x'^\mu}{\partial x^\alpha} B^\alpha = \frac{\partial x^\nu}{\partial x'^\mu} \frac{\partial x'^\mu}{\partial x^\alpha} A_\nu B^\alpha \\ &= \frac{\partial x^\nu}{\partial x^\alpha} A_\nu B^\alpha = \delta^\nu_\alpha A_\nu B^\alpha = A_\alpha B^\alpha = A \cdot B, \end{aligned} \tag{8.76}$$

where the Kronecker delta δ^ν_μ is given by

$$\delta^\mu_\nu = \frac{\partial x'^\mu}{\partial x'^\nu} = \frac{\partial x^\mu}{\partial x^\nu} = \begin{cases} 1 & \mu = \nu \\ 0 & \mu \neq \nu \end{cases}, \tag{8.77}$$

which is a rank-2 tensor with the unusual property that its components take the same value in all coordinate systems.

8.7 Tensors of Higher Rank

Three types of rank-2 tensors $(0,2)$, $(1,1)$, and $(2,0)$ may be distinguished; they have the transformation laws

$$T'{}_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta}, \tag{8.78}$$

$$T'^\nu{}_\mu = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x'^\nu}{\partial x^\beta} T^\beta{}_\alpha, \tag{8.79}$$

$$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha} \frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}. \tag{8.80}$$

These transformation rules may be generalized easily to tensors of any rank. In such a generalization,

Table 8.1 Some tensor transformation laws

Tensor	Transformation law
Scalar	$\phi' = \phi$
Dual vector	$A'_\mu = \dfrac{\partial x^\nu}{\partial x'^\mu} A_\nu$
Vector	$A'^\mu = \dfrac{\partial x'^\mu}{\partial x^\nu} A^\nu$
Covariant rank-2	$T'_{\mu\nu} = \dfrac{\partial x^\alpha}{\partial x'^\mu} \dfrac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta}$
Contravariant rank-2	$T'^{\mu\nu} = \dfrac{\partial x'^\mu}{\partial x^\alpha} \dfrac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}$
Mixed rank-2	$T'^{\nu}{}_{\mu} = \dfrac{\partial x^\alpha}{\partial x'^\mu} \dfrac{\partial x'^\nu}{\partial x^\beta} T^{\beta}{}_{\alpha}$

1. Each upper index on the left side requires a right-side "factor" of the form $\partial x'^\mu / \partial x^\nu$ (prime in the numerator), and
2. each lower index on the left side requires a right-side "factor" of the form $\partial x^\mu / \partial x'^\nu$ (prime in the denominator).

This "position of the left-side index equals position of the right-side primed coordinate" rule for the partial derivative factors is a useful aid in remembering the forms of the tensor transformation equations. Transformation laws for tensors through rank 2 are summarized in Table 8.1.

8.8 The Metric Tensor

By analogy with the discussion in Section 8.3.7, a rank-2 tensor of special importance is the metric tensor $g_{\mu\nu}$, because it is associated with the line element

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{8.81}$$

that determines the *geometry of the manifold.* The metric tensor is symmetric in its indices ($g_{\mu\nu} = g_{\nu\mu}$) and satisfies the $(0,2)$ tensor transformation law

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta}. \tag{8.82}$$

The contravariant form of the metric tensor $g^{\mu\nu}$ is defined by the requirement

$$g_{\mu\alpha} g^{\alpha\nu} = \delta^\nu_\mu, \tag{8.83}$$

so $g_{\mu\nu}$ and $g^{\mu\nu}$ are *matrix inverses.* Contractions with the metric tensor may be used to raise and lower (any number of) tensor indices; for example,

$$A^{\mu} = g^{\mu\nu} A_{\nu} \qquad A_{\mu} = g_{\mu\nu} A^{\nu} \qquad T^{\mu}{}_{\nu} = g_{\nu\alpha} T^{\mu\alpha} \qquad T^{\alpha}{}_{\beta\gamma} = g^{\alpha\mu} g_{\gamma\varepsilon} T_{\mu\beta}{}^{\varepsilon}. \tag{8.84}$$

Thus, the scalar product of vectors may be expressed as

$$A \cdot B = g_{\mu\nu} A^{\mu} B^{\nu} \equiv A_{\nu} B^{\nu} = g^{\mu\nu} A_{\mu} B_{\nu} \equiv A^{\nu} B_{\nu}. \tag{8.85}$$

Because such products are scalars, they are unchanged by coordinate-system transformations. A specific example of such a conserved quantity is an invariant length such as the line element of Eq. (8.81).

In mixed-tensor expressions like the third or fourth ones in Eq. (8.84) the relative horizontal order of upper and lower indices can be important. For example, in $T^{\mu}{}_{\nu} = g_{\nu\alpha} T^{\mu\alpha}$ the notation indicates that the mixed tensor on the left side of the equation was obtained by lowering the rightmost index of $T^{\mu\alpha}$ on the right side (since in $T^{\mu}{}_{\nu}$ the lower index ν is to the right of the upper index μ). This distinction is immaterial if the tensor is symmetric under exchange of indices but which index is lowered or raised by contraction matters for tensors that are antisymmetric under index exchange (see Section 8.9): $T^{\mu}{}_{\nu} = g_{\nu\alpha} T^{\mu\alpha}$ and $T_{\nu}{}^{\mu} = g_{\nu\alpha} T^{\alpha\mu}$ are equivalent if T is symmetric, but different if T is antisymmetric.

Vectors and dual vectors are distinct entities that are defined in different spaces (see Section 8.4.2). However, Eqs. (8.83)–(8.85) and the discussion in Section 8.3.5, and Section 8.4.8 make it clear that for the *special case of a manifold with metric,* indices on any tensor may be raised or lowered at will by contraction with the metric tensor, as in Eq. (8.84). Defining a metric establishes a relationship that permits vectors and dual vectors to be treated as if they were (in effect) different representations of the same vector. Our discussion will usually proceed as if A^{μ} and A_{μ} are different forms of the same vector that are related by contraction with the metric tensor, while secretly remembering that they really are different, and that it is only for metric spaces that this conflation is not likely to land us in trouble.

The preceding discussion makes clear why the convention in much of nonrelativistic physics to ignore the mathematical distinction between vectors and dual vectors causes few problems. For example, the gradient operator is commonly termed a vector in elementary physics but Section 8.5.2 shows that it is in truth a dual vector. However, physics assumes metric spaces, so vectors and dual vectors are related trivially through the metric tensor and no lasting harm is done by calling the gradient a vector if the bookkeeping is done correctly in equations. Specifically, the gradient dual vector may be converted to the corresponding vector by contracting with the metric tensor to raise the index. The issue is particularly simple if the problem is formulated in euclidean space using cartesian coordinates, in which case the metric tensor is just the unit matrix and the components of the vector and dual vector are the same. The mathematician is required to be more circumspect because the definition of a manifold does not automatically imply existence of a metric to enable this identification.

8.9 Symmetric and Antisymmetric Tensors

The symmetry of tensors under exchanging pairs of indices is often important. An arbitrary rank-2 tensor can always be decomposed into a symmetric and antisymmetric part according to the identity

$$T_{\alpha\beta} = \tfrac{1}{2}(T_{\alpha\beta} + T_{\beta\alpha}) + \tfrac{1}{2}(T_{\alpha\beta} - T_{\beta\alpha}), \tag{8.86}$$

where the first term is clearly symmetric and the second term antisymmetric under exchange of indices. For completely symmetric and completely antisymmetric rank-2 tensors

$$T_{\alpha\beta} = \pm T_{\beta\alpha} \qquad T^{\alpha\beta} = \pm T^{\beta\alpha},$$

where the plus sign holds if the tensor is symmetric and the minus sign if it is antisymmetric. More generally, a tensor of rank two or higher is said to be *symmetric* in any two of its indices if exchanging those indices leaves the tensor invariant and *antisymmetric* (or *skew-symmetric*) in any two indices if it changes sign upon switching those indices.

8.10 Algebraic Tensor Operations

Various algebraic operations are permitted for tensors in equations. These valid operations include:

1. *Multiplication by a scalar:* A tensor may be multiplied by a scalar (meaning that each component is multiplied by the scalar) to produce a tensor of the same rank. For example, $aA^{\mu\nu} = B^{\mu\nu}$, where a is a scalar and $A^{\mu\nu}$ and $B^{\mu\nu}$ are rank-2 contravariant tensors.
2. *Addition or subtraction:* Two tensors of the same type may be added or subtracted (meaning that their components are added or subtracted) to produce a new tensor of the same type. For example, $A^\mu - B^\mu = C^\mu$, where A^μ, B^μ, and C^μ are vectors.
3. *Multiplication:* Two or more tensors may be multiplied by forming products of their components. The rank of the resultant tensor will be the product of the ranks of the tensor factors. For example, $A_{\mu\nu} = U_\mu V_\nu$, where $A_{\mu\nu}$ is a rank-2 covariant tensor and U_μ and V_ν are dual vectors.
4. *Contraction:* For a tensor or tensor product with covariant rank n and contravariant rank m, a tensor of covariant rank $n-1$ and contravariant rank $m-1$ may be formed by setting one upper and one lower index equal and taking the implied sum. For example, $A = A^\mu{}_\mu$, where A is a scalar and $A^\mu{}_\nu$ is a mixed rank-2 tensor, or $A_\mu = g_{\mu\nu}A^\nu$, where the metric $g_{\mu\nu}$ is a rank-2 covariant tensor, A^ν is a vector, and A_μ is a dual vector.

In addition to these algebraic manipulations, it will often be necessary to integrate or differentiate in tensor expressions. This *tensor calculus* is addressed in the following section.

8.11 Tensor Calculus on Curved Manifolds

To formulate physical theories in terms of tensors requires the ability to manipulate tensors mathematically. In addition to the algebraic rules for tensors described in preceding sections, we must formulate a prescription to integrate tensor equations and one to differentiate them. Tensor calculus is mostly a straightforward generalization of normal calculus but additional complexity arises for two reasons:

1. It must be ensured that integration and differentiation preserve the symmetries and invariances embodied in the tensor equations.
2. To preserve the utility of the tensor formalism, it must be ensured that the results of these operations on tensor quantities are themselves tensor quantities.

As will now be shown, the first requirement implies a simple modification of the rules for ordinary integration while the second implies a less-simple modification with far-reaching mathematical and physical implications for the rules of partial differentiation.

8.11.1 Invariant Integration

Under a change of coordinates the volume element for integration over the spacetime coordinates changes according to

$$d^4x' = \det\left(\frac{\partial x'}{\partial x}\right) d^4x = Jd^4x, \tag{8.87}$$

where $d^4x \equiv dx^0dx^1dx^2dx^3$ and $J \equiv \det(\partial x'/\partial x)$ is the Jacobian determinant (determinant of the 4×4 matrix of partial derivatives relating the x and x' coordinates; see Example 8.9). Notice that the right side of Eq. (8.82) may be viewed as a triple matrix product and recall that the determinant of a matrix product is the product of determinants. Thus Eq. (8.82) implies that the determinant of the metric tensor $g \equiv \det g_{\mu\nu}$ transforms as $g' = J^{-2}g$. Hence $J = \sqrt{|g|}/\sqrt{|g'|}$, where the absolute value signs are necessary because the determinant of $g_{\mu\nu}$ is negative in 4D spacetime with Lorentzian metric signature [see Eq. (9.6)], and this result may be substituted into Eq. (8.87) to give $\sqrt{|g'|}\,d^4x' = \sqrt{|g|}\,d^4x$. This implies that an *invariant volume element,*

$$dV = \sqrt{|g|}\,d^4x, \tag{8.88}$$

must be employed in tensor integration to ensure that the results are invariant under a change of coordinate system. A simple demonstration of using invariant integration to determine an area for a 2-dimensional curved manifold is given in Example 8.10.

Example 8.10 The metric for a 2-dimensional spherical surface (the 2-sphere S^2) is specified by the line element $d\ell^2 = R^2d\theta^2 + R^2\sin^2\theta d\phi^2$, which may be written as the matrix equation

$$d\ell^2 = (d\theta \quad d\phi)\begin{pmatrix} R^2 & 0 \\ 0 & R^2\sin^2\theta \end{pmatrix}\begin{pmatrix} d\theta \\ d\phi \end{pmatrix}.$$

The area of the 2-sphere may then be calculated as

$$\begin{aligned} A &= \int_0^{2\pi} d\phi \int_0^{\pi} \sqrt{\det g_{ij}}\, d\theta \\ &= \int_0^{2\pi} d\phi \int_0^{\pi} R^2 \sin\theta d\theta = 4\pi R^2, \end{aligned}$$

where the metric tensor g_{ij} is the 2×2 matrix in the first equation for the line element. In this 2-dimensional example the sign of the determinant is positive, so no absolute value is required under the radical in Eq. (8.88).

Thus, the extension of ordinary integration to integration over tensor fields requires only that the volume element be made invariant according to the prescription in Eq. (8.88). What about the derivatives of tensor quantities? This is a more complicated issue that we must now address.

8.11.2 Partial Derivatives

Before proceeding it will be useful to introduce a more compact way to write partial derivatives. Two shorthand notations are in common use, as illustrated by the following examples.[12]

$$\partial_\mu \phi = \phi_{,\mu} \equiv \frac{\partial \phi(x)}{\partial x^\mu} \qquad \partial'_\mu \phi' = \phi'_{,\mu} \equiv \frac{\partial \phi'(x')}{\partial x'^\mu} \qquad \partial^\mu \phi = \phi^{,\mu} \equiv \frac{\partial \phi(x)}{\partial x_\mu}. \qquad (8.89)$$

All three of the notations exhibited for partial derivatives in these equations will be used at various places in this book. Let us now consider the covariance of the partial derivative operation applied to tensors.

The transformation law for the derivative of a scalar is given by Eq. (8.70), which is just the transformation law (8.71); therefore the derivative of a scalar is a dual vector and scalars and their first derivatives have well-defined tensorial properties. So far, so good, but now consider the derivative of a dual vector,

$$\begin{aligned} A'_{\mu,\nu} \equiv \frac{\partial A'_\mu}{\partial x'^\nu} &= \frac{\partial}{\partial x'^\nu}\left(A_\alpha \frac{\partial x^\alpha}{\partial x'^\mu}\right) \\ &= \frac{\partial A_\alpha}{\partial x'^\nu}\frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu} \\ &= \frac{\partial A_\alpha}{\partial x^\beta}\frac{\partial x^\beta}{\partial x'^\nu}\frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu} \\ &= A_{\alpha,\beta}\frac{\partial x^\beta}{\partial x'^\nu}\frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu}. \end{aligned} \qquad (8.90)$$

[12] Higher-order derivatives can be denoted by additional subscripts. For example $\partial^2\phi/\partial x^\mu \partial x^\nu = \phi_{,\mu\nu}$.

The first term in the last line transforms as a (rank-2) tensor but the second term does not since it involves second derivatives, ultimately because the partial-derivative matrix implementing the transformation is *position dependent* in curved spacetime. In flat space it is possible to choose coordinates where the second term vanishes but in curved spacetime it cannot be transformed away globally. By a similar procedure it is found that the partial derivatives of vectors and all higher-order tensors exhibit a similar pathology:

> With the exception of derivatives of scalars, ordinary partial differentiation of tensors is *not a covariant operation in curved spacetime* because it fails to preserve the tensor structure of equations.

This will complicate the formalism immensely because the utility of the tensor framework rests on the preservation of tensor structure under transformations. It is desirable to define a new covariant derivative operation that does this automatically. In general the terms that violate the tensor transformation laws for partial derivatives of tensors will involve second derivatives, as in Eq. (8.90). The non-tensorial contributions can be eliminated systematically by introducing additional fields on the manifold, which can be done in more than one way, each leading to a different form of covariant differentiation. Three common approaches are (1) *covariant derivatives* and (2) *absolute derivatives,* which use derivatives of the metric tensor field to cancel non-tensorial terms, and (3) *Lie derivatives,* which use derivatives of an auxiliary vector field defined on the manifold to the same end. We will discuss here only covariant derivatives, which will be introduced in Section 8.11.3

8.11.3 Covariant Derivatives

For the manifolds important in general relativity, the most common approach to converting partial differentiation into an operation that preserves tensor structure is to use particular linear combinations of metric-tensor derivatives to create new non-tensorial terms that *exactly cancel* the non-tensorial terms arising from taking the partial derivative. Notice that if the *Christoffel symbols* $\Gamma^{\lambda}_{\alpha\beta}$ are introduced and required to obey a transformation law

$$\Gamma'^{\lambda}_{\alpha\beta} = \Gamma^{\kappa}_{\mu\nu}\frac{\partial x^{\mu}}{\partial x'^{\alpha}}\frac{\partial x^{\nu}}{\partial x'^{\beta}}\frac{\partial x'^{\lambda}}{\partial x^{\kappa}} + \frac{\partial^2 x^{\mu}}{\partial x'^{\alpha}\partial x'^{\beta}}\frac{\partial x'^{\lambda}}{\partial x^{\mu}}, \tag{8.91}$$

it may be shown that

$$\left(A'_{\mu,\nu} - \Gamma'^{\lambda}_{\mu\nu}A'_{\lambda}\right) = \left(A_{\alpha,\beta} - \Gamma^{\kappa}_{\alpha\beta}A_{\kappa}\right)\frac{\partial x^{\alpha}}{\partial x'^{\mu}}\frac{\partial x^{\beta}}{\partial x'^{\nu}}. \tag{8.92}$$

Comparing with Eq. (8.78), the quantity in brackets is seen to transform as a rank-2 covariant tensor. This suggests the utility of introducing a new derivative operation: the *covariant derivative* of a dual vector is defined to be

$$A_{\mu;\nu} \equiv A_{\mu,\nu} - \Gamma^{\lambda}_{\mu\nu}A_{\lambda}, \tag{8.93}$$

where now a subscript comma denotes ordinary partial differentiation and a subscript semicolon denotes covariant differentiation with respect to the variables following it. It will be useful to introduce also an alternative notation for the covariant derivative,

$$\nabla_\nu A_\mu = A_{\mu;\nu} \equiv \partial_\nu A_\mu - \Gamma^\lambda_{\mu\nu} A_\lambda. \tag{8.94}$$

This covariant derivative of a dual vector then transforms as a covariant tensor of rank 2: neither of its terms is a tensor but their *difference* is. Likewise, the covariant derivative of a vector can be introduced in either of the notations

$$A^\lambda{}_{;\mu} = A^\lambda{}_{,\mu} + \Gamma^\lambda_{\alpha\mu} A^\alpha \qquad \nabla_\mu A^\lambda = \partial_\mu A^\lambda + \Gamma^\lambda_{\alpha\mu} A^\alpha, \tag{8.95}$$

where the result of (8.95) is a mixed rank-2 tensor, and the covariant derivatives of the three possible rank-2 tensors through

$$A_{\mu\nu;\lambda} = A_{\mu\nu,\lambda} - \Gamma^\alpha_{\mu\lambda} A_{\alpha\nu} - \Gamma^\alpha_{\nu\lambda} A_{\mu\alpha}, \tag{8.96}$$

$$A^\mu{}_{\nu;\lambda} = A^\mu{}_{\nu,\lambda} + \Gamma^\mu_{\alpha\lambda} A^\alpha{}_\nu - \Gamma^\alpha_{\nu\lambda} A^\mu{}_\alpha, \tag{8.97}$$

$$A^{\mu\nu}{}_{;\lambda} = A^{\mu\nu}{}_{,\lambda} + \Gamma^\mu_{\alpha\lambda} A^{\alpha\nu} + \Gamma^\nu_{\alpha\lambda} A^{\mu\alpha}, \tag{8.98}$$

or in alternative notation

$$\nabla_\lambda A_{\mu\nu} = \partial_\lambda A_{\mu\nu} - \Gamma^\alpha_{\mu\lambda} A_{\alpha\nu} - \Gamma^\alpha_{\nu\lambda} A_{\mu\alpha}, \tag{8.99}$$

$$\nabla_\lambda A^\mu{}_\nu = \partial_\lambda A^\mu{}_\nu + \Gamma^\mu_{\alpha\lambda} A^\alpha{}_\nu - \Gamma^\alpha_{\nu\lambda} A^\mu{}_\alpha, \tag{8.100}$$

$$\nabla_\lambda A^{\mu\nu} = \partial_\lambda A^{\mu\nu} + \Gamma^\mu_{\alpha\lambda} A^{\alpha\nu} + \Gamma^\nu_{\alpha\lambda} A^{\mu\alpha}, \tag{8.101}$$

where the derivatives in Eqs. (8.96)–(8.101) define rank-3 tensors. In the general case the covariant derivative of a tensor is a tensor of one rank higher than the tensor being differentiated and the heuristic for constructing it is to form the ordinary partial derivative and add one Christoffel symbol term having the sign and form for a dual vector for each lower index, and one having the sign and form for a vector for each upper index of the tensor (see Example 8.11).

In general relativity the Christoffel symbols are equivalent to *connection coefficients* (hence the same notation will be employed for both), which have a geometrical significance on the manifold and can be defined in terms of the derivatives of the metric tensor as

$$\Gamma^\sigma_{\lambda\mu} = \tfrac{1}{2} g^{\nu\sigma} \left(\frac{\partial g_{\mu\nu}}{\partial x^\lambda} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\lambda}}{\partial x^\nu} \right).$$

Thus the correction terms in Eqs. (8.93)–(8.101) that cancel non-tensorial character are indeed composed of derivatives of the metric tensor, as promised above.

Rules for covariant differentiation: Most of the rules for partial differentiation carry over with suitable generalization for covariant differentiation. For instance, the ordinary (Leibniz) rule for differentiating a product applies also to covariant differentiation,

$$(A_\mu B_\nu)_{;\lambda} = A_{\mu;\lambda} B_\nu + A_\mu B_{\nu;\lambda}. \tag{8.102}$$

The most important exception concerns the results of successive covariant differentiations. Partial derivative operators normally commute: if two are applied successively, the outcome does not depend on the order in which they are applied. However, covariant derivative operators generally do not commute and two successive covariant differentiations may give results that depend on the order in which they are applied. It may be shown that covariant derivatives, and their non-commuting nature, arise naturally from a prescription for parallel transport of vectors in curved spaces.

Example 8.11 The Leibniz differentiation rule for the product of two vectors may be used to derive the expressions (8.96)–(8.98) for the covariant derivatives of rank-2 tensors from those given for vectors in Eqs. (8.93) and (8.95). For example,

$$\begin{aligned}(U^\alpha V^\beta)_{;\gamma} &= U^\alpha V^\beta{}_{;\gamma} + U^\alpha{}_{;\gamma} V^\beta \\ &= U^\alpha V^\beta{}_{,\gamma} + U^\alpha(\Gamma^\beta_{\rho\gamma} V^\rho) + U^\alpha{}_{,\gamma} V^\beta + (\Gamma^\alpha_{\rho\gamma} U^\rho) V^\beta \\ &= (U^\alpha V^\beta)_{,\gamma} + \Gamma^\beta_{\rho\gamma}(U^\alpha V^\rho) + \Gamma^\alpha_{\rho\gamma}(U^\rho V^\beta),\end{aligned}$$

where Eq. (8.95) and $U^\alpha V^\beta{}_{,\gamma} + U^\alpha{}_{,\gamma} V^\beta = (U^\alpha V^\beta)_{,\gamma}$ were used in the second and third lines, respectively. This is equivalent to

$$A^{\alpha\beta}{}_{;\gamma} = A^{\alpha\beta}{}_{,\gamma} + \Gamma^\beta_{\rho\gamma} A^{\alpha\rho} + \Gamma^\alpha_{\rho\gamma} A^{\rho\beta},$$

where $A^{\alpha\beta} \equiv U^\alpha V^\beta$, which is Eq. (8.98) for the covariant derivative of a contravariant rank-2 tensor.

Carrying out similar manipulations as in this example for the rank-2 tensors $A_{\mu\nu}$ and $A^\nu{}_\mu$ suggests the rule given after Eq. (8.101): differentiate in the normal way and add one Christoffel symbol term having the sign and form for a dual vector for each lower index, and one having the sign and form for a vector for each upper index of the tensor.

Implications of covariant differentiation: One rather important consequence of requiring that differentiation preserve tensor structure in the manner described above is that the covariant derivative of the metric tensor vanishes at all points of a manifold,[13]

$$\nabla_\alpha g_{\mu\nu} = g_{\mu\nu;\alpha} = 0. \tag{8.103}$$

This means that in manipulating tensor equations the operation of covariant differentiation commutes with the operation (8.84) of raising or lowering an index by contraction with the metric tensor. For example,

$$g_{\alpha\beta} \nabla_\gamma V^\beta = \nabla_\gamma (g_{\alpha\beta} V^\beta) = \nabla_\gamma V_\alpha.$$

Thus the order of covariant differentiation and contraction with the metric tensor can be interchanged without altering the result.

[13] The validity of Eq. (8.103) is a consequence of particular assumptions concerning the nature of parallel transport in curved spacetime.

8.12 Invariant Equations

The properties of tensors elaborated above ensure that any equation will be form-invariant under general coordinate transformations if it equates tensor components having the same upper and lower indices. For example, if the quantities $A^{\mu}{}_{\nu}$ and $B^{\mu}{}_{\nu}$ each transform as mixed rank-2 tensors according to Eq. (8.79) and $A^{\mu}{}_{\nu} = B^{\mu}{}_{\nu}$ in the x coordinate system, then in the x' coordinate system $A'^{\mu}{}_{\nu} = B'^{\mu}{}_{\nu}$. Likewise, an equation that equates any tensor to zero (that is, sets all its components to zero) in some coordinate system is covariant under general coordinate transformations, implying that the tensor is equal to zero in all coordinate systems. However, equations such as $A^{\nu}{}_{\mu} = 10$ or $A^{\mu} = B_{\mu}$ might hold in particular coordinate systems but generally are not valid in all coordinate systems because they equate tensors of different kinds (a mixed rank-2 tensor with a scalar in the first example and a dual vector with a vector in the second).

The preceding discussion suggests that invariance of a theory under general coordinate transformations will be guaranteed by carrying out the following steps.

1. Formulate all quantities in terms of tensors, with tensor types matching on the two sides of any equation, and with all algebraic manipulations corresponding to valid tensor operations (addition, multiplication, contraction, . . .).
2. Redefine any integration to be invariant integration, as discussed in Section 8.11.1.
3. Replace all partial derivatives with the covariant derivatives introduced in Section 8.11.3.
4. Take care to remember that a covariant differentiation generally does not commute with a second covariant differentiation, so the order matters.

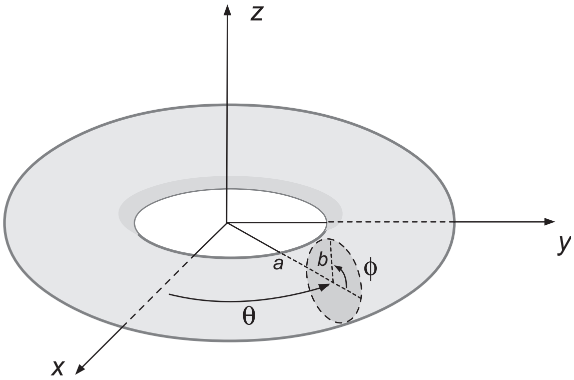
This prescription in terms of tensors will provide a powerful formalism for dealing with mathematical relations that would be much more formidable in standard notation.

Background and Further Reading

Significant parts of this chapter have been adapted from Refs. [11] and [12]. The discussion of coordinate transformations in euclidean space is based on the presentation of Ref. [5].

Problems

8.1 Consider the following torus, parameterized by the angles θ and ϕ.



Points on the torus are defined by

$$x = (a + b\cos\phi)\cos\theta \qquad y = (a + b\cos\phi)\sin\theta \qquad z = b\sin\phi,$$

where a and b are constants. Construct the tangent basis vectors for θ and ϕ, and the corresponding metric tensor.

8.2 (a) Verify explicitly that the Lorentz transformation of Eq. (9.25)

$$\begin{aligned} cdt' &= c\cosh\xi\, dt + \sinh\xi\, dx \\ dx' &= c\sinh\xi\, dt + \cosh\xi\, dx \\ dy' &= dy \\ dz' &= dz \end{aligned}$$

leaves invariant the Minkowski line element ds^2.

(b) Use the Lorentz transformations (9.31) expressed in differential form to obtain the velocity transformation rules consistent with Lorentz invariance. Show that for the special case of two inertial frames moving along the x axis with relative velocity v, the velocity transformation law is

$$u' = \frac{u - v}{1 - uv/c^2},$$

and that this embodies the constancy of the speed of light in all inertial frames, but reduces to the result expected from Galilean invariance for small v.

8.3 Use tangent and dual basis vectors to construct metric tensor components g^{ij} and g_{ij}, and the line element, for the coordinate system (u, v, w) of Example 8.3. Verify that the matrices g^{ij} and g_{ij} found for this problem are inverses of each other.

8.4 Show that if index raising and lowering operations are *defined* by expressions like

$$T^{\mu}{}_{\nu} = g_{\nu\alpha}T^{\mu\alpha} \qquad T_{\mu\nu} = g_{\mu\alpha}g_{\nu\beta}T^{\alpha\beta},$$

then these are valid tensor operations. That is, show that if these relations are true in one coordinate system they are true in all coordinate systems.

9 Special Relativity

In this chapter we shall build on the mathematical foundation of Ch. 8 to describe the special theory of relativity and use that to demonstate the Lorentz invariance of the Maxwell equations. Let's begin by noting the natural scientific and historical affinity of the Maxwell equations and the special theory of relativity.

9.1 Maxwell's Equations and Special Relativity

Scientifically, electrical charges may be viewed as classical objects (for example, as having positions and momenta that are simultaneously well defined) so that quantum mechanics is not required, but the motion of these classical charges could be described by Newtonian mechanics at low velocities, or by special relativity at velocities that are significant fractions of the speed of light c.[1] Historically, classical electromagnetism and special relativity have been intertwined because Einstein was influenced strongly by the beauty and symmetry properties of the Maxwell equations in his formulation of the special theory of relativity. In particular he was motivated by comparing the Lorentz invariance of the Maxwell equations with the Galilean invariance of classical mechanics to propose that it was classical mechanics, not electromagnetism, that required revision if one wanted the laws of electromagnetism and of classical particle motion to be mutually consistent. This led him to propose the radical notion that the speed of light is constant in all inertial frames, which is consistent with Maxwell's equations but not with Newtonian mechanics, and which forms the basis of the special theory of relativity.

9.2 Minkowski Space

A manifold equipped with a prescription for measuring distances is termed a *metric space* and the mathematical function that specifies distances is termed the *metric* for the space. In this section those ideas are applied to flat 4-dimensional spacetime, which is commonly termed *Minkowski space.* Although many concepts will be similar to those for euclidean

[1] If strong gravity is important, as it would be in various electromagnetic phenomena in astrophysics (for example, accretion disks for black holes), one must extend the description of charged-particle motion to curved spacetime. This requires use of the *general theory of relativity*. Here we will consider only relativistic electrodynamics in flat spacetime, so special but not general relativity will be necessary.

manifolds, fundamentally new features will enter. Many of these new features are associated with the *indefinite metric* of Minkowski space.

9.2.1 The Indefinite Metric of Spacetime

Although Minkowski space is flat it is not euclidean, for it does not possess a euclidean metric. A metric that can be put into a diagonal form in which the signs of the diagonal entries can all be chosen positive is termed *positive definite.* In contrast, we have see in Ch. 8 that the Minkowski metric has an essential property that the diagonal entries cannot all be chosen positive. Such a metric is termed *indefinite,* and it leads to properties of Minkowski space differing fundamentally from those of euclidean spaces.

9.2.2 Scalar Products and the Metric Tensor

In a particular inertial frame we may introduce unit vectors e_0, e_1, e_2, and e_3 that point along the t, x, y, and z axes, respectively, such that any 4-vector A may be expressed in the form,

$$A = A^0 e_0 + A^1 e_1 + A^2 e_2 + A^3 e_3. \tag{9.1}$$

Thus (A^0, A^1, A^2, A^3) are the (contravariant) components of the 4-vector A.[2] The scalar product of 4-vectors is given by

$$A \cdot B = B \cdot A = (A^\mu e_\mu) \cdot (B^\nu e_\nu) = e_\mu \cdot e_\nu A^\mu B^\nu. \tag{9.2}$$

Introducing the definition

$$\eta_{\mu\nu} \equiv e_\mu \cdot e_\nu, \tag{9.3}$$

we may express the scalar product as

$$A \cdot B = \eta_{\mu\nu} A^\mu B^\nu. \tag{9.4}$$

The metric tensor $\eta_{\mu\nu}$ is just a special case of the general metric tensor for spacetime that is generally denoted $g_{\mu\nu}$; however, in flat spacetime $g_{\mu\nu}$ is a constant matrix independent of the coordinates and it is common to denote it by the special symbol $\eta_{\mu\nu}$.

9.2.3 The Line Element

The line element ds^2 in Minkowski space measures the square of the distance between points with infinitesimal separation and is given by

$$ds^2 = -c^2 d\tau^2 = \eta_{\mu\nu} dx^\mu dx^\nu = -c^2 dt^2 + dx^2 + dy^2 + dz^2, \tag{9.5}$$

[2] Recall that in our notation non-bold symbols are being used to denote 4-vectors and bold symbols are reserved for 3-vectors, and that a notation such as A^μ may stand either for the full 4-vector, or for a component of it, depending on the context.

where τ is the proper time (the time measured by a clock carried in an inertial frame; thus, it is the time measured between events that are at the same spatial point), and where the metric tensor of flat spacetime may be represented by the constant diagonal matrix

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \equiv \text{diag}\ (-1,1,1,1). \tag{9.6}$$

Then Eq. (9.5) for the line element may be written as the matrix equation

$$ds^2 = (cdt\ \ dx\ \ dy\ \ dz) \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cdt \\ dx \\ dy \\ dz \end{pmatrix}, \tag{9.7}$$

where ds^2 represents the square of the spacetime interval between x and $x+dx$ with

$$x = (x^0, x^1, x^2, x^3) = (ct, x^1, x^2, x^3). \tag{9.8}$$

A point in Minkowski space defines an *event* and the path followed by an object in spacetime is termed the *worldline* for the object. This 4-dimensional spacetime with indefinite metric is termed a *Lorentzian manifold* (or sometimes a *pseudo-euclidean manifold*).

Example 9.1 Given a Minkowski vector with components (A^0, A^1, A^2, A^3), what are the components of the corresponding dual vector? From Eq. (8.84) with $\eta_{\mu\nu}$ substituted for the metric tensor, the indices may be lowered through the contraction $A_\mu = \eta_{\mu\nu}A^\nu$. Therefore, using the metric tensor (9.6) the elements of the corresponding dual vector are $A_\mu = (-A^0, A^1, A^2, A^3)$. This illustrates explicitly that vectors and dual vectors generally are not equivalent in non-euclidean manifolds, but that they are in one-to-one correspondence though contraction with the metric tensor.

9.2.4 Invariance of the Spacetime Interval

Special relativity follows from two assumptions: (1) the speed of light is constant for all observers and (2) the laws of physics cannot depend on spacetime coordinates. The postulate that the speed of light is a constant is equivalent to a statement that the spacetime interval ds^2 of Eq. (9.5) is an invariant that is unchanged by transformations between inertial systems (the Lorentz transformations to be discussed below). This is not true for the euclidean spatial interval $dx^2 + dy^2 + dz^2$, nor is it true for the time interval c^2dt^2; it is true only for the particular combination of spatial and time intervals defined by Eq. (9.5). Because of this invariance, Minkowski space is the natural manifold for the formulation of special relativity.

Example 9.2 The metric can be used to determine the relationship between the time coordinate t and the proper time τ. From Eq. (9.5)

$$\begin{aligned} d\tau^2 = \frac{-ds^2}{c^2} &= \frac{1}{c^2}(c^2dt^2 - dx^2 - dy^2 - dz^2) \\ &= dt^2 \left\{ 1 - \frac{1}{c^2}\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 \right] \right\} \\ &= \left(1 - \frac{v^2}{c^2}\right) dt^2. \end{aligned} \tag{9.9}$$

where v is the magnitude of the velocity. Therefore, the proper time that elapses between coordinate times t_1 and t_2 is

$$\tau_{12} = \int_{t_1}^{t_2} \left(1 - \frac{v^2}{c^2}\right)^{1/2} dt\,. \tag{9.10}$$

The proper time interval τ_{12} is shorter than the coordinate time interval $t_2 - t_1$ because the square root in the integrand of Eq. (9.10) is always less than one. This is the special-relativistic *time dilation effect,* stated in general form. For the special case of constant velocity, (9.10) yields

$$\Delta\tau = \left(1 - \frac{v^2}{c^2}\right)^{1/2} \Delta t, \tag{9.11}$$

which is the formulation of special-relativistic time dilation that is found commonly in textbooks. From this example it is clear that the origin of time dilation in special relativity lies in the geometry of spacetime, specifically in the indefinite nature of the Minkowski metric.

The first postulate of special relativity (constant speed of light for all observers) is ensured by the invariance of the interval (9.5) under transformations between inertial frames. As was suggested by the discussion in Ch. 8, the second postulate (coordinate invariance of physical law) can be ensured by formulating the equations of special relativity in terms of tensors defined in Minkowski space, which we now address.

9.3 Tensors in Minkowski space

In Minkowski space the transformations between coordinate systems are particularly simple because they are *independent of spacetime coordinates.* Furthermore, in flat space the correction terms disappear and partial derivatives are equivalent to covariant derivatives. Therefore, the derivatives appearing in the general definitions of Table 8.1 for tensors are constants and the transformation of a coordinate vector x^μ may be expressed as

$x'^{\mu} = \Lambda^{\mu}{}_{\nu} x^{\nu}$, where the matrix $\Lambda^{\mu}{}_{\nu}$ does not depend on the spacetime coordinates. Hence, for flat spacetime the tensor transformation laws simplify to

$$
\begin{aligned}
\phi' &= \phi && \text{Scalar} && (9.12)\\
A'^{\mu} &= \Lambda^{\mu}{}_{\nu} A^{\nu} && \text{Vector} && (9.13)\\
A'_{\mu} &= \Lambda_{\mu}{}^{\nu} A_{\nu} && \text{Dual vector} && (9.14)\\
T'^{\mu\nu} &= \Lambda^{\mu}{}_{\gamma} \Lambda^{\nu}{}_{\delta} T^{\gamma\delta} && \text{Contravariant rank-2 tensor} && (9.15)\\
T'_{\mu\nu} &= \Lambda_{\mu}{}^{\gamma} \Lambda_{\nu}{}^{\delta} T_{\gamma\delta} && \text{Covariant rank-2 tensor} && (9.16)\\
T'^{\mu}{}_{\nu} &= \Lambda^{\mu}{}_{\gamma} \Lambda_{\nu}{}^{\delta} T^{\gamma}{}_{\delta} && \text{Mixed rank-2 tensor} && (9.17)
\end{aligned}
$$

and so on. In addition, for flat spacetime it is possible to choose a coordinate system for which non-tensorial terms like the second term of Eq. (8.90) can be transformed away so *covariant derivatives are equivalent to partial derivatives* in Minkowski space. In the transformation laws (9.17) the $\Lambda^{\mu}{}_{\nu}$ are elements of *Lorentz transformations* that we will now discuss in more detail.

9.4 Lorentz Transformations

Inertial frames enjoy a privileged role in Newtonian mechanics. Newton's first law is unchanged in special relativity and inertial frames can be constructed in the same way as for Newtonian mechanics. What is different about the inertial frames of special relativity is that because of the requirements imposed by the constant speed of light and principle of relativity postulates, the transformations between inertial frames are no longer the Galilean transformations of Newtonian mechanics but rather the *Lorentz transformations.* Hence, the inertial frames of special relativity are often termed *Lorentz frames.* Rotations are an important class of transformations in euclidean space because they *change the direction but preserve the length* of an arbitrary 3-vector. It is desirable to generalize this idea to investigate abstract rotations in the 4-dimensional Minkowski space that change the direction but preserve the length of 4-vectors. As we will now demonstrate, such rotations in Minkowski space are just the Lorentz transformations alluded to above.

9.4.1 Rotations in Euclidean Space

First we consider a rotation of the coordinate system in euclidean space, as illustrated in Fig. 8.4. The condition that the length of an arbitrary vector be unchanged by this transformation corresponds to the requirement that the transformation matrix R implementing the rotation [see Eq. (8.45)] act on the metric tensor g_{ij} in the following way

$$R g_{ij} R^{\mathrm{T}} = g_{ij}, \tag{9.18}$$

where R^{T} denotes the transpose of R. For euclidean space the metric tensor is just the unit matrix so the requirement (9.18) reduces to $RR^{\mathrm{T}} = 1$, which is the condition that R

be an orthogonal matrix. Thus, we have obtained the well-known result that rotations in euclidean space are implemented by orthogonal matrices in a rather pedantic manner. But Eq. (9.18) is valid generally, not just for euclidean spaces. Therefore, it may be used as guidance for constructing more general rotations in Minkowski space.

9.4.2 Generalized 4D Minkowski Rotations

By analogy with the above discussion of rotations in euclidean space, which left the length of 3-vectors invariant, let us now seek a set of transformations that leave the length of a 4-vector invariant in the Minkowski space. The coordinate transformation may be written in matrix form,

$$dx'^{\mu} = \Lambda^{\mu}{}_{\nu} dx^{\nu}, \tag{9.19}$$

where the transformation matrix $\Lambda^{\mu}{}_{\nu}$ is expected to satisfy the analog of Eq. (9.18) for the Minkowski metric $\eta_{\mu\nu}$,

$$\Lambda \eta_{\mu\nu} \Lambda^{\mathrm{T}} = \eta_{\mu\nu}, \tag{9.20}$$

or explicitly in terms of components, $\Lambda_{\mu}{}^{\rho} \Lambda^{\sigma}{}_{\nu} \eta_{\rho\sigma} = \eta_{\mu\nu}$.[3] This property may now be used to construct the elements of the transformation matrix $\Lambda^{\mu}{}_{\nu}$. The possible transformations include rotations about the spatial axes (corresponding to rotations within inertial systems) and transformations between inertial systems moving at different constant velocities that are termed Lorentz boosts.

Two inertial frames may differ in displacement, rotational orientation, and uniform velocity. This corresponds to 10 possible transformations between inertial frames: three velocity boosts along the spatial axes, three rotations about the three spatial axes, and four translations in the space and time directions. These 10 transformations form a group called the *Poincaré group* (see Box 9.1). The six Lorentz transformations correspond to the velocity boosts and spatial rotations, and they form a group called the *Lorentz group* that is a subgroup of the Poincaré group, also discussed in Box 9.1. Consider first the simple case of rotations about the z axis.

9.4.3 Lorentz Spatial Rotations

Rotations about a single spatial axis in Minkowski space correspond to a 2-dimensional problem with euclidean metric, so the condition (9.18) may be written as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{9.21}$$

[3] Note that in this discussion we are using the (common) convention that $\eta_{\mu\nu}$ is either a symbol standing for the full tensor or a specific component of the tensor distinguished by indices μ and ν, depending on context. In a matrix equation like (9.20) the order of the factors matters because matrices don't generally commute, but when the matrix equation is written out in terms of sums over component products as in $\Lambda_{\mu}{}^{\rho} \Lambda^{\sigma}{}_{\nu} \eta_{\rho\sigma} = \eta_{\mu\nu}$ the order of factors can be rearranged at will, since the components of the matrices are just numbers that commute with each other. The matrices Λ are symmetric so horizontal placement of indices isn't crucial, but a typical convention is to define $\Lambda^{\mu}{}_{\nu} = \partial x'^{\mu}/\partial x^{\nu}$ and $\Lambda_{\mu}{}^{\nu} = \partial x^{\nu}/\partial x'^{\mu}$ (compare Table 8.1).

Box 9.1 Symmetries and groups

A group is a set $G = \{x, y, \ldots\}$ for which a binary operation $a \cdot b = c$ called *group multiplication* is defined that has the following properties

(i) *Closure:* If x and y are elements of G, then $x \cdot y$ is an element of G also.
(ii) *Identity:* An identity element e exists such that $e \cdot x = x \cdot e = x$ for $x \in G$.
(iii) *Existence of an Inverse:* For every group element x there is an inverse x^{-1} in the set such that $xx^{-1} = e$.
(iv) *Associativity:* Multiplication is associative: $(x \cdot y) \cdot z = x \cdot (y \cdot z)$ for $x, y, z \in G$.

For groups of transformations multiplication corresponds to applying first one and then the other transformation. The group definition requires associativity but not commutivity. A group consisting of commutative elements only is *abelian*; otherwise, it is *nonabelian*.

Example: The Lorentz group

There are six independent Lorentz transformations: three rotations about the spatial axes parameterized by real angles, and three boosts along the spatial axes parameterized by boost velocities. Because rotation angles and boost velocities can take continuous real values, the set of Lorentz transformations is infinite. The Lorentz transformations form a group:

1. Two successive transformations are equivalent to some other transformation.
2. Every Lorentz transformation has an inverse that is the transformation in the opposite direction (for example, $v \to -v$ for boosts).
3. The identity corresponds to no transformation.
4. It doesn't matter how three successive transformations are grouped, so multiplication is associative, but the *order* matters, so the Lorentz group is nonabelian.

An important class of groups corresponds to transformations that are continuous (analytical) in their parameters. These are called *Lie groups.* The Lorentz group is a Lie group. Groups also can be classified according to whether their parameter spaces are closed and bounded (*compact groups*) or not (*noncompact groups*). Because boost velocities can approach c asymptotically but never reach it, the Lorentz-group parameter space is not closed (the limit $v = c$ is not part of the set) and the group is noncompact.

Example: The Poincaré group

The Poincaré group is formed by adding to the Lorentz transformations the uniform translations along the four spacetime axes. It is a non-abelian, noncompact Lie group, and contains the Lorentz group as a *subgroup* (a subset of group elements that satisfy the same group postulates as the parent group).

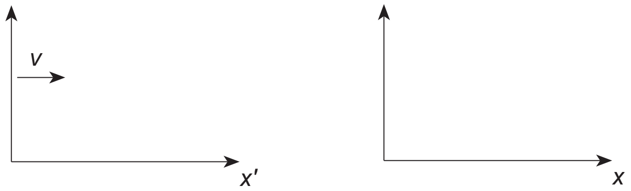


Fig. 9.1 A Lorentz boost along the positive x axis by a velocity v.

where a, b, c, and d parameterize the transformation matrices. Carrying out the matrix multiplications explicitly on the left side gives

$$\begin{pmatrix} a^2+b^2 & ac+bd \\ ac+bd & c^2+d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and comparison of the two sides of the equation implies the conditions

$$a^2+b^2=1 \qquad c^2+d^2=1 \qquad ac+bd=0.$$

Obviously one choice of parameters that satisfies these conditions is

$$a=\cos\phi \qquad b=\sin\phi \qquad c=-\sin\phi \qquad d=\cos\phi.$$

This leads to the standard result

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = R\begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}\begin{pmatrix} x^1 \\ x^2 \end{pmatrix}, \tag{9.22}$$

which is Eq. (8.44) restricted to rotations about a single axis. Now we shall apply this same technique to determine the elements of a Lorentz boost transformation.

9.4.4 Lorentz Boost Transformations

Consider a boost from one inertial system to a second one moving in the positive direction at uniform velocity along the x axis, as illustrated in Fig. 9.1. Since the y and z coordinates will not be affected, this is a 2-dimensional problem in the time coordinate t and the spatial coordinate x. The transformation is of the general form

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} cdt \\ dx \end{pmatrix}, \tag{9.23}$$

and the condition (9.20) can be written out explicitly as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{9.24}$$

which is identical in form to the rotation case, except for the indefinite metric. Multiplying the matrices on the left side and comparing with the right side gives the conditions

$$a^2-b^2=1 \qquad -c^2+d^2=1 \qquad -ac+bd=0,$$

Box 9.2 Minkowski rotations

The respective derivations make clear that the appearance of hyperbolic functions in Eq. (9.25) instead of trigonometric functions as in Eq. (9.22) traces to the role of the indefinite metric diag$(-1,1)$ in Eq. (9.24) relative to the positive-definite metric diag$(1,1)$ in Eq. (9.21). The boost transformations are in a sense "rotations" in Minkowski space, but these rotations have unusual properties relative to normal rotations in euclidean space since they mix space and time, and may be viewed as rotations through imaginary angles. These properties follow from the metric because the invariant interval is neither the length of vectors in space nor the length of time intervals, but rather the specific mixture of space and time intervals implied by the Minkowski line element (9.5) with indefinite metric (9.6). This is quite different from Newtonian mechanics, where time is a universal parameter common to all observers and the Galilean transformations conserve only the *space interval.*

which clearly are satisfied by the parameterization

$$a = \cosh\xi \qquad b = \sinh\xi \qquad c = \sinh\xi \qquad d = \cosh\xi,$$

where ξ is a hyperbolic variable taking the values $-\infty \le \xi \le \infty$. Therefore, we may write the boost transformation as

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix} \begin{pmatrix} cdt \\ dx \end{pmatrix}. \tag{9.25}$$

A geometrical interpretation of this result is discussed in Box 9.2.

The Lorentz boost transformation (9.25) can be put into a more familiar form by exchanging the boost parameter ξ for the boost velocity. For convenience, let's replace the differentials in Eq. (9.25) with finite space and time intervals ($dt \to t$ and $dx \to x$). The velocity of the boosted system is $v = x/t$. From Eq. (9.25), the origin ($x' = 0$) of the boosted system is given by

$$x' = ct\sinh\xi + x\cosh\xi = 0.$$

Therefore, $x/t = -c\sinh\xi/\cosh\xi$, from which it may be concluded that

$$\beta \equiv \frac{v}{c} = \frac{x}{ct} = -\frac{\sinh\xi}{\cosh\xi} = -\tanh\xi. \tag{9.26}$$

This relationship between ξ and β is plotted in Fig. 9.2. Utilizing the identity $1 = \cosh^2\xi - \sinh^2\xi$ and the definition

$$\gamma \equiv \left(1 - \frac{v^2}{c^2}\right)^{-1/2} \tag{9.27}$$

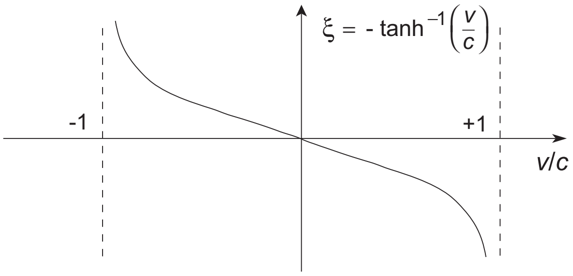


Fig. 9.2 Dependence of the Lorentz parameter ξ on $\beta = v/c$.

of the *Lorentz γ-factor,* we may write

$$\cosh\xi = \sqrt{\frac{\cosh^2\xi}{1}} = \frac{1}{\sqrt{1-\sinh^2\xi/\cosh^2\xi}} = \frac{1}{\sqrt{1-v^2/c^2}} = \gamma, \tag{9.28}$$

and from this result and (9.26),

$$\sinh\xi = -\beta\cosh\xi = -\beta\gamma. \tag{9.29}$$

Inserting (9.28)–(9.29) into (9.25) for finite intervals gives

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} \tag{9.30}$$

and writing this matrix expression out explicitly gives the Lorentz boost equations (for the specific case of a positive boost along the x axis) in standard textbook form,

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) \qquad x' = \gamma(x - vt) \qquad y' = y \qquad z' = z, \tag{9.31}$$

with the inverse transformation corresponding to the replacement $v \to -v$. By inspection, these reduce to the Galilean boost equations (8.46) if $v/c \to 0$ and it is easy to verify that the Lorentz transformations leave invariant the spacetime interval ds^2.

9.5 Lightcone Diagrams

The line element (9.5) defines a cone, implying that Minkowski spacetime may be classified according to the *lightcone diagram* exhibited in Fig. 9.3. The *lightcone* is a 3-dimensional surface in the 4-dimensional spacetime and intervals relative to the origin may be characterized according to whether they are inside of, outside of, or on the lightcone. Assuming the metric signature $(-+++)$ employed here, the standard terminology is

- If $ds^2 < 0$ the interval is termed *timelike*.
- If $ds^2 > 0$ the interval is termed *spacelike*.
- If $ds^2 = 0$ the interval is called *lightlike* (or *null*).

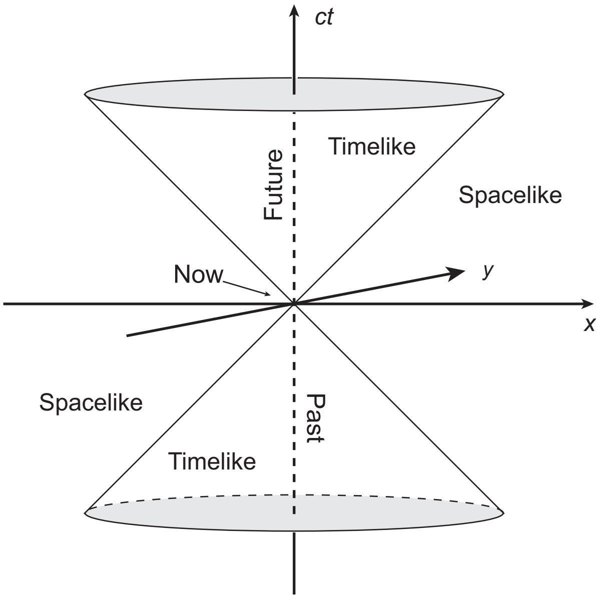


Fig. 9.3 Lightcone diagram for flat spacetime in two space and one time dimensions. The future lightcone is the surface swept out by a spherical light pulse emitted from the origin.

These regions for flat spacetime are labeled in Fig. 9.3. This classification can be extended also to surfaces. For example, a *spacelike surface* is a collection of points for which any pair of points has a spacelike separation and a *lightlike surface* is a collection of points for which any pair of points has a lightlike separation. The lightcone classification makes clear the distinction between Minkowski spacetime and a mere 4-dimensional euclidean space in that two points in the Minkowski spacetime may be separated by a distance that when squared could be positive, negative, or zero. This is not possible in a euclidean metric. Notice in particular that for lightlike particles, which have worldlines confined to the lightcone, the square of the spacetime interval between any two points on a worldline is *zero.*

Example 9.3 The Minkowski line element (9.5) in one space and one time dimension [which often is termed $(1+1)$ dimensions] is given by $ds^2 = -c^2dt^2 + dx^2$. Thus, if the spacetime interval is lightlike, $ds^2 = 0$ and

$$-c^2dt^2 + dx^2 = 0 \quad \longrightarrow \quad \left(\frac{dx}{dt}\right)^2 = c^2 \quad \longrightarrow \quad v = \pm c.$$

This result can be generalized easily to the full space, leading to the conclusion that events in Minkowski space separated by a null interval ($ds^2 = 0$) are connected by signals moving at light velocity, $v = c$. If the time axis (ct) and space axes have the same scales, this means that the worldline of a freely-propagating photon (or any massless particle, necessarily moving at light velocity) always makes ±45° angles in the lightcone diagram. By similar

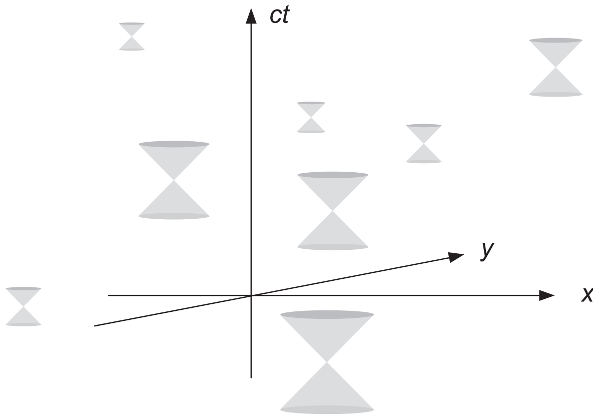


Fig. 9.4 Lightcones are local concepts. Each point of spacetime should be imagined to have its own lightcone.

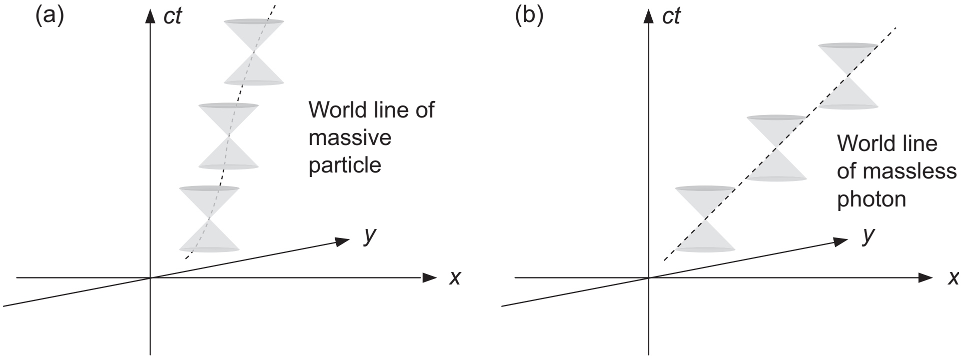


Fig. 9.5 Worldlines for (a) massive (timelike) particles and (b) massless (lightlike) particles. For massive particles the trajectory must always lie inside the local light cone at each point; for massless particles it must always lie on the lightcone at each point.

arguments, events at timelike separations (inside the lightcone) are connected by signals with $v < c$, and those with spacelike separations (outside the lightcone) could be connected only by causality-violating signals with $v > c$.

The lightcone illustrated in Fig. 9.3 was placed at the origin for illustration but each point in the spacetime has its own lightcone, as illustrated in Fig. 9.4. From Example 9.3, the tangent to the worldline of any particle at a point defines the local velocity of the particle at that point and constant velocity implies straight worldlines. Therefore, as illustrated in Fig. 9.5(b), light travels in a straight line in flat spacetime and always on the lightcone because it has constant local velocity, $v = c$, while the worldline for any massive particle must lie inside the local lightcone because it must always have $v \leq c$ (in the jargon, a worldline for a massive particle is always *timelike*). The worldline for the massive particle in Fig. 9.5(a) is curved, indicating an acceleration since the velocity is changing with time. For non-accelerated massive particles the worldline would be straight, but always within the local lightcone.

In the Galilean relativity of Newtonian mechanics an event picks out a hyperplane of simultaneity in the spacetime diagram consisting of all events occurring at the same time as the event. All observers agree on what constitutes this set of simultaneous events because in Galilean relativity simultaneity is independent of the observer. In Einstein's relativity, *simultaneity depends on the observer* and hyperplanes of constant coordinate time have no invariant meaning. However, *all observers agree on the lightcones associated with events,* because the speed of light is an invariant for all observers. Thus, the *lightcones define an invariant spacetime structure* permitting unambiguous causal classification of events.

9.6 The Causal Structure of Spacetime

The causal properties of Minkowski spacetime are encoded in its invariant lightcone structure. Each spacetime point lies at the apex of its lightcone ("Now" in Fig. 9.3), as illustrated in Fig. 9.4. From Example 9.3, the lightcone defines a set of points that are connected to the origin by signals moving at light velocity c, events inside the lightcone may be connected to the origin by signals moving at $v < c$, and events outside the lightcone may be connected to the origin only by signals having $v > c$. Thus, the event at the origin of a local lightcone may influence any event within its forward lightcone ("Future" in Fig. 9.3) through signals propagating at $v < c$. Likewise, the event at the origin may be influenced by events within its backward lightcone ("Past" in Fig. 9.3) through signals with speeds less than that of light. Conversely, events at spacelike separations may not be influenced, or have an influence on, the event at the origin except by signals that require $v > c$. Finally, events on the lightcone are connected by signals that travel exactly at c. Thus, events at spacelike separations are causally disconnected and the lightcone is a surface separating the knowable from the unknowable for an observer at the apex of the lightcone.

This lightcone structure of spacetime ensures that all velocities obey locally the constraint $v \leq c$. Since velocities are defined and measured locally, covariant field theories in either flat or curved spacetime are guaranteed to respect the speed limit $v \leq c$, irrespective of whether velocities appear to exceed c globally. For example, in the Hubble expansion of the Universe galaxies beyond a certain distance (the horizon) appear to recede at velocities in excess of c because of the expansion of space, and light coming to us from near the horizon is stretched in wavelength and takes longer to propagate to us than it would in a flat, non-expanding spacetime. However, all local measurements in that expanding, possibly curved spacetime would determine the velocity of light to be c, in accordance with the axioms of special relativity and the associated lightcone structure of spacetime.

Example 9.4 Time machines are of enduring interest in science fiction and in the public imagination. The local lightcones of Minkowski spacetime embody the causal structure of

Box 9.3 Time Machines and Causality Paradoxes

When time travel comes up it is usually about going *backward* in time. Traveling *forward* in time requires no special talent and it is easy to arrange various scenarios consistent with relativity where a person could travel into a future time even faster than normal (at least as thought experiments; I leave procurement and engineering details to you!) {har03} [13]. For example, in the twin paradox it is possible to arrange for the traveler (whether twin or not) to arrive back at Earth centuries in the future relative to clocks that remain on Earth—a kind of time travel. Similar options exist using the gravitational time dilation in strong gravity. (You'll need a black hole and an unlimited fuel budget, or a planet-size batch of incredibly strong and thin material; but again I leave procurement and engineering to you!) However, the real question is, could you go *back in time* to explore your earlier history?

No! Not according to current understanding. To bend a forward-going timelike worldline continuously into a backward-going one requires going outside the local lightcone, requiring that $v > c$. If *closed timelike loops* were permitted, travel to earlier times might be possible. However, they are forbidden if there are no negative energy densities and the Universe has the topology in evidence. Thus, the determined time traveler has two options: find some negative energy, or find structures with an exotic spacetime topology allowing closed timelike loops. However, negative energy is probably forbidden in classical gravity (it is unclear whether quantum mechanics could provide any loopholes), and there is no evidence at present for exotic spacetime topologies with closed timelike loops. Hence, I would advise against taking a strong investment position in (if you will) time-travel futures! These statements are based entirely on classical gravity considerations; it is unknown at present whether they could be modified by some future understanding of quantum gravity.

special relativity, and it will be seen later that general relativity inherits the local lightcone structure of Minkowski space. Therefore, as explored further in Box 9.3, lightcone diagrams provide a simple way to answer the question of whether special or general relativity allow going back in time to prevent your own birth, which in turns prevents you from going back in time to prevent your own birth, ...

From the preceding discussion we may conclude that the axioms of special relativity are fundamentally at odds with the Newtonian concept of absolute simultaneity, since the demand that light have the same speed for all observers necessarily means that the apparent temporal order of two events depends upon the observer. However, the abolishment of absolute simultaneity introduces *no causal ambiguity* because all observers agree on the lightcone structure of spacetime and hence all observers will agree that event A can cause event B only if A lies in the past lightcone of B, for example.

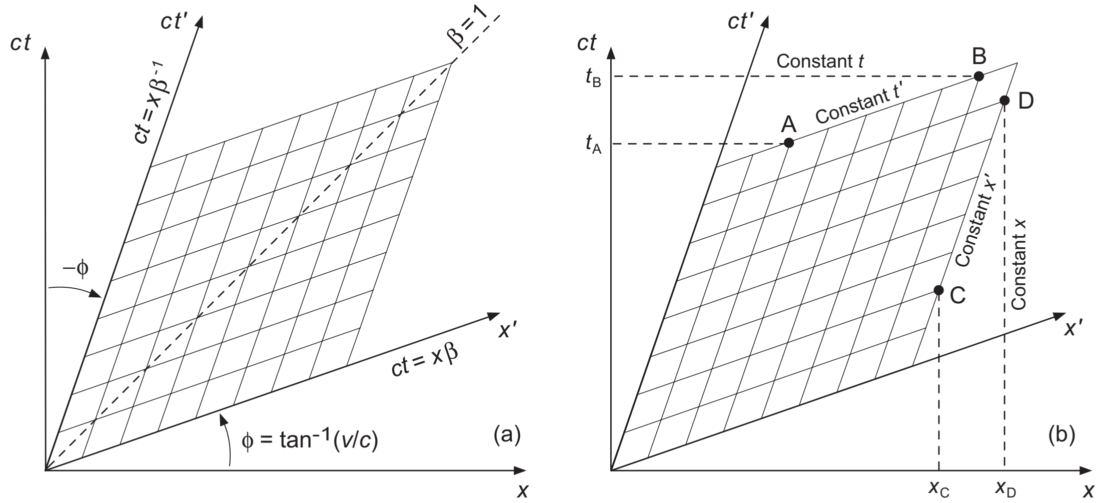


Fig. 9.6 (a) Lorentz boost transformation in a spacetime diagram. (b) Comparison of events in boosted and unboosted reference frames.

9.7 Lorentz Transformations in Spacetime Diagrams

It is instructive to examine the action of Lorentz transformations in the spacetime (lightcone) diagram. If consideration is restricted to boosts only in the x direction, the relevant part of the spacetime diagram in some inertial frame corresponds to a plot with axes ct and x, as illustrated in Fig. 9.6.

9.7.1 Lorentz Boosts and the Lightcone

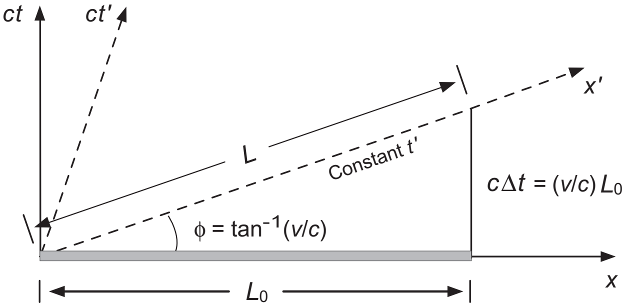
What happens to the axes in Fig. 9.6 under a Lorentz boost? From the first two of Eqs. (9.31),

$$ct' = c\gamma\left(t - \frac{vx}{c^2}\right) \qquad x' = \gamma(x - vt). \tag{9.32}$$

The t' axis corresponds to $x' = 0$, which implies from the second of Eqs. (9.32) that $ct = x\beta^{-1}$, with $\beta = v/c$, is the equation of the t' axis in the (ct, x) coordinate system. Likewise, the x' axis corresponds to $t' = 0$, which implies from the first of Eqs. (9.32) that $ct = x\beta$. The x' and ct' axes for the boosted system are also shown in Fig. 9.6(a) for a boost corresponding to a positive value of β. The time and space axes are rotated by the same angle, but in *opposite directions* by the boost (a consequence of the indefinite Minkowski metric). The angle of rotation is related to the boost velocity through $\tan\phi = v/c$.

Many characteristic features of special relativity are apparent from Fig. 9.6. For example, relativity of simultaneity follows directly, as illustrated in Fig. 9.6(b). Points A and B lie on the same t' line, so they are simultaneous in the boosted frame. But from the dashed projections on the ct axis, in the unboosted frame event A occurs before event B. Likewise, points C and D lie at the same value of x' in the boosted frame and so are spatially congruent, but in the unboosted frame $x_C \neq x_D$. Relativistic time dilation and space contraction effects follow rather directly from these observations, as illustrated in the example below.

Example 9.5 Consider the following schematic representation of the spacetime diagram illustrated in Fig. 9.6, where a rod of length L_0, as measured in its own rest frame (t,x), is oriented along the x axis.



The frame (t',x') is boosted by a velocity v along the $+x$ axis relative to the (t,x) frame. Therefore, in the primed frame the rod will have a velocity v in the negative x' direction. Determining the length L observed in the primed frame requires that the positions of the ends of the rod be measured *simultaneously in that frame.* The axis labeled x' corresponds to constant t' [see Fig. 9.6(b)], so the distance marked as L is the length in the primed frame. This distance *seems* longer than L_0, but this is deceiving because a slice of Minkowski spacetime is being represented on a piece of euclidean paper. Much as a Mercator projection of the globe onto a euclidean sheet of paper gives misleading distance information (Greenland isn't really larger than Brazil), the metric must be trusted to determine the correct distance in a space. From the Minkowski indefinite metric and the triangle in the figure above, $L^2 = L_0^2 - (c\Delta t)^2$. But from Eq. (9.32) it was found that the equation for the x' axis is $c\Delta t = (v/c)L_0$, from which it follows that

$$L = (L_0^2 - (c\Delta t)^2)^{1/2} = \left(L_0^2 - \left(\frac{v}{c}L_0\right)^2\right)^{1/2} = L_0(1 - v^2/c^2)^{1/2},$$

which is the familiar length-contraction formula of special relativity: L is *shorter* than L_0, even though it appears to be longer in the figure above.

9.7.2 Spacelike and Timelike Intervals

As noted previously, the spacetime interval between any two events may be classified in a relativistically invariant way as timelike, lightlike, or spacelike by constructing the lightcone at one of the points, as illustrated in Fig. 9.7(a). This geometry and that of Fig. 9.6

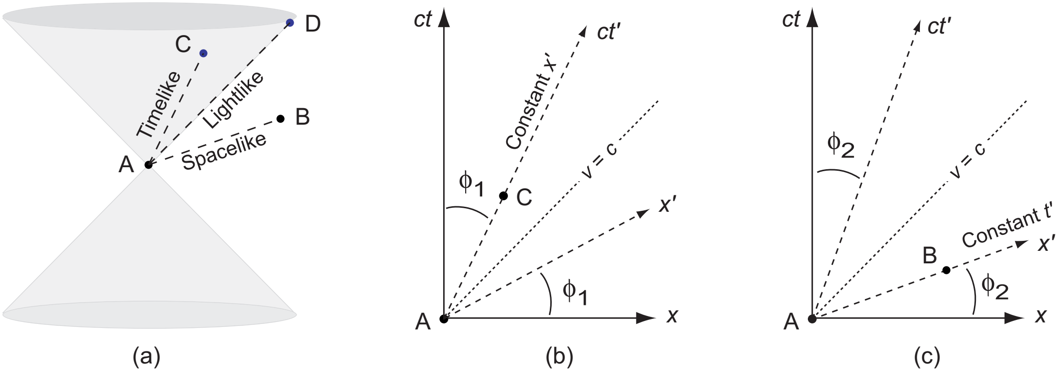


Fig. 9.7 (a) Timelike, lightlike (null), and spacelike separations for events. (b) A Lorentz transformation that brings the timelike separated points A and C of (a) into spatial congruence (they lie along a line of constant x' in the primed coordinate system). (c) A Lorentz transformation that brings the spacelike separated points A and B of (a) into coincidence in time (they lie along a line of constant t' in the primed coordinate system).

then suggest another important distinction between events at spacelike separations [the line AB in Fig. 9.7(a)] and timelike separations [the line AC in Fig. 9.7(a)]:

(i) If two events have a timelike separation, a Lorentz transformation exists that can bring them into spatial congruence. Figure 9.7(b) illustrates geometrically a coordinate system (ct', x'), related to the original system by an x-axis Lorentz boost of $v/c = \tan\phi_1$, in which A and C have the same coordinate x'.

(ii) If two events have a spacelike separation, a Lorentz transformation exists that can synchronize the events. Figure 9.7(c) illustrates an x-axis Lorentz boost by $v/c = \tan\phi_2$ to a system in which A and B have the same time t'.

Maximum values of ϕ_1 and ϕ_2 are limited by the $v = c$ line. The Lorentz transformation to bring A into spatial congruence with C exists only if C lies to the left of $v = c$ (C separated by a timelike interval from A). Likewise, the Lorentz transformation to synchronize A with B exists only if B lies to the right of $v = c$ (B separated by a spacelike interval from A).

9.8 Lorentz Invariance of Maxwell's Equations

We conclude this chapter by examining the Lorentz invariance of the Maxwell equations that describe classical electromagnetism. There are several motivations. First, it provides a nice example of how useful Lorentz invariance and Lorentz tensors can be. Second, the properties of the Maxwell equations influenced Einstein strongly in his development of the special theory of relativity. Finally, there are many useful parallels between general relativity and the Maxwell theory, particularly for weak gravity where the Einstein field equations may be linearized.

9.8.1 Maxwell Equations in Non-Covariant Form

In free space, using Heaviside–Lorentz, $c = 1$ units (see Appendix B.3), the Maxwell equations may be written as

$$\nabla \cdot \boldsymbol{E} = \rho \tag{9.33}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} + \nabla \times \boldsymbol{E} = 0 \tag{9.34}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{9.35}$$

$$\nabla \times \boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{j}, \tag{9.36}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, ρ is the charge density, and $\boldsymbol{j}$ is the current vector, with the density and current required to satisfy the equation of continuity

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{j} = 0. \tag{9.37}$$

The Maxwell equations are consistent with the special theory of relativity. However, in the form (9.33)–(9.37) this covariance is *not manifest,* since these equations are formulated in terms of 3-vectors and separate derivatives with respect to space and time, not in terms of Minkowski tensors. It is useful in a number of contexts to reformulate the Maxwell equations in a manner that is manifestly covariant with respect to Lorentz transformations. The usual route to accomplishing this begins by replacing the electric and magnetic fields by new variables.

9.8.2 Scalar and Vector Potentials

The electric and magnetic fields appearing in the Maxwell equations may be eliminated in favor of a vector potential $\boldsymbol{A}$ and a scalar potential ϕ through the definitions

$$\boldsymbol{B} \equiv \nabla \times \boldsymbol{A} \qquad \boldsymbol{E} \equiv -\nabla \phi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{9.38}$$

The vector identities

$$\nabla \cdot (\nabla \times \boldsymbol{B}) = 0 \qquad \nabla \times \nabla \phi = 0, \tag{9.39}$$

may then be used to show that the second and third Maxwell equations are satisfied identically, and the identity

$$\nabla \times (\nabla \times \boldsymbol{A}) = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}, \tag{9.40}$$

may be used to write the remaining two Maxwell equations as the coupled second-order equations

$$\nabla^2 \phi + \frac{\partial}{\partial t} \nabla \cdot \boldsymbol{A} = -\rho \tag{9.41}$$

$$\nabla^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \nabla \left(\nabla \cdot \boldsymbol{A} + \frac{\partial \phi}{\partial t} \right) = -\boldsymbol{j}. \tag{9.42}$$

These equations may then be decoupled by exploiting a fundamental symmetry of electromagnetism termed *gauge invariance*.

9.8.3 Gauge Transformations

Because of the identity $\boldsymbol{\nabla}\times\boldsymbol{\nabla}\phi = 0$, the simultaneous transformations

$$\boldsymbol{A} \to \boldsymbol{A}+\boldsymbol{\nabla}\chi \qquad \phi \to \phi - \frac{\partial\chi}{\partial t} \tag{9.43}$$

for an arbitrary scalar function χ do not change the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields; thus, they leave the Maxwell equations invariant. The transformations (9.43) are termed (classical) *gauge transformations.* This freedom of gauge transformation may be used to decouple Eqs. (9.41)–(9.42). For example, if a set of potentials $(\boldsymbol{A},\phi)$ that satisfy

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}+\frac{\partial\phi}{\partial t}=0, \tag{9.44}$$

is chosen, the equations decouple to yield

$$\boldsymbol{\nabla}^2\phi - \frac{\partial^2\phi}{\partial t^2} = -\rho \qquad \boldsymbol{\nabla}^2\boldsymbol{A} - \frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\boldsymbol{j}, \tag{9.45}$$

which may be solved independently for $\boldsymbol{A}$ and ϕ.

A constraint of the sort (9.44) is termed a *gauge condition* and the imposition of such a constraint is termed *fixing the gauge.* This particular choice of gauge that leads to the decoupled equations (9.45) is termed the *Lorentz gauge*. Another common gauge is the *Coulomb gauge*, with a gauge-fixing condition

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}=0, \tag{9.46}$$

which leads to the equations

$$\boldsymbol{\nabla}^2\phi = -\rho \qquad \boldsymbol{\nabla}^2\boldsymbol{A} - \frac{\partial^2\boldsymbol{A}}{\partial^2 t} = \boldsymbol{\nabla}\frac{\partial\phi}{\partial t} - \boldsymbol{j}. \tag{9.47}$$

Let us utilize the shorthand notation for derivatives introduced in Eq. (8.89):

$$\begin{aligned} \partial^\mu &\equiv \frac{\partial}{\partial x_\mu} = (\partial^0,\partial^1,\partial^2,\partial^3) = \left(-\frac{\partial}{\partial x^0},\boldsymbol{\nabla}\right), \\ \partial_\mu &\equiv \frac{\partial}{\partial x^\mu} = (\partial_0,\partial_1,\partial_2,\partial_3) = \left(\frac{\partial}{\partial x^0},\boldsymbol{\nabla}\right), \end{aligned} \tag{9.48}$$

where, for example, $\partial^1 = \partial/\partial x_1$ and

$$\boldsymbol{\nabla} \equiv (\partial^1,\partial^2,\partial^3) \tag{9.49}$$

is the 3-divergence. A compact and covariant formalism then results from introducing the 4-vector potential A^μ, the 4-current j^μ, and the d'Alembertian operator $\Box$ through

$$A^\mu \equiv (\phi,\boldsymbol{A}) = (A^0,\boldsymbol{A}) \qquad j^\mu \equiv (\rho,\boldsymbol{j}) \qquad \Box \equiv \partial_\mu\partial^\mu. \tag{9.50}$$

Then a gauge transformation takes the form

$$A^\mu \to A^\mu - \partial^\mu\chi \equiv A'^\mu \tag{9.51}$$

and the preceding examples of gauge-fixing constraints become[4]

$$\partial_\mu A^\mu = 0 \quad \text{(Lorentz gauge)} \qquad \boldsymbol{\nabla}\cdot\boldsymbol{A} = 0 \quad \text{(Coulomb gauge)}. \tag{9.52}$$

The operator $\Box$ is Lorentz invariant since

$$\Box' = \partial'_\mu \partial'^\mu = \Lambda_\mu{}^\nu \Lambda^\mu{}_\lambda \partial_\nu \partial^\lambda = \partial_\mu \partial^\mu = \Box.$$

Thus, the Lorentz-gauge wave equation may be expressed in the manifestly covariant form

$$\Box A^\mu = j^\mu \tag{9.53}$$

and the continuity equation (9.37) becomes

$$\partial_\mu j^\mu = 0. \tag{9.54}$$

The covariance of the Maxwell wave equation (9.53) in the Lorentz gauge, coupled with the gauge invariance of electromagnetism, ensures that the Maxwell equations are covariant in all gauges. However, as was seen in the example of the Coulomb gauge, the covariance may not be manifest for a particular choice of gauge.

9.8.4 Maxwell Equations in Manifestly Covariant Form

The Maxwell equations may be cast in a form that is manifestly covariant by appealing to Eqs. (9.38) to construct the components of the electric and magnetic fields in terms of the potentials. Proceeding in this manner, we find that the six independent components of the 3-vectors $\boldsymbol{E}$ and $\boldsymbol{B}$ are elements of an antisymmetric rank-2 *electromagnetic field tensor*

$$F^{\mu\nu} = -F^{\nu\mu} = \partial^\mu A^\nu - \partial^\nu A^\mu, \tag{9.55}$$

which may be expressed in matrix form as

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & -B^3 & B^2 \\ E^2 & B^3 & 0 & -B^1 \\ E^3 & -B^2 & B^1 & 0 \end{pmatrix}. \tag{9.56}$$

That is, the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$ are *vectors* in three-dimensional euclidean space but in Minkowski space their six components together form an antisymmetric *rank-2 tensor.* Now let us employ the Levi–Civita symbol $\varepsilon_{\alpha\beta\gamma\delta}$, which has the value $+1$ for $\alpha\beta\gamma\delta = 0123$ and cyclic permutations, -1 for odd permutations, and zero if any two indices are equal, and which further satisfies $\varepsilon_{\alpha\beta\gamma\delta} = -\varepsilon^{\alpha\beta\gamma\delta}$. If the *dual field tensor* $\mathscr{F}^{\mu\nu}$ is then defined by

$$\mathscr{F}^{\mu\nu} = \tfrac{1}{2}\varepsilon^{\mu\nu\gamma\delta}F_{\gamma\delta} = \begin{pmatrix} 0 & -B^1 & -B^2 & -B^3 \\ B^1 & 0 & E^3 & -E^2 \\ B^2 & -E^3 & 0 & E^1 \\ B^3 & E^2 & -E^1 & 0 \end{pmatrix}, \tag{9.57}$$

[4] This notation shows explicitly that the Lorentz condition is a covariant constraint because it is formulated in terms of 4-vectors; however, the Coulomb gauge condition is not covariant because it is formulated in terms of only three of the components of a 4-vector.

the Maxwell equations (9.33) and (9.36) may be written

$$\partial_\mu F^{\mu\nu} = j^\nu, \tag{9.58}$$

and the Maxwell equations (9.34) and (9.35) may be written as

$$\partial_\mu \mathscr{F}^{\mu\nu} = 0. \tag{9.59}$$

The Maxwell equations in this form are manifestly covariant because they are formulated exclusively in terms of Lorentz tensors.

Problems

9.1 The primed axes (x', t') plotted in the coordinate system of the unprimed axes in Fig. 9.6 do not appear to be orthogonal, but they must be since the unprimed axes are orthogonal (their scalar product vanishes) and the scalar product is preserved under Lorentz transformations. Show generally that two vectors in a spacetime diagram are orthogonal if each makes the same angle with respect to the lightcone. Use this result to show that a lightlike vector is necessarily orthogonal to itself.

9.2 If a Lorentz transformation is denoted by $\Lambda^{\mu}{}_{\nu}$, prove that the Lorentz transformation given by $\Lambda_{\mu}{}^{\nu} = \eta_{\mu\alpha}\eta^{\nu\beta}\Lambda^{\alpha}{}_{\beta}$ is its inverse.

9.3 (a) Prove that the Maxwell field tensor $F^{\mu\nu}$ is invariant under a gauge transformation of the 4-vector potential A^{μ}. (b) By appealing to the definitions (9.38), show that the Maxwell field tensor $F^{\mu\nu}$ has components given by Eqs. (9.55) and (9.56).

9.4 Show that the Maxwell equations written in the covariant form (9.58) and (9.59) are equivalent to the non-covariant form of the Maxwell equations given by Eqs. (9.33)–(9.36).

10 Electromagnetism as a Gauge Theory

A Appendix A **Mathematics Review**

This Appendix reviews mathematics required in this book. Readers are assumed to have some prior acquaintance with this and mostly we list equations and concepts without proof. A book such as Griffiths [8] may be consulted for proofs and more detail.

A.1 Vectors and other Tensors

Every physics student learns at her mother's knee that vectors have magnitude and direction, so they are specified by more than one number, and that this is indicated graphically by an arrow, with length indicating magnitude and orientation indicating direction. This view of vectors as directed line segments works in introductory physics but it can lead to erroneous views. For example, the directed line segment invites one to think of vectors as connecting two points in a manifold. This is wrong: a vector is *defined at a single point*; it does *not* connect two points. We can often get away with this sloppy thinking for flat manifolds, but how is one to interpret an extended straight line segment in a curved manifold? A more rigorous definition is required for advanced applications, particularly if the space is curved and/or described by non-cartesian coordinates.

At a somewhat more rigorous level, vectors are a special case of *tensors*, which we may think of (with the naive pragmatism of physicists unconstrained by rigorous mathematical training) as objects that obey particular transformation laws under a change of coordinate system, and that require n indices when expanded in a basis, with n termed the *rank* of the tensor. General transformation laws for 4D spacetime tensors with $n < 3$ are

$$\phi'(x') = \phi(x) \qquad \text{(scalar)}, \tag{A.1a}$$

$$V'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} V^{\nu}(x) \qquad \text{(vector)}, \tag{A.1b}$$

$$V'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} V_{\nu}(x) \qquad \text{(dual vector)}, \tag{A.1c}$$

$$T'_{\mu\nu}(x') = \frac{\partial x^{\alpha}}{\partial x'^{\mu}} \frac{\partial x^{\beta}}{\partial x'^{\nu}} T_{\alpha\beta}(x) \qquad \text{(covariant rank-2 tensor)}, \tag{A.1d}$$

$$T'^{\nu}{}_{\mu}(x') = \frac{\partial x^{\alpha}}{\partial x'^{\mu}} \frac{\partial x'^{\nu}}{\partial x^{\beta}} T^{\beta}{}_{\alpha}(x) \qquad \text{(mixed rank-2 tensor)}, \tag{A.1e}$$

$$T'^{\mu\nu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\alpha}} \frac{\partial x'^{\nu}}{\partial x^{\beta}} T^{\alpha\beta}(x) \qquad \text{(contravariant rank-2 tensor)}. \tag{A.1f}$$

In these equations, unprimed coordinates refer to the original coordinate system, primed coordinates refer to the transformed coordinate system, and all partial derivatives in the general case depend on the spacetime coordinates and are understood to be evaluated at a specific spacetime point labeled by x in one coordinate system and by x' in the other coordinate system.[1] Note that we are using the usual conventions of special relativity, where the manifold is considered to be 4D spacetime [Minkowski space with coordinates ct, x, y, z)] and the use of greek indices signifies that the index can range over the time and all three spatial components.

Furthermore, in Eqs. (A.1) (and in various other places in these lectures) we employ the *Einstein summation convention*, where a repeated index, once in an upper position and once in a lower position signifies an implicit summation on that index. For example, from Eqs. (A.1) for vectors and contravariant rank-2 tensors, respectively,

$$V'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} V^{\nu}(x) \equiv \sum_{\nu} \frac{\partial x'^{\mu}}{\partial x^{\nu}} V^{\nu}(x),$$

$$T'^{\mu\nu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\alpha}} \frac{\partial x'^{\nu}}{\partial x^{\beta}} T^{\alpha\beta}(x) \equiv \sum_{\alpha}\sum_{\beta} \frac{\partial x'^{\mu}}{\partial x^{\alpha}} \frac{\partial x'^{\nu}}{\partial x^{\beta}} T^{\alpha\beta}(x).$$

Note that a repeated index on the right side does not appear on the left side of the equation because it has been summed over, and indices that aren't repeated on the right side of the equation appear in the same positions (upper or lower) on the left side of the equation.

Before proceeding, let us concede that we are being rather sloppy mathematically by referring to objects carrying indices and obeying particular transformation laws as tensors. The indexed quantities appearing in Eqs. (A.1) are actually *tensor components* that have been expressed in a particular basis. Tensors are *geometrical objects,* meaning that their properties are *independent of expression in a particular basis*, as explained further in Box A.1. For example,

1. A *rank-0 tensor* or *scalar* transforms as $\phi'(x') = \phi(x)$ (it is unchanged by a coordinate transformation) and requires a single real number to specify it. "Ordinary numbers", such the age of your dog in years, are scalars.
2. There are two kinds of *rank-1 tensors*. Loosely in physics applications they are often both termed vectors, and the distinction is not of much practical importance for cartesian coordinates in non-curved spaces. However, in the general case of non-cartesian coordinate systems in possibly curved spaces, mathematical consistency requires that one must distinguish

   a. *vectors* (also called *contravariant vectors*), which transform as Eq. (A.1b),

$$V'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} V^{\nu}(x) \equiv \sum_{\nu} \frac{\partial x'^{\mu}}{\partial x^{\nu}} V^{\nu}(x),$$

   and require a single index in an upper position, V^{α}, when expanding in a basis, from

[1] The partial derivatives in Eqs. (A.1) are generally functions of the spacetime coordinate but for the special case of flat Minkowski space and special relativity they are constants taking the same value at each spacetime point. A transformation where points in spacetime are merely relabeled in a new coordinate system is termed a *passive transformation*. A simple example is a plot displayed in cartesian coordinates (x, y, z) that is replotted in spherical polar coordinates (r, θ, ϕ). The physical content is unchanged but each point formerly labeled by values of the old coordinates (x, y, z) is now labeled by values of the new coordinates (r, θ, ϕ).

b. *dual vectors* (also called *covariant vectors* or *1-forms*), which transform as Eq. (A.1c),

$$V'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} V_\nu(x),$$

and require a single index in a lower position, V_α, when expanded in a basis.

3. *Rank-2 tensors* require two indices, when expanding in a basis. Generalizing the distinction between vectors and dual vectors for rank-1 tensors when expanding in a basis, the two indices required for rank-2 tensor components can be in upper or lower positions, so there are three kinds of rank-2 tensors.

a. *Contravariant rank-2 tensors*, which transform as Eq. (A.1f),

$$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha} \frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta},$$

and require two indices in an upper position, $T^{\alpha\beta}$, when expanding in a basis.

b. *Covariant rank-2 tensors*, which transform as Eq. (A.1d),

$$T'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta},$$

and require a two indices in a lower position, $T_{\alpha\beta}$, when expanding in a basis.

c. *Mixed rank-2 tensors*, which transform as Eq. (A.1e),

$$T'^{\nu}{}_{\mu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x'^\nu}{\partial x^\beta} T^{\beta}{}_{\alpha},$$

and require a one index in a lower position and one in an upper position, $T^{\beta}{}_{\alpha}$, when expanding in a basis.

In a similar way, tensors of higher rank can be defined. For most physics applications tensors of rank 0, 1, and 2 are sufficient, but some applications require tensors of rank greater than two. For example, the curvature of 4D spacetime in general relativity is described by a rank-4 tensor called the *Riemann curvature tensor*, which may be viewed as the generalization of gaussian curvature from 2D space to 4D spacetime.

A.2 Vector Algebra

The *scalar product* of two vectors is

$$\boldsymbol{A} \cdot \boldsymbol{B} = AB\cos\theta \qquad \text{(scalar product)}, \qquad \text{(A.2)}$$

where θ is the angle between the vectors $\boldsymbol{A}$ and $\boldsymbol{B}$. The *vector product* or *cross product* of two vectors is

$$\boldsymbol{A} \times \boldsymbol{B} = AB\sin\theta\hat{\boldsymbol{n}} = \begin{vmatrix} \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}} \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix} \qquad \text{(vector or cross product)}, \qquad \text{(A.3)}$$

Box A.1 Geometrical Definition of Tensors

Tensors in Eqs. (A.1) were introduced in terms of the the transformation properties of their components when they are expressed a basis. For real physics or engineering problems it is often simplest to work with the components of tensors expressed in a basis rather than with the tensors themselves, so this is of practical utility. However, viewing tensors in terms of basis components obscures considerable mathematical beauty and elegance associated with tensor properties being independent of expression in any particular basis. Thus mathematicians prefer to define tensors *geometrically* (independent of expression in a particular basis), in terms of linear maps to the real numbers. For example,

1. a dual vector is mathematically an operator that accepts a vector as input and returns a real number, or
2. a dual vector is an object with components (when expanded in a basis) that transforms as Eq. (A.1c),

$$V'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} V_\nu(x),$$

at a point labeled by x in one coordinate system and by x' in a second coordinate system.

The two approaches embody different tradeoffs between utility and elegance, but lead to the same physical results if manipulated correctly. A more extensive discussion of these ideas and a general introduction to spacetime tensors in possibly curved spacetime as the basis for describing special and general relativity may be found in Ref. [11].

where $|\;|$ indicates the determinant, $\hat{\boldsymbol{n}}$ is a unit vector perpendicular to the plane containing $\boldsymbol{A}$ and $\boldsymbol{B}$ (with ambiguity of up and down resolved by the *right-hand rule*), θ is the angle between $\boldsymbol{A}$ and $\boldsymbol{B}$, and the cartesian unit vectors are denoted by $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$. *Note*: The cross product doesn't commute: $\boldsymbol{A}\times\boldsymbol{B} \neq \boldsymbol{B}\times\boldsymbol{A}$, but the scalar product does: $\boldsymbol{A}\cdot\boldsymbol{B} = \boldsymbol{B}\cdot\boldsymbol{A}$.

The *scalar triple product* is

$$\boldsymbol{A}\cdot(\boldsymbol{B}\times\boldsymbol{C}) = \boldsymbol{B}\cdot(\boldsymbol{C}\times\boldsymbol{A}) = \boldsymbol{C}\cdot(\boldsymbol{A}\times\boldsymbol{B}) = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}. \tag{A.4}$$

The *vector triple product* is

$$\boldsymbol{A}\times(\boldsymbol{B}\times\boldsymbol{C}) = \boldsymbol{B}(\boldsymbol{A}\cdot\boldsymbol{C}) - \boldsymbol{C}(\boldsymbol{A}\cdot\boldsymbol{B}). \tag{A.5}$$

A.3 Useful Vector Identities

Some identities that are useful in manipulating vector equations are collected here, with $\boldsymbol{A}$ assumed to be an arbitrary vector and f assumed to be an arbitrary scalar.

$$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0, \tag{A.6}$$

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla} f) = 0, \tag{A.7}$$

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \nabla^2\boldsymbol{A}. \tag{A.8}$$

Proofs may be found in standard books such as Griffiths [8].

A.4 Vector Calculus

Let's now consider the extension of the principles of calculus to vector equations, first for derivatives and then for integrals.

A.4.1 First Derivatives

The *gradient* $\boldsymbol{\nabla}$ of a scalar function F,

$$\boldsymbol{\nabla} F = \frac{\partial F}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial F}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial F}{\partial z}\hat{\boldsymbol{z}} \qquad \text{(gradient)}, \tag{A.9}$$

is a vector quantity. The gradient can be viewed as a vector operator

$$\boldsymbol{\nabla} = \hat{\boldsymbol{x}}\frac{\partial}{\partial x} + \hat{\boldsymbol{y}}\frac{\partial}{\partial y} + \hat{\boldsymbol{z}}\frac{\partial}{\partial z} \qquad \text{(gradient operator)} \tag{A.10}$$

that acts upon a scalar argument following it. The notation $\boldsymbol{\nabla}_x$ means explicitly that the gradient acts on the coordinate $\boldsymbol{x}$ while $\boldsymbol{\nabla}_{x'}$ means that the gradient acts on the coordinates $\boldsymbol{x}'$. (In the text we will often use the abbreviations $\boldsymbol{\nabla} \equiv \boldsymbol{\nabla}_x$ and $\boldsymbol{\nabla}' \equiv \boldsymbol{\nabla}_{x'}$. Useful identities involving the gradient and Laplacian:

$$\boldsymbol{\nabla}_x\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}, \tag{A.11a}$$

$$\boldsymbol{\nabla}_{x'}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = \frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}, \tag{A.11b}$$

$$\boldsymbol{\nabla}_x\left(\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\right) = 4\pi\delta(\boldsymbol{x}-\boldsymbol{x}'), \tag{A.11c}$$

$$\boldsymbol{\nabla}_{x'}\left(\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\right) = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}'), \tag{A.11d}$$

$$\boldsymbol{\nabla}_x^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = \boldsymbol{\nabla}_{x'}^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}'), \tag{A.11e}$$

It will be useful to define the completely antisymmetric rank-3 tensor (Levi–Civita symbol) by

$$\varepsilon_{ijk} = \begin{cases} 1 & (\text{ if } i,j,k \text{ is cyclic permutation of } 1,2,3), \\ -1 & (\text{if } i,j,k \text{ is cyclic permutation of } 1,2,3), \\ 0 & (\text{otherwise}), \end{cases} \tag{A.12}$$

The ε_{ijk} obey two important identities

$$\begin{aligned} \varepsilon_{ijk}\varepsilon_{lmn} &= \delta_{il}\delta_{jm}\delta_{kn} + \delta_{jl}\delta_{km}\delta_{in} + \delta_{kl}\delta_{im}\delta_{jn} \\ &\quad - \delta_{jl}\delta_{im}\delta_{kn} - \delta_{kl}\delta_{jm}\delta_{in} - \delta_{il}\delta_{km}\delta_{jn}, \end{aligned} \tag{A.13}$$

and

$$\sum_i \varepsilon_{ijk}\varepsilon_{imn} = \delta_{jm}\delta_{kn} - \delta_{km}\delta_{jn}. \tag{A.14}$$

The *divergence* of a vector $\boldsymbol{V}$ is

$$\boldsymbol{\nabla}\cdot V = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z} \qquad \text{(divergence)}. \tag{A.15}$$

Note: The divergence of a vector is a scalar.

The *curl* $(\boldsymbol{\nabla}\times)$ of a vector $\boldsymbol{V}$ is

$$\begin{aligned} \boldsymbol{\nabla}\times\boldsymbol{V} &= \begin{vmatrix} \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ V_x & V_y & V_z \end{vmatrix} \\ &= \left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z}\right)\hat{\boldsymbol{x}} + \left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x}\right)\hat{\boldsymbol{y}} + \left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y}\right)\hat{\boldsymbol{z}} \qquad \text{(curl)}, \end{aligned} \tag{A.16}$$

which can also be written as

$$(\boldsymbol{A}\times\boldsymbol{B})_i = \sum_{jk} \varepsilon_{ijk} A_j B_k \tag{A.17}$$

Note: The curl of a vector is a vector.

Derivatives of products are common in realistic problems and product rules for vector derivatives can be proved that are analogous to product rules for ordinary derivatives. For example,

$$\boldsymbol{\nabla}(fg) = f\boldsymbol{\nabla}g + g\boldsymbol{\nabla}f, \tag{A.18a}$$

$$\boldsymbol{\nabla}(\boldsymbol{A}\cdot\boldsymbol{B}) = \boldsymbol{A}\times(\boldsymbol{\nabla}\times\boldsymbol{B}) + \boldsymbol{B}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) + (\boldsymbol{A}\cdot\boldsymbol{\nabla})\boldsymbol{B} + (\boldsymbol{B}\cdot\boldsymbol{\nabla})\boldsymbol{A}, \tag{A.18b}$$

$$\boldsymbol{\nabla}\cdot(f\boldsymbol{A}) = f(\boldsymbol{\nabla}\cdot\boldsymbol{A}) + \boldsymbol{A}\cdot(\boldsymbol{\nabla}f), \tag{A.18c}$$

$$\boldsymbol{\nabla}\cdot(\boldsymbol{A}\times\boldsymbol{B}) = \boldsymbol{B}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) - \boldsymbol{A}\cdot(\boldsymbol{\nabla}\times\boldsymbol{B}), \tag{A.18d}$$

$$\boldsymbol{\nabla}\times(f\boldsymbol{A}) = f(\boldsymbol{\nabla}\times\boldsymbol{A}) - \boldsymbol{A}\times(\boldsymbol{\nabla}f), \tag{A.18e}$$

$$\boldsymbol{\nabla}\times(\boldsymbol{A}\times\boldsymbol{B}) = (\boldsymbol{B}\cdot\boldsymbol{\nabla})\boldsymbol{A} - (\boldsymbol{A}\cdot\boldsymbol{\nabla})\boldsymbol{B} + \boldsymbol{A}(\boldsymbol{\nabla}\cdot\boldsymbol{B}) - \boldsymbol{B}(\boldsymbol{\nabla}\cdot\boldsymbol{A}), \tag{A.18f}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are arbitrary vectors and f and g are arbitrary scalar functions,

A.4.2 Second Derivatives

As shown above in Section A.4.1, the first derivatives that can be constructed using the operator $\boldsymbol{\nabla}$ defined in Eq. (A.10) are

1. the *gradient* of a scalar, $\boldsymbol{\nabla} f$,
2. the *divergence* of a vector, $\boldsymbol{\nabla} \cdot \boldsymbol{A}$, and
3. the *curl* of a vector, $\boldsymbol{\nabla} \times \boldsymbol{A}$.

Second derivatives can be constructed by applying two first-derivative operators in succession. Lets consider the possibilities for second derivatives in vector equations.

1. The *divergence of a gradient*, $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} f)$, gives in cartesian coordinates

$$\begin{aligned}\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} f) &= \left(\hat{\boldsymbol{x}}\frac{\partial}{\partial x} + \hat{\boldsymbol{y}}\frac{\partial}{\partial y} + \hat{\boldsymbol{z}}\frac{\partial}{\partial z}\right) \cdot \left(\hat{\boldsymbol{x}}\frac{\partial f}{\partial x} + \hat{\boldsymbol{y}}\frac{\partial f}{\partial y} + \hat{\boldsymbol{z}}\frac{\partial f}{\partial z}\right) \\ &= \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.\end{aligned}$$

This second derivative operator appears often and it is useful to define it as ∇^2; this is called the *Laplacian operator*, and we rewrite the preceding result as

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}. \tag{A.19}$$

Notice that *the Laplacian applied to a scalar gives a scalar*.

2. The *curl of a gradient is always equal to zero*, $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} f) = 0$, by the identity (A.7).

3. The *gradient of a divergence* $\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A})$ is a valid mathematical operation but it is relatively uncommon in physical problems and has no special name.

4. The *divergence of a curl always vanishes*, $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{A}) = 0$, by the identity (A.6).

5. The *curl of a curl* evaluates to

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}$$

[see identity (A.8)], which gives nothing new since the first term on the right is just a number and the second is the Laplacian of a vector, which we considered in Eq. (A.19).

Thus we need consider only two kinds of second derivatives: (1) the Laplacian in Eq. (A.19), which will play a fundamental role in electromagnetism, and the gradient of the divergence, which isn't very common. By similar procedures we could evaluate third derivatives in vector equations, but we will omit them since they seldom occur for the types of problems that we shall encounter.

A.4.3 Integrals

A *line integral* (or *path integral*) of a vector function $\boldsymbol{V}$ between points $\boldsymbol{A}$ and $\boldsymbol{B}$ is given by

$$\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{V} \cdot d\boldsymbol{l} \qquad d\boldsymbol{l} \equiv \hat{\boldsymbol{x}}\,dx + \hat{\boldsymbol{y}}\,dy + \hat{\boldsymbol{z}}\,dz \qquad \text{(line integral).} \tag{A.20}$$

The integral is carried out over a prescribed path from $\boldsymbol{A}$ to $\boldsymbol{B}$, with $d\boldsymbol{l}$ the infinitesimal displacement vector along the specified path (with magnitude equal to the length of the infinitesimal displacement and direction tangent to the path at that point). If the line integral is carried out over a closed loop ($\boldsymbol{B} = \boldsymbol{A}$), the line integral is called a *cyclic integral* and denoted

$$\oint \boldsymbol{V} \cdot d\boldsymbol{l} \qquad \text{(closed line integral).} \tag{A.21}$$

A *surface integral* of a vector function $\boldsymbol{V}$ over a surface S is denoted by

$$\int_S \boldsymbol{V} \cdot d\boldsymbol{s} \qquad \text{(surface integral),} \tag{A.22}$$

where $d\boldsymbol{s} \equiv \boldsymbol{n}\, ds$ is a vector associated with an infinitesimal patch of surface area having magnitude equal to the area of the patch and direction given by the normal $\boldsymbol{n}$ to the surface at the patch (implying that the sign is ambiguous since, there are two normals—"up" and "down"—at each point). If the surface is closed, the surface integral is denoted

$$\oint_S \boldsymbol{V} \cdot d\boldsymbol{s} \qquad \text{(integral over closed surface),} \tag{A.23}$$

A *volume integral of a scalar function* over a volume V is indicated by

$$\int_V F\, d\tau \qquad \text{(scalar volume integral),} \tag{A.24}$$

where F is a scalar function and $d\tau$ is an infinitesimal volume element. (For example, in cartesian coordinates $d\tau = dx\, dy\, dz$.) A *volume integral for a vector function* $\boldsymbol{F}$ is given by

$$\begin{aligned} \int \boldsymbol{F}\, d\tau &= \int_V (F_x \hat{\boldsymbol{x}} + F_y \hat{\boldsymbol{y}} + F_z \hat{\boldsymbol{z}}) \\ &= \hat{\boldsymbol{x}} \int_V F_x\, d\tau + \hat{\boldsymbol{y}} \int F_y\, d\tau + \hat{\boldsymbol{z}} \int F_z\, d\tau \qquad \text{(vector volume integral),} \end{aligned} \tag{A.25}$$

where the unit vectors $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$ have been pulled out of the integrals in the last step because they are constants.

A.5 Fundamental Theorems

Let us now list some fundamental theorems of calculus and of vector calculus that will be of use. The *fundamental theorem of (ordinary) calculus* is

$$\int_a^b \left(\frac{\partial f}{\partial x} \right) dx = \int_a^b F(x)\, dx = f(b) - f(a) \quad \text{(fundamental theorem of calculus),} \tag{A.26}$$

where $F(x) \equiv df/dx$. Thus the fundamental theorem of calculus tells us how to integrate $F(x)$: find a function $f(x)$ that has a derivative equal to $F(x)$, which implies that *integration and differentiation are inverse operations*. In vector calculus there are three fundamental types of derivatives:

1. the *gradient* of Eq. (A.9),

2. the *divergence* of Eq. (A.15), and
3. the *curl* of Eq. (A.16).

For each type of vector derivative one can formulate a "fundamental theorem" having the same general structure as the fundamental theorem of calculus given in Eq. (A.26):

> The integral of a derivative of a function over some region is determined by the values of the function on the boundaries of the region.

For *gradients*, this takes the form

$$\int_{\boldsymbol{A}}^{\boldsymbol{B}} (\boldsymbol{\nabla} F) \cdot d\boldsymbol{l} = F(\boldsymbol{B}) - F(\boldsymbol{A}) \qquad \text{(fundamental theorem for gradients)}, \tag{A.27}$$

for a scalar function $F(x,y,z)$ evaluated on a path from $\boldsymbol{A}$ to $\boldsymbol{B}$. Notice that Eq. (A.27) has the same structure as Eq. (A.26): the integral (a line integral here) of a derivative (a gradient here) is determined by the values of the function at the boundaries ($\boldsymbol{A}$ and $\boldsymbol{B}$ here).

For *divergences*, the fundamental theorem takes the form [see Eqs. (A.23) and (A.24)],

$$\int_V (\boldsymbol{\nabla} \cdot \boldsymbol{A})\, d\tau = \oint_S \boldsymbol{A} \cdot d\boldsymbol{s} \qquad \text{(divergence theorem)}, \tag{A.28}$$

where $\boldsymbol{A}$ is a vector field and $d\boldsymbol{s} = \boldsymbol{n}\, da$ with $\boldsymbol{n}$ the normal to the surface. This may be termed the *fundamental theorem for divergences*[2] but, as we have indicated, it is commonly termed *the divergence theorem* in the literature and we will typically adopt that terminology.

The *fundamental theorem for curls* is given by

$$\int_S (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot d\boldsymbol{s} = \oint_P \boldsymbol{A} \cdot d\boldsymbol{l} \qquad \text{(Stokes' theorem)}, \tag{A.29}$$

where the left side is a surface integral over S and the right side is a line integral on the perimeter of the surface. This is the fundamental theorem for curls[3] but, as we have indicated, this is commonly called *Stokes' theorem* in the literature. We will typically use that terminology. Further discussion of Stokes' theorem is given in Box 2.3.

A.6 Laws, Theorems, and Definitions

Some important laws, theorems, and definitions of classical electromagnetism are listed here. In some cases both the integral and differential forms of the laws are given. Unless otherwise noted, SI units are assumed in all equations.

[2] Note that it has the same structure as Eq. (A.26): the integral of a derivative (here the divergence) over a region (here the volume V) is determined by the value of the function on the boundaries (the boundary of the volume V is the surface S. In this case the boundary term is a surface integral.

[3] Analogous to Eq. (A.26), the integral of a derivative (here the curl) over a region (here a patch of surface S) is determined by the value of the function on the boundary (the perimeter P of the surface patch). Notice that the boundary term in this case is itself an integral.

Coulomb's law (forces): The force between two charges is

$$\boldsymbol{F} = \frac{q_1 q_2}{4\pi\varepsilon_0} \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{|\boldsymbol{x}_1 - \boldsymbol{x}_2|^3} \tag{A.30}$$

where the charges q_1 and q_2 are located at the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

Coulomb's law (electric fields): For a continuous charge distribution,

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \int \rho(\boldsymbol{x}') \frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3} d^3x', \tag{A.31}$$

where $\rho(\boldsymbol{x})$ is the charge density.

Gauss's law: For a continuous charge distribution $\rho(\boldsymbol{x})$,

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0} \qquad \oint \boldsymbol{E} \cdot \boldsymbol{n}\, da = \frac{1}{\varepsilon_0} \int_V \rho(\boldsymbol{x})\, d^3x, \tag{A.32}$$

where the integration is over the volume V contained within the surface S (see Fig. 2.3).

Divergence theorem: For a vector field defined within a volume V that is enclosed by a surface S,

$$\oint_S \boldsymbol{A} \cdot \boldsymbol{n}\, da = \int_V \boldsymbol{\nabla} \cdot \boldsymbol{A}\, d^3x, \tag{A.33}$$

where the left side is the surface integral of the outwardly directed normal component of the vector $\boldsymbol{A}$ and the right side is the volume integral of the divergence of $\boldsymbol{A}$ [equivalent to fundamental theorem for divergences given in Eq. (A.28)].

Stokes' theorem: For a 3D vector field $\boldsymbol{A}$ (see Box 2.3),

$$\oint_C \boldsymbol{A} \cdot d\boldsymbol{r} = \int_S (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot \boldsymbol{n}\, ds \equiv \int_S (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot d\boldsymbol{s}, \tag{A.34}$$

where S is the 2D surface enclosed by the 1D boundary C and the outward normal to the surface is $\boldsymbol{n}$ [equivalent to fundamental theorem for curls given in Eq. (A.29)].

Scalar potential: For an electric field $\boldsymbol{E}$ and scalar potential Φ,

$$\boldsymbol{E} = -\boldsymbol{\nabla}\Phi \qquad \Phi(\boldsymbol{x}) = -\int_{\mathscr{O}}^{\boldsymbol{x}} \boldsymbol{E}(\boldsymbol{x}) \cdot d\boldsymbol{l}, \tag{A.35}$$

where $\mathscr{O}$ is an arbitrary reference point.

Vector potential: The vector potential $\boldsymbol{A}$ is defined through

$$\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}, \tag{A.36}$$

where $\boldsymbol{B}$ is the magnetic field. Eqs. (A.35) and (A.36) then allow the electric and magnetic fields to be eliminated in favor of vector and scalar potentials.

Poisson's equation: For a scalar field Φ,

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon_0}, \tag{A.37}$$

where the charge density is ρ.

Laplace's equation: For a scalar field Φ,

$$\nabla^2\Phi = 0. \tag{A.38}$$

This is Poisson's equation with zero charge density.

Lorentz force: For electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{B}$,

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}), \tag{A.39}$$

where q is the charge and $\boldsymbol{v}$ is the velocity of the particle.

A.7 Curvilinear Coordinate Systems

It is often advantageous to formulate electromagnetic problems in non-cartesian coordinates. We include some relevant equations in this Appendix for spherical coordinates and cylindrical coordinates.

A.7.1 Spherical Coordinates

The standard spherical coordinates (r, θ, ϕ) are related to cartesian coordinates (x, y, z) by

$$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\cos\theta. \tag{A.40}$$

Spherical unit vectors are related to cartesian unit vectors by

$$\begin{aligned} \hat{\boldsymbol{r}} &= \sin\theta\cos\phi\,\hat{\boldsymbol{x}} + \sin\theta\sin\phi\,\hat{\boldsymbol{y}} + \cos\theta\,\hat{\boldsymbol{z}}, \\ \hat{\boldsymbol{\theta}} &= \cos\theta\cos\phi\,\hat{\boldsymbol{x}} + \cos\theta\sin\phi\,\hat{\boldsymbol{y}} - \sin\theta\,\hat{\boldsymbol{z}}, \\ \hat{\boldsymbol{\phi}} &= -\sin\phi\,\hat{\boldsymbol{x}} + \cos\phi\,\hat{\boldsymbol{y}}, \end{aligned} \tag{A.41}$$

or in matrix form, the transformation from cartesian to spherical coordinates is

$$\begin{pmatrix} \hat{\boldsymbol{r}} \\ \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{x}} \\ \hat{\boldsymbol{y}} \\ \hat{\boldsymbol{z}} \end{pmatrix}. \tag{A.42}$$

The transformation matrix is orthogonal so the inverse matrix is the transpose (interchange rows and columns) and the matrix transformation from spherical to cartesian coordinates is given by

$$\begin{pmatrix} \hat{\boldsymbol{x}} \\ \hat{\boldsymbol{y}} \\ \hat{\boldsymbol{z}} \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & \cos\phi \\ \cos\theta & -\sin\theta & 0 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{r}} \\ \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix}. \tag{A.43}$$

This means that an arbitrarary vector $\boldsymbol{V}$ can be decomposed as

$$\boldsymbol{V} = (\hat{\boldsymbol{r}} \cdot \boldsymbol{V})\hat{\boldsymbol{r}} + (\hat{\boldsymbol{\theta}} \cdot \boldsymbol{V})\hat{\boldsymbol{\theta}} + (\hat{\boldsymbol{\phi}} \cdot \boldsymbol{V})\hat{\boldsymbol{\phi}}. \tag{A.44}$$

The spherical line element is

$$d\boldsymbol{l} = dr\,\hat{\boldsymbol{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\phi\,\hat{\boldsymbol{\phi}}, \tag{A.45}$$

and the spherical volume element is

$$d\tau = r^2 \sin\theta\,dr\,d\theta\,d\phi. \tag{A.46}$$

The vector derivatives are given by

$$\text{Gradient:}\quad \boldsymbol{\nabla} F = \frac{\partial F}{\partial r}\hat{\boldsymbol{r}} + \frac{1}{r}\frac{\partial F}{\partial \theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial F}{\partial \theta}\hat{\boldsymbol{\phi}}, \tag{A.47}$$

$$\text{Divergence:}\quad \boldsymbol{\nabla}\cdot\boldsymbol{V} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 V_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta V_\theta) + \frac{1}{r\sin\theta}\frac{\partial V_\phi}{\partial\phi}, \tag{A.48}$$

$$\begin{aligned}\text{Curl:}\quad \boldsymbol{\nabla}\times\boldsymbol{V} &= \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta V_\phi) - \frac{\partial V_\theta}{\partial\phi}\right]\hat{\boldsymbol{r}} \\ &\quad + \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial V_r}{\partial\phi} - \frac{\partial}{\partial r}(rV_\phi)\right]\hat{\boldsymbol{\theta}} + \frac{1}{r}\left[\frac{\partial}{\partial r}(rV_\theta - \frac{\partial V_r}{\partial\theta}\right]\hat{\boldsymbol{\phi}},\end{aligned} \tag{A.49}$$

$$\begin{aligned}\text{Laplacian:}\quad \nabla^2 F &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial F}{\partial r}\right) \\ &\quad + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial F}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 F}{\partial\phi^2},\end{aligned} \tag{A.50}$$

when expressed in spherical coordinates.

A.7.2 Cylindrical Coordinates

Standard cylindrical coordinates (ρ,ϕ,z) are related to cartesian coordinates (x,y,z) by

$$x = \rho\cos\phi \qquad y = \rho\sin\phi \qquad z = z, \tag{A.51}$$

the line element is

$$d\boldsymbol{l} = d\rho\,\hat{\boldsymbol{\rho}} + \rho\,d\phi\,\hat{\boldsymbol{\phi}} + dz\,\hat{\boldsymbol{z}}, \tag{A.52}$$

and the volume element is

$$d\tau = \rho\,d\rho\,d\phi\,dz. \tag{A.53}$$

The vector derivatives are given by

$$\text{Gradient:}\quad \boldsymbol{\nabla} F = \frac{\partial F}{\partial \rho}\hat{\boldsymbol{\rho}} + \frac{1}{\rho}\frac{\partial F}{\partial \phi}\hat{\boldsymbol{\phi}} + \frac{\partial F}{\partial z}\hat{\boldsymbol{z}}, \tag{A.54}$$

$$\text{Divergence:}\quad \boldsymbol{\nabla}\cdot\boldsymbol{V} = \frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho V_\rho) + \frac{1}{\rho}\frac{\partial V_\phi}{\partial \phi} + \frac{\partial V_z}{\partial z}, \tag{A.55}$$

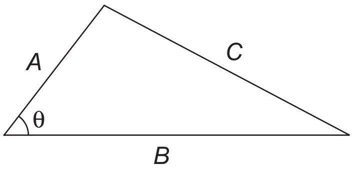
$$\begin{aligned}\text{Curl:}\quad \boldsymbol{\nabla}\times\boldsymbol{V} = &\left(\frac{1}{\rho}\frac{\partial V_z}{\partial \phi} - \frac{\partial V_\phi}{\partial z}\right)\hat{\boldsymbol{\rho}}\\ &+\left(\frac{\partial V_\rho}{\partial z} - \frac{\partial V_z}{\partial \rho}\right)\hat{\boldsymbol{\phi}} + \frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho V_\phi - \frac{\partial V_\rho}{\partial \phi}\right]\hat{\boldsymbol{z}},\end{aligned} \tag{A.56}$$

$$\text{Laplacian:}\quad \nabla^2 F = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial F}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 F}{\partial \phi^2} + \frac{\partial^2 F}{\partial z^2}, \tag{A.57}$$

when expressed in cylindrical coordinates.

A.8 Miscellaneous Equations

LAW OF COSINES:



$$C^2 = A^2 + B^2 - 2AB\cos\theta \qquad \text{(Law of Cosines).} \tag{A.58}$$

INTEGRATION BY PARTS:

$$\int_a^b f\left(\frac{dg}{dx}\right)dx = fg|_a^b - \int_a^b g\left(\frac{df}{dx}\right)dx. \tag{A.59}$$

A.9 The Dirac Delta Function

In one dimension the *Dirac delta function* is a (mathematically improper) function written $\delta(x-a)$ and having the properties that

$$\delta(x-a) = \begin{cases} 0 & (\text{if } x \neq a), \\ \infty & (\text{if } x = a), \end{cases} \tag{A.60}$$

with the normalization

$$\int_{-\infty}^{+\infty} \delta(x-a)dx = 1. \tag{A.61}$$

The Dirac delta function can be viewed intuitively (but non-rigorously) as the limit of a peaked curve that becomes higher and higher as it is made narrower and narrower, in such as way that the *area under the curve remains constant*.[4] For an arbitrary continuous function $f(x)$,

$$f(x)\delta(x-a) = f(a)\delta(x-a) \tag{A.62}$$

and insertion of a Dirac delta function in an integral over a function picks out the value of the integrand at $x = a$,

$$\int_{-\infty}^{+\infty} f(x)\delta(x-a)dx = f(a). \tag{A.63}$$

In n dimensions the Dirac delta function $\delta^n(\boldsymbol{x}-\boldsymbol{X})$ can be written as a product of n 1D delta functions; for example, in a 3D space parameterized by cartesian coordinates $\boldsymbol{x} = (x_1, x_2, x_3)$,

$$\delta^3(\boldsymbol{x}-\boldsymbol{X}) \equiv \delta(x_1 - X_1)\delta(x_2 - X_2)\delta(x_3 - X_3). \tag{A.64}$$

This vanishes everywhere except at $\boldsymbol{x} = \boldsymbol{X}$, and generalizing Eqs. (A.60) and (A.61) to 3D,

$$\int_{\Delta V} \delta^3(\boldsymbol{x}-\boldsymbol{X})\, d^3x = \begin{cases} 1 & (\text{if } \Delta V \text{ contains } \boldsymbol{x} = \boldsymbol{X}), \\ 0 & (\text{if } \Delta V \text{ doesn't contain } \boldsymbol{x} = \boldsymbol{X}), \end{cases} \tag{A.65}$$

where ΔV is the integration volume over d^3x, while Eq. (A.63) generalizes to

$$\int_{\Delta V} f(\boldsymbol{x})\, \delta^3(\boldsymbol{x}-\boldsymbol{a})\, d^3x = f(\boldsymbol{a}). \tag{A.66}$$

Thus, just as in 1D the 3D Dirac delta function $\delta^3(\boldsymbol{x}-\boldsymbol{a})$ picks out the point $\boldsymbol{x} = \boldsymbol{a}$ in the integration. Notice from the preceding definitions that a Dirac delta function in n dimensions has the dimensionality of inverse volume in n-dimensional space. Applying the Laplacian operator ∇^2 to $|\boldsymbol{x}-\boldsymbol{x}'|^{-1}$ gives a result proportional to a 3D δ-function,

$$\nabla^2 \left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} \right) = -4\pi\delta^3(\boldsymbol{x}-\boldsymbol{x}'), \tag{A.67}$$

which can be of considerable utility in evaluating electrostatic integrals (see Example 2.5). By evaluating

$$\int_a^b \frac{d\delta(x)}{dx} f(x)\, dx$$

by “parts” for $f(x)$ an arbitrary function, one can show that the delta function *anticommutes with derivative operators such as* $\boldsymbol{\nabla}$ under an integral; schematically,

$$\frac{d}{dx}\delta = -\delta\frac{d}{dx}. \tag{A.68}$$

[4] The Dirac delta function was first introduced by Dirac to aid in normalization of probability integrals in quantum mechanics. It was met by considerable initial skepticism from mathematicians. However, contrary to what might be inferred from our loose discussion here, the Dirac delta function can be placed on a mathematically rigorous footing by viewing it as a *generalized function* or *distribution* (meaning a quantity that only makes sense when it is integrated over) over the real numbers, which has a value of zero everywhere except at zero, and has an integral over the entire real number line that is equal to one. See Lighthill [20] for a more complete exposition.

(One can show that this “parts” operation is legitimate, even though δ is not a true function.)

B Appendix B **Electromagnetic Units**

We have systematically tended to use SI units throughout this book. In this Appendix we give examples of important equations expressed in other units, most notably in the gaussian or CGS system.

B.1 Maxwell Equations in SI Units

In SI units the vacuum Maxwell equations are

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)}, \tag{B.1a}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{B.1b}$$

$$\nabla \cdot \boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{B.1c}$$

$$\nabla \times \boldsymbol{B} - \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} = \mu_0 \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{B.1d}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, ρ is the charge density, $\boldsymbol{J}$ is the current vector, ε_0 is the permittivity of free space [defined for SI units in Eq. (2.3)], and μ_0 is the permeability of free space (which are related by $\varepsilon_0\mu_0 = 1/c^2$). The corresponding Lorentz force in SI units is

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \qquad \text{(Lorentz force)}, \tag{B.2}$$

where q is the charge and $\boldsymbol{b}$ the velocity of a test charge.

B.2 Maxwell Equations in Gaussian (CGS) Units

In gaussian (CGS) units the vacuum Maxwell equations are

$$\nabla \cdot \boldsymbol{E} = 4\pi\rho \qquad \text{(Gauss's law)}, \tag{B.3a}$$

$$\nabla \times \boldsymbol{E} + \frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{B.3b}$$

$$\nabla \cdot \boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{B.3c}$$

$$\nabla \times \boldsymbol{B} - \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} = \frac{4\pi}{c}\boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{B.3d}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, ρ is the charge density, and $\boldsymbol{J}$ is the current vector. The corresponding Lorentz force in CGS units is

$$\boldsymbol{F} = q\left(\boldsymbol{E} + \frac{1}{c}\boldsymbol{v} \times \boldsymbol{B}\right) \qquad \text{(Lorentz force)}, \tag{B.4}$$

where q is the charge and $\boldsymbol{v}$ the velocity of a test charge.

B.3 Maxwell Equations in Heaviside–Lorentz Units

In Heaviside–Lorentz units (common in high energy physics) the vacuum Maxwell equations are given by

$$\nabla \cdot \boldsymbol{E} = \rho \qquad \text{(Gauss's law)}, \tag{B.5a}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{B.5b}$$

$$\nabla \cdot \boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{B.5c}$$

$$\nabla \times \boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{B.5d}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, ρ is the charge density, and $\boldsymbol{J}$ is the current vector. The corresponding Lorentz force in CGS units is

$$\boldsymbol{F} = q\left(\boldsymbol{E} + \frac{1}{c}\boldsymbol{v} \times \boldsymbol{B}\right) \qquad \text{(Lorentz force)}, \tag{B.6}$$

where q is the charge and $\boldsymbol{v}$ the velocity of a test charge.

References

[1] Arfken, G. 1970. *Mathematical Methods for Physicists*. New York: Academic Press.

[2] Barrera, R. G., Estevez, G. A., and Giraldo, J. 1985. *European J. Physics*, **6**, 287.

[3] Chaichian, M., Merches, I., Radu, D., and Tureanu, A. 2016. *Electrodynamics: An Intensive Course*. Berlin: Springer.

[4] Corson, D., and Lorrain, P. 1962. *Introduction to Electromagnetic Fields and Waves*. San Francisco: W. H. Freeman and Company.

[5] Foster, J., and Nightingale, J. D. 2006. *A Short Course in General Relativity*. New York: Springer.

[6] Garg, A. 2012. *Classical Electromagnetism in a Nutshell*. Princeton: Princeton University Press.

[7] Goldhaber, A. S., and Nieto, M. M. 1971. *Rev. Mod. Phys.*, **43**, 277.

[8] Griffiths, D. J. 2024. *Introduction to Electrodynamics*. Cambridge: Cambridge University Press.

[9] Guidry, M. W. 1999. *Gauge Field Theories: An Introduction with Application*. New York: Wiley Interscience.

[10] Guidry, M. W. 2018. *Sci. Bull.*, **63**, 2 (DOI: 10.1016/j.scib.2017.11.021).

[11] Guidry, M. W. 2019. *Modern General Relativity: Black Holes, Gravitational Waves, and Cosmology*. Cambridge: Cambridge University Press.

[12] Guidry, M. W., and Sun, Y. 2022. *Symmetry, Broken Symmetry, and Topology in Modern Physics: A First Course*. Cambridge: Cambridge University Press.

[13] Hartle, James B. 2003. *Gravity, An Introduction to Einstein's General Relativity*. Addison–Wesley.

[14] Hunt, B. J. 2012 (11). *Phys. Today.*, **65**, 48.

[15] Jackson, J. D. 1999. *Classical Electrodynamics*. New York: Wiley.

[16] Jackson, J. D., and Okun, L. B. 2001. *Phys. Rev. Mod. Phys.*, **75**, 663.

[17] Kamberaj, H. 2022. *Electromagnetism*. Springer.

[18] Kobzarev, I. Yu., and Okun, L. B. 1968. *Sov. Phys. Usp.*, **11**, 338.

[19] Kragh, H. 1991. *Applied Optics*, **30**, 4688.

[20] Lighthill, M. J. 1958. *Introduction to Fourier Analysis and Generalised Functions*. Cambridge: Cambridge University Press.

[21] Lorrain, P., and Corson, D. R. 1978. *Electromagnetism: Principles and Applications*. New York: W. H. Freeman and Company.

[22] Maxwell, J. C. 1865. A Dynamical Theory of the Electromagnetic Field. *Philosophical Transactions of the Royal Society of London*, **155**, 459 (DOI: 10.1098/rstl.1865.0008).

[23] Purcell, E. M., and Morin, D. J. 2013. *Electricity and Magnetism*. Cambridge: Cambridge University Press.

[24] Schutz, Bernard. 1980. *Geometrical Methods of Mathematical Physics*. Cambridge: Cambridge University Press.

[25] Wald, R. M. 2022. *Advanced Classical Electromagnetism*. Princeton: Princeton University Press.

[26] Williams, E. R, Faller, J. E, and Hill, H. A. 1971. *Phys. Rev. Lett.*, **26**, 721.

[27] Zangwill, A. 2013. *Modern Electrodynamics*. Cambridge: Cambridge University Press.

Index