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 [19]. 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, material in the book by Griffiths [13]). These lectures are a work in progress, so please do not distribute them without notifying me.

Solutions to all problems at the ends of chapters are contained in the *Instructor Solutions Manual*, which is available only to instructors. A number of problems at the ends of chapters are marked ***, indicating that those problems are contained in the *Student's Solutions Manual*, available to any students taking the course.

Mike Guidry
Knoxville, Tennessee
June 14, 2025

1 Overview

The first inkling of the properties that we now attribute to electricity and magnetism traces to the distant past when humans began to realize and remark upon observed physical phenomena such as the behavior of naturally occurring electricity in amber and the properties of naturally occurring magnetism in lodestones. Although these phenomena were known qualitatively to the ancient Greeks, they remained mysterious for centuries and the modern quantitative understanding of electricity and magnetism emerged over a period of only a little over a century, beginning in the late 1700s.

1.1 The Synthesis of Classical Electromagnetism

At the risk of slighting the contributions of many, one could say that the explosion in knowledge of electricity and magnetism beginning in the late 18th century, and the forging of that knowledge into a theoretically and mathematically coherent framework that we now call electromagnetism over the next century or so, can be illustrated by citing a few landmark achievements.

1. *Henry Cavendish* (1731-1810) did pioneering experiments on electrostatics in the early 1770s, including establishing that electrical forces varied as one over the square of the distance. However, his work was not widely published until after Coulomb's publication in 1785.
2. *Charles-Augustin de Coulomb* (1736-1806) published his extensive work on electrostatics beginning in 1785, including the eponymous inverse-square law for electrical interaction between charges.
3. *Hans Christian Ørsted* (1777-1827) discovered the magnetic effect of the electric current, establishing the first connection between electric and magnetic phenomena.
4. *André–Marie Ampère* (1775-1827) extended the work of Ørsted, finding in 1823 the circuit law between electric current passing through a loop and magnetic field around the loop, and establishing the formula describing the interaction of two currents.
5. *Michael Faraday* (1791-1867) did his highly influential work studying time-varying currents and fields in the mid-1800s. He discovered electromagnetic induction in 1831, which proved to be a crucial step in Maxwell's unification of electric and magnetic phenomena into modern electromagnetism.
6. *James Clerk Maxwell* (1831-1879) published his famous 1865 paper (read to the Royal Society in 1864), which unified electricity and magnetism and synthesized in equations a dynamical theory of what could now be termed the *electromagnetic field* [27, 30].

7. *Oliver Heaviside* (1850-1925) reformulated Maxwell's equations in their more modern vector calculus form in the late 1800s.
8. Heinrich Rudolph Herz (1857-1894) produced the transverse electromagnetic waves predicted by Maxwell's theory in the laboratory in 1888, and studied their wave properties (refraction, reflection, ...) This placed Maxwell's theory on firm experimental footing and established experimentally the connection between electromagnetism and optics suggested by Maxwell's theory.
9. *Joseph John Thomson* (1856-1940) discovered the electron in 1897. This was the first step in elaborating how electromagnetic waves interacted with matter at a microscopic level, though a full microscopic theory of atoms and molecules awaited the invention of quantum mechanics in ∼1925-1926. This, for example, allowed the classical Drude model for conduction electrons described in Section 7.3.2 to be constructed, even though there wasn't yet (in the year 1900) a theory of atoms in any modern form.
10. *Albert Einstein* (1879-1955) published his *special theory of relativity* in 1905. It was inspired by Maxwell's theory, since Einstein was strongly motivated by a desire to unify particle motion and electromagnetism, which he realized could be accomplished only if both were described by Lorentz-invariant theories.

One could add to this list the *invention of quantum mechanics* in 1925-1926, which provided a firm basis for describing electromagnetism interacting with matter. But this book is about *classical electromagnetism,* and the quantum revolution that shook physics beginning in 1925 is primarily part of a *quantum theory of electromagnetism*, which merits a separate book and is necessarily beyond our present scope.

1.2 The Maxwell Equations

Thus, by the late 1800s the basics of classical electromagnetic theory could be summarized concisely in the *Maxwell equations*,[1] which may be written in free space as

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

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

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

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

[1] Maxwell's original formulation of what we now call the Maxwell equations was in terms of the vector and scalar potentials instead of the electric and magnetic fields (see Sections 2.6 and 8.6 for the relationship between the potentials and the fields), required 20 equations instead of four, and did not use the modern vector calculus notation of Eqs. (1.1), 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 the new vector calculus notation. One cannot overstate the importance of compact, concise, and evocative notion in the advancement of mathematical physics, and Ref. [18] may be consulted for the history of this important evolution in notation for electromagnetic equations.

using modern notation and SI units (see the discussion of units in Section 2.1), where

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

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

when expressed in SI units, where N is newtons and A is amperes. It should be noted that the SI-system constants ε_0 and μ_0 are *not independent* but are constrained by $\mu_0\varepsilon_0 = 1/c^2$, where c is the speed of light.[2]

An important property of the Maxwell equations is that they obey a set of symmetries, which are summarized in Box 1.1. In Ch. 12 we shall elaborate on the relationship of these symmetries to the conservation laws obeyed by classical electromagnetism.

1.3 Charge Conservation

The charge density ρ and the current vector $\boldsymbol{J}$ appearing in the Maxwell equations are not independent but are constrained by the *continuity equation,*

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot\boldsymbol{J} = 0 \qquad \text{(Continuity equation)}, \tag{1.3}$$

which ensures *conservation of charge* by requiring that variation of electrical charge in some arbitrary volume is caused by flow of electrical current through the surface of that volume. The local conservation of charge implied by Eq. (1.3) is not a separate feature, but is implied by the Maxwell equations themselves (see Problem 1.1). As indicated in Box 1.1, this conservation of charge by the equations of electromagnetism is associated with the *gauge symmetry of electromagnetism*, which will enter our later discussion in a number of forms.

1.4 Maxwell and the Displacement Current

A feature of the Maxwell equations with profound implications is that Maxwell realized that the original form of Ampére's law (which lacked the $\partial\boldsymbol{E}/\partial t$ term of Eq. (1.1d)) was inconsistent with Eq. (1.3) because it was valid only for stationary charge densities. As

[2] This relationship between ε_0 and μ_0 is a consequence of the units chosen for the electric field and magnetic field in the SI system.

Box 1.1 Symmetries of the Maxwell Equations

Symmetry plays a fundamental role in physics [17]. The Maxwell equations (1.1) exhibit some symmetries that have important implications for electromagnetism.

1. *Invariance under Space and Time Translations:* Invariance under translations in space and time is associated with conservation of momentum and energy, respectively. Satisfaction of these invariances by the Maxwell equations implies that *energy and momentum can be assigned 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. [17].
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 be consistent with the Maxwell equations if run backward or forward.
5. *Lorentz Invariance:* Electromagnetism isn't Galilean invariant but is Lorentz invariant (it is consistent with special relativity), while Newtonian mechanics is Galilean invariant but not Lorentz invariant (it is inconsistent with special relativity). The Maxwell equations (1.1) are *Lorentz covariant* (their validity is unaltered by Lorentz transformations). However, this isn't clear from the notation because space and time enter on an equal footing in special relativity and the use in Eqs. (1.1) of 3-vectors instead of 4-vectors, and of separate derivatives for space and time, obscures this covariance. Later we will re-write the Maxwell equations in a manifestly Lorentz-covariant form.
6. *Gauge Invariance:* Consistency of the Maxwell equations with Eq. (1.3) requires *charge to be conserved locally,* which implies *local gauge invariance of the electromagnetic field* (see Section 1.3). We shall have much to say about gauge symmetry and its far-reaching implications in later chapters.
7. *Electric and Magnetic Field Asymmetry:* The Maxwell equations exhibit a large *asymmetry* between electric and magnetic fields [most obvious in CGS units; see Appendix B.2]. If a magnetic charge $4\pi\rho_{\mathrm{m}}$ were added to the right side of Eq. (B.3c) and a magnetic current $(4\pi/c)\boldsymbol{J}_{\mathrm{m}}$ added to the right side of Eq. (B.3b), the Maxwell equations would become symmetric under $\boldsymbol{E} \to \boldsymbol{B}$ and $\boldsymbol{B} \to -\boldsymbol{E}$. However, *no magnetic charges have ever been observed*.

discussed in Section 11.1.1, 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.

1.5 Classical Electromagnetic Forces

The four Maxwell equations of Eqs. (1.1), supplemented by appropriate boundary conditions,[3] may be solved 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 tangible connection to experimental results in classical physics. This is remedied by introducing the empirically justified *Lorentz force law*,

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

which implies that the electromagnetic force acting on a particle having 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 second law relating momentum change to force,

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

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

1.6 Internal Consistency of Electromagnetism

Classical mechanics is internally self-consistent, in the sense that there are phenomena outside its domain (the very small requires quantum theory; the very fast requires special relativity; strong gravity requires general relativity), but it gives internally consistent results

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

when restricted to it domain of validity. Classical electromagnetism is different. In Eq. (2.45) we shall show that the energy U of a continuous charge distribution $\rho(\boldsymbol{x})$ is the work W required to assemble it,

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

where $\boldsymbol{E}$ is the electric field associated with the charge distribution and the integration is over all of space. But from Eq. (2.9) the electric field of a continuous charge distribution behaves as

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

and U diverges as $|\boldsymbol{x}-\boldsymbol{x}'|^{-4}$ upon approaching a point charge located at $\boldsymbol{x}'$. This infinite *self-energy* of a classical point charge may be viewed as resulting from the charge acting on itself and is clearly unphysical; the first thought is that it must be excluded from the energy of a charge distribution. But further problems arise if one considers *accelerated charges*, which radiate electromagnetic waves and must lose energy; this can be thought of most simply as a radiative damping force acting on the charged particle, but this again forces us to think of the charge acting on itself, with the short-range limit diverging.

We conclude that there is no consistent way to deal with such self-energies in classical electromagnetism and the theory is internally inconsistent.[4] However, as a practical matter we may ask at what order of magnitude for distance and time scales does one expect such inconsistencies lead to actual problems? Consider the self-energy of an electron. As far as is known experimentally elementary particle like electrons are point-like, without spatial extent. But if the electron is assumed to have a finite radius $r_{\rm e}$, the electrostatic energy becomes of order $e^2/r_{\rm e}$, where e is the electron charge. The only other energy scale in the problem is the electron restmass energy $m_{\rm e}c^2$, and if we equate these two energy scales and solve for $r_{\rm e}$,

$$r_{\rm e} = \frac{e^2}{mc^2} = 2.82\times10^{-13}\ \text{cm}, \tag{1.6}$$

which is known as the *classical radius of the electron*. (In comparison, the Bohr radius of the hydrogen atom is 0.529×10^{-8} cm.) A corresponding timescale is given by the time for light to travel this distance,

$$\tau_e = \frac{r_{\rm e}}{c} = \frac{e^2}{mc^3} \simeq 10^{-21}\ \text{s}, \tag{1.7}$$

Unless we consider motion of charges that changes over a length scale of order $r_{\rm e}$, or a timescale of order $\tau_{\rm e}$, the infinities associated with self energies are of no practical significance. In fact, one expects quantum effects to become important on a length given by the

[4] Similar problems with infinities associated with matrix elements evaluated at a point (such as the infinite self-energies of point particles) arise in quantum field theory. There a scheme called the *renormalization program* has been developed to systematically subtract away the infinities in observable quantities and bury them in quantities that are not observable. Not everyone is comfortable philosophically with the "sweeping under the rug" nature of the quantum renormalization scheme, but everyone agrees that renormalization succeeds in removing infinities from observables, and that the resulting (finite) predictions of quantum field theory are in spectacular agreement with experimental observables.

Compton wavelength $r_C = \hbar/mc$, which is

$$\frac{r_C}{r_e} = \frac{\hbar/mc}{e^2/mc^2} = \frac{\hbar c}{e^2} \simeq 137$$

times larger than the classical electron radius. We conclude that inconsistencies associated with self-energies of charged particles are irrelevant for our discussion.

1.7 Classical Electromagnetic Solutions

Thus the understanding of classical electromagnetism can be summarized in a succinct admonition: to be blunt, "shut up and calculate", or to be less blunt and more precise,

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

While this admonition is not wrong, it risks 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 been altered dramatically by modern advances in quantum field theory (which were often inspired and guided by the theoretical understanding of classical electrodynamics).

The following chapters 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 more broadly in the context of modern theoretical physics.

1.8 Electromagnetism in Modern Physics

Let us make some general remarks about the relationship of classical electromagnetism with modern theoretical physics.

1.8.1 Gauge Symmetry

One of the remarkable findings of modern physics is that the already beautiful edifice of Maxwell's equations for electromagnetism that we shall elaborate in these lectures hides within it a deceptively powerful symmetry called *(local) gauge invariance*—in essence a

symmetry under local phase transformations—that was introduced in Box 1.1 and, 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.

In quantum field theory local gauge symmetries are generated by a set of quantum operators. If the quantum operators all commute among themselves, the gauge symmetry is said to be *abelian*; if they do not all commute, the gauge symmetry is said to be *non-abelian* (see the discussion of symmetries and group theory in Box 16.1). 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*). In Ch. 18 we shall explore in more depth this view of classical electromagnetism as an abelian gauge field theory and its relationship to quantum electrodynamics. 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*.

1.8.2 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

$$e \equiv |q_e| = 1.60217733 \times 10^{-19}\,\mathrm{C} \qquad \text{[in SI units]}. \tag{1.8}$$

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, so the very existence of ourselves and the visible Universe is strong evidence for quantization of electrical charge.

> Dirac proposed long ago that, if fundamental magnetic charges (magnetic monopoles) existed, there could be a topological reason for charge quantization. However, no reproducible observations of magnetic monopoles have ever been reported, so Dirac's idea remains only conjecture. As we have noted in Box 1.1, if magnetic monopoles did exist the Maxwell equations (1.1) would require modifications that would make them more symmetric with respect to electric and magnetic fields.

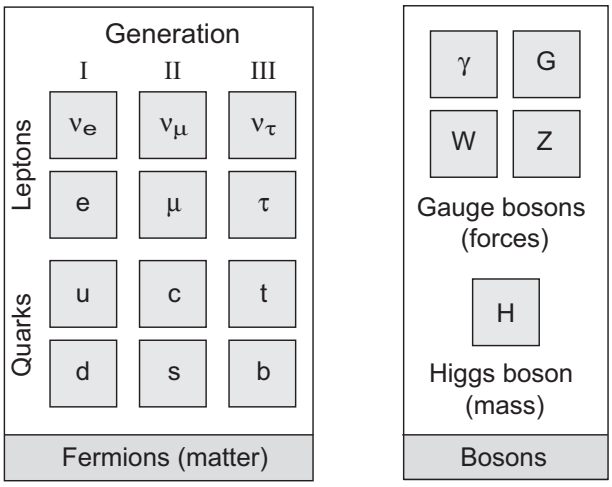


Fig. 1.1

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 in this diagram.

Since there is no evidence for the existence of free magnetic charge (magnetic monopoles), we must conclude that there is presently no convincing fundamental explanation for the observed quantization of electrical charge in integer multiples of the electron charge.

1.8.3 Charges of Known Elementary Particles

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 all matter is formed from fermions $(e, \nu_e, u, \ldots)$, while interactions between elementary particles are mediated by the exchange of gauge bosons $(\gamma, G, W^{\pm}, Z^0)$, and masses for bare (that is, non-interacting) particles arise from couplings to the Higgs boson H. For reasons that remain elusive, 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. To give one 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, the (electrical) charge Q of each quark is given by

Table 1.1 Quantum number assignments for quarks [14]

	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)	$\frac{2}{3}e$	$-\frac{1}{3}e$	$-\frac{1}{3}e$	$\frac{2}{3}e$	$-\frac{1}{3}e$	$\frac{2}{3}e$
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 $Q/e = T_3 + \frac{1}{2}(B+S+c+b+t)$, where e the absolute value of the electron charge.

$$Q/e = T_3 + \frac{1}{2}(B+S+c+b+t), \tag{1.9}$$

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

Example 1.1 From Eq. (1.9) and Table 1.1, the charge of the down quark d is

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

This quantization of charge in integer multiples of $\pm\frac{1}{3}$ of the absolute value of the electron charge 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,*[5] and

[5] More precisely, the gauge theory of the strong interactions, quantum chromodynamics (QCD), is thought to exhibit *color confinement:* particles carrying the strong-interaction gauge charge (called whimsically "color", with no relationship to the usual meaning of color) are *confined* to the interior of the hadrons such as neutrons and protons, and can never appear as free particles. Finding solutions for the equations of QCD is extremely difficult because it is a strongly interacting, highly non-linear field theory in the limit that would lead to confinement. Thus it is not simple to prove that QCD is color-confining. However, modern large-scale computer

indeed no free particles with fractional charges have ever been observed reproducibly in experiments.

Example 1.2 A proton has a *udu* quark structure (two up quarks and one down quark) and charge is an additive quantum number, so the charge of the proton is the sum of the charges of its quarks. From Table 1.1, the charge of a proton is

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

in terms of the fundamental charge unit e given by Eq. (1.8), while a neutron has a *udd* quark structure and

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

so neutrons have zero net electrical charge.

Thus both neutrons and protons carry integer charges, even though their constituent quarks have fractional charges.

Modern high energy physics proposes a variety of elementary particles, some of which have electrical charges that are fractions of the fundamental unit (1.8) set by the charge on the electron. However, the *physically observable particles* entering into classical electromagnetism (and into relativistic quantum field theory) all appear to have electrical charges that are integer multiples of the fundamental charge unit given by Eq. (1.8).

Hence we shall develop the theory of classical electromagnetism assuming that charged particles carry an electrical charge quantized in integer units of the electron charge, even though we presently have scant fundamental explanation for why this should be so.

Background and Further Reading

The history of classical electromagnetism is summarized in various parts of Jackson [19]. A view of classical electromagnetism with emphasis on the relationship of modern quantum field theory to classical electromagnetism may be found in Wald [40]. 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. [14, 17].

simulations of QCD (a discipline called *lattice gauge theory*) have increasingly converged on the conclusion that QCD exhibits confinement of particles like quarks carrying a color charge.

Problems

1.1 Prove that Maxwell's equations (1.1) are consistent with the continuity equation (1.3) that ensures local conservation of electrical charge.

1.2 Use Eq. (1.9) and Table 1.1 to confirm that the top quark is expected to carry an electrical charge of $\frac{2}{3}$. Should you worry about seeing a particle with charge $\frac{2}{3}$ in your experimental apparatus? ***

1.3 A family of subatomic particles called the Δ-resonance has four distinct members $(\Delta^{++}, \Delta^{+}, \Delta^{0}, \Delta^{-})$, which have a valence quark stucture (uuu, uud, udd, ddd), respectively, where u is the up quark and d is the down quark, with properties summarized in Table 1.1. Use this table to confirm the electrical charge assigments

$$Q(\Delta^{++}) = +2e \qquad Q(\Delta^{+}) = +1e \qquad Q(\Delta^{0}) = 0\,e \qquad Q(\Delta^{-}) = -1e$$

for these particles. Thus, show that the four members of the Δ-resonance (each of which is observable in high-energy physics experiments) have charges that are integer multiples of e, even though each consists of valence quarks having charges that are fractions of e (which are not observable because of color confinement for quarks).

2 Electrostatics in Vacuum

The fundamental problem that is to be solved in electrodynamics is illustrated schematically in Fig. 2.1(b). 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 basics and consider the simpler case of *electrostatics,* where the source charges are assumed to be fixed in spatial position (but the test charge may move). Electrostatics problems can be divided into two broad categories [34, 42]:

- summation problems, and
- boundary value problems

In a *summation problem* the charge density $\rho(\boldsymbol{x})$ is specified initially at every point in space at a given time, reducing the task of solving the electrostatics problem to evaluating an integral. The integral might be easy or difficult to evaluate but conceptually a summation problem has a clearly defined solution. If the charge density cannot be specified initially at every point in space, one is dealing with a more challenging *boundary value problem*. This will be the case if matter in any form is present, because the Coulomb force [see Eq. (2.1)] will cause charge to redistribute itself in the matter until equilibrium is reached.[1] It is remarkable that the electrostatics problem can be solved uniquely for every spatial point even in this latter case, provided that a model of the polarizable matter is specified, with an adequate model specifying both the behavior of the electric field in matter, and boundary or matching conditions for the Maxwell equations. This chapter will introduce summation problems for electrostatics, while Ch. 3 will address boundary value problems and their solutions.

2.1 Coulomb's Law

Let us assume initially that the source charges and the test charges in Fig. 2.1(b) are embedded in vacuum, so that it isn't necessary to worry about a polarized medium altering the interactions between charges (that will be the subject of extensive discussion in later chapters). Let us now consider a quantitative description of this idealized electrostatics problem, utilizing experimental data, physical intuition, and mathematics for guidance.

Consider first the force acting between two isolated charges at rest with respect to each

[1] For historical reasons, in conductors this charge redistribution is called *electrostatic induction* and in nonconductors it is called *electrostatic polarization*.

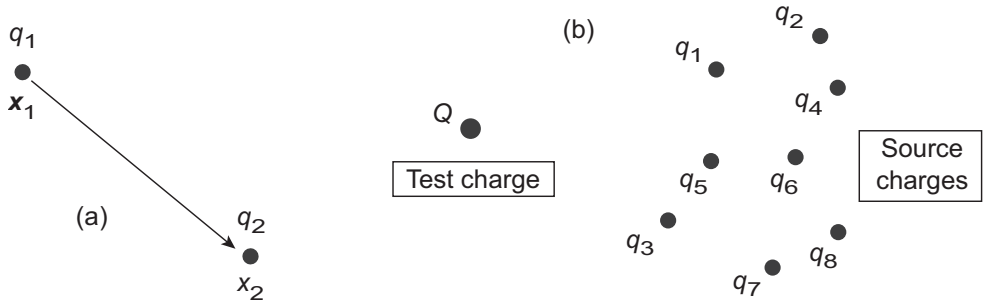


Fig. 2.1 (a) Coulomb interaction between two isolated and stationary test charges. (b) The prototype electrostatics problem: interaction in vacuum of a test charge Q with a set of stationary source charges q_n.

other, as illustrated in Fig. 2.1(a). 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*,

$$\boldsymbol{F} = kq_1q_2\,\frac{\boldsymbol{x}_1-\boldsymbol{x}_2}{|\boldsymbol{x}_1-\boldsymbol{x}_2|^3} \qquad \text{(Coulomb's law)}, \tag{2.1}$$

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 appearing in Coulomb's law depends on the system of units that is in use. Three common choices are *the SI system*, *Gaussian or electrostatic units*, and *Heaviside–Lorentz* units.

1. 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

$$\varepsilon_0 \simeq 8.85\times10^{-12}\,\frac{\mathrm{C}^2}{\mathrm{N\ m}^2} = 8.85\times10^{-12}\,\frac{\mathrm{F}}{\mathrm{m}}, \tag{2.3}$$

where the farad (F) is the derived SI unit of electrical capacitance. The constants μ_0, ε_0, and the speed of light c that may appear in SI units are related by[2]

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

[2] The three constants that enter into 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 Eq. (2.4). This has the potentially confusing implication that the form of equations written in SI units can be changed by using $\mu_0\varepsilon_0c^2 = 1$ to interchange constants appearing in them. As will be discussed in Section 8.2, SI units also partially obscure that under typical laboratory conditions magnetic phenomena are intrinsically much weaker than electrostatic phenomena, but that the two become comparable in strength if the characteristic speeds of charged particles approach that of light.

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

$$\boldsymbol{F} = \frac{q_1 q_2}{4\pi\varepsilon_0} \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{|\boldsymbol{x}_1 - \boldsymbol{x}_2|^3} \qquad \text{(Coulomb's law, SI units).} \tag{2.5}$$

The Maxwell equations and the Lorentz force law in SI units are summarized in Appendix B.1. We will use SI units in these lectures by default, specifying explicitly if we choose other units in a few limited situations.

2. In *Gaussian* or *CGS units*, also termed *electrostatic units (esu)*, $k = 1$ 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 Gaussian system the unit charge is called a *statcoulomb*. The Maxwell equations and the Lorentz force law are expressed in Gaussian units in Appendix B.2. There are some advantages to Gaussian units and many older books use them often, but more modern books tend to use SI units because of the affinity with standardized MKS units for non-electromagnetic quantities.
3. In some disciplines (for example, elementary particle physics) *Heaviside–Lorentz units* may be used, where $k = 1/4\pi$. The Maxwell equations and the Lorentz force law are given in Heaviside–Lorentz units in Appendix B.3.

In some cases one may encounter specialized variations on units where fundamental constants like the speed of light c or Planck's constant $\hbar$ (in quantum problems) are set to one, though that is probably more common in quantum field theory than in classical field theory.

Beginning with the time of Cavendish and Coulomb, the empirical validity of Coulomb's law (2.5) has been tested by a variety of techniques applied over a range of distance scales. As discussed in Box 2.1, well-established experimental limits suggest that for classical electromagnetism we can safely assume the photon to be massless and the deviation from Coulomb's law to be negligible.

2.2 The Electric Field

In experiments investigating 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 abstracting broader implications from measurements. Perhaps no abstraction has been more powerful in the development of physics than that of a *field*, which is simply an instance of something that is defined at every point of spacetime. Let us introduce an *electric field* $\boldsymbol{E}$ acting on a test charge q by *defining* the force $\boldsymbol{F}$ acting on the test charge to be

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

Thus an electric field $\boldsymbol{E} = \boldsymbol{F}/q$ may be interpreted as an object that can exert a force per unit test charge for a test charge located at any point in in spacetime. The electric field is a function of position, but *it doesn't depend on whether there is a test charge located at P*. The field exists, independent of whether there is a test charge exists upon which it acts. Since force is a vector and charge is a scalar, the electric field is a *vector field* $\boldsymbol{E}(t,\boldsymbol{x})$ defined

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 those early experiments, 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 photon mass (zero for an inverse square force). Then deviations from the inverse square law may be reported as an upper limit on μ or 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 $\varepsilon = (2.7 \pm 3.1) \times 10^{-16}$ [41].
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 [12, 19, 22].

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.

at each point of spacetime $(t, \boldsymbol{x})$, but we ignore any time dependence for now. 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. In the SI system the unit of charge is the coulomb (C) and the electric field $\boldsymbol{E}$ has units of volts per meter. The importance of the field concept in the development of modern physics is elaborated in Box 2.2.

2.3 The Principle of Superposition

Now let us address the more general case in Fig. 2.1(b) 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 verified fact that

Box 2.2 Significance of Fields in Electromagnetism

Equation (2.6) suggests measuring forces to infer fields, implying that fields are derivative. 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, which are the central concepts of classical electromagnetism. The importance of fields is most obvious in relativistic quantum field theory (QFT) , which is not our subject here, but the field concept traces historically to the introduction of electric and magnetic fields in classical electromagnetism, which is our subject.

Action at a Distance

Originally it was believed that forces associated with charges, currents, and magnets acted *instantaneously over any distance* (this was termed *action at a distance*).[a] The modern view—shaped by experimental measurement and the development of relativistic QFT—is that fields are every bit as fundamental as particles. In this picture, forces are mediated by fields, and the lightspeed limit of 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

For example, if a charge moves the fields associated with the charge change, but the 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 those changes can propagate only at a finite speed (less than or equal to the speed of light).

This abstract view was resisted initially because fields could exist in vacuum, and did not describe tangible matter. Modern physics takes quite a different view. The introduction of electric and magnetic fields allows a simple mathematical description of electromagnetic phenomena, but the fields are *not just mathematical abstractions;* they are *real physical entities* ("as real as a rinoceros" [11]), carrying concrete physical properties such as energy, momentum, and angular momentum. This is most clear in QFT, where these physical attributes become properties of *quanta of the field* (photons) and electromagnetic interactions are viewed as being mediated by exchange of virtual photons. These photons associated with the electromagnetic field have observable consequences: at high enough energy the collision of two photons can produce matter in the form of an electron and positron pair; real as a rinoceros indeed! Relativistic QFT implies that all non-gravitational fundamental interactions (electromagnetic, strong, and weak) are mediated by the gauge bosons of Fig. 1.1, which are the quanta of fields that generalize the gauge symmetry of photons. Such is the rich legacy of classical electromagnetic fields.

[a] Likewise, Newtonian physics assumes gravity to act instantaneously on a distant object. Replacement of Newtonian gravity with general relativity (a field theory consistent with the limiting speed of light) eliminated action at a distance for gravity, just as relativistic QFT eliminated it for electromagnetism.

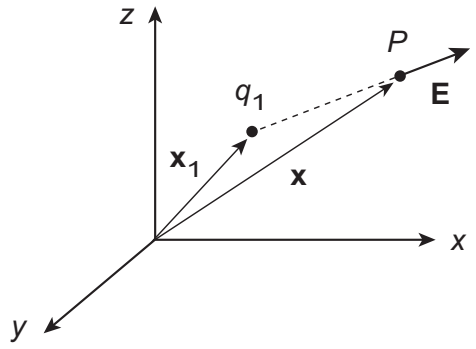


Fig. 2.2

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

1. if the *charged-particle interactions are not too large*, and
2. if *quantum effects can be ignored*,

then 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 traced to the linear dependence of the Maxwell equations (1.1) on 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) 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 we have used Eq. (2.7) expressed in SI 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

[3] In the presence of strong electric fields such as those generated by powerful lasers, and in various *vacuum polarization phenomena* observed in quantum field theory experiments, this linear superposition principle may fail. However, we shall restrict ourselves to the *classical linear regime*, where such effects may be neglected by hypothesis. It is fortunate for development of electromagnetic theory that the experiments that laid the foundations for 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.

sum over discrete charges in Eq. (2.8) may be replaced by an integral over a continuous charge distribution,

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \rho(\boldsymbol{x}')\,\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x' = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{R^2}\,\hat{\boldsymbol{R}}\,d^3x' \tag{2.9}$$

where in the last step a compact notation

$$\boldsymbol{R} \equiv \boldsymbol{x}-\boldsymbol{x}' \qquad R = |\boldsymbol{R}| = \left|\boldsymbol{x}-\boldsymbol{x}'\right| \qquad \hat{\boldsymbol{R}} = \frac{\boldsymbol{R}}{|\boldsymbol{R}|} = \frac{\boldsymbol{R}}{R} = \frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|} \tag{2.10}$$

such that

$$\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3} = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|^2}\cdot\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|} = \frac{1}{R^2}\,\hat{\boldsymbol{R}}, \tag{2.11}$$

has been introduced that will prove useful in solving problems. We will generally be working with continuous charge distributions in one, two, or three spatial dimensions.

1. If the charge is spread along a line in 1D with a charge per unit length of λ, then the volume element in Eq. (2.9) becomes a *line charge* $\lambda\,dl'$, where dl' is an infinitesimal length element.
2. If the charge is spread over an area in 2D with a charge density per unit area of σ, the volume element becomes a *surface charge* $\sigma\,da'$, where da' is a differential element of area.
3. If the charge is spread over a volume in 3D with charge per unit volume ρ, the volume element is a *volume charge* $\rho\,d^3x'$, which we sometimes abbreviate as $\rho\,d\tau'$.

Thus, for continuous charge distributions in three, two, and one dimensions, respectively, the electric field is given explicitly by

$$\boldsymbol{E}(\boldsymbol{x})_{3\mathrm{D}} = \frac{1}{4\pi\varepsilon_0}\int \rho(\boldsymbol{x}')\,\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x' = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{R^2}\,\hat{\boldsymbol{R}}\,d^3x', \tag{2.12a}$$

$$\boldsymbol{E}(\boldsymbol{x})_{2\mathrm{D}} = \frac{1}{4\pi\varepsilon_0}\int \sigma(\boldsymbol{x}')\,\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,da' = \frac{1}{4\pi\varepsilon_0}\int \frac{\sigma(\boldsymbol{x}')}{R^2}\,\hat{\boldsymbol{R}}\,da', \tag{2.12b}$$

$$\boldsymbol{E}(\boldsymbol{x})_{1\mathrm{D}} = \frac{1}{4\pi\varepsilon_0}\int \lambda(\boldsymbol{x}')\,\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,dl' = \frac{1}{4\pi\varepsilon_0}\int \frac{\lambda(\boldsymbol{x}')}{R^2}\,\hat{\boldsymbol{R}}\,dl', \tag{2.12c}$$

in terms of volume densities $\rho(\boldsymbol{x}')$, surface densities $\sigma(\boldsymbol{x}')$, and line densities $\lambda(\boldsymbol{x}')$, respectively.

Example 2.1 Let us determine the electric field produced by a straight line charge with uniform charge density λ. First consider the field produced at the point P of Fig. 2.3 if the line charge is of finite length $2L$. As shown in Problem 2.10, from Eq. (2.12c) the electric

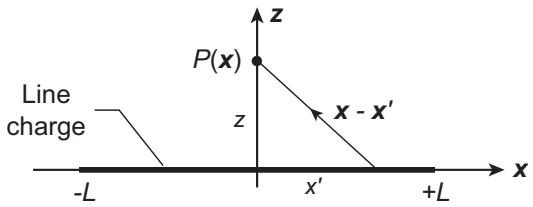


Fig. 2.3 Charged line segment of length $2L$ used to calculate the electric field in Example 2.1.

field produced at the point P is

$$\begin{aligned}\boldsymbol{E} &= \frac{1}{4\pi\varepsilon_0}\int_{-L}^{+L}\frac{\lambda}{R^2}\hat{\boldsymbol{R}}\,dx' \\ &= \frac{1}{4\pi\varepsilon_0}\int_{-L}^{+L}\frac{\lambda}{x^2+z^2}\cdot\frac{z\hat{\boldsymbol{z}}-x\hat{\boldsymbol{x}}}{\sqrt{x^2+z^2}}\,dx \\ &= \frac{1}{2\pi\varepsilon_0}\frac{\lambda L}{z(z^2+L^2)^{1/2}}\,\hat{\boldsymbol{z}},\end{aligned} \tag{2.13}$$

where the compact notation of Eq. (2.10) has been used. If we identify this charge segment with a straight wire, the electric field produced by a wire of infinite length can be obtained by taking the limit $L\to\infty$ of Eq. (2.13), which gives

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda}{z}\,\hat{\boldsymbol{z}} \tag{2.14}$$

for the electric field produced by an infinitely long straight wire oriented along the x axis. The electric field is perpendicular to the wire, with magnitude proportional to the inverse of the distance from the wire.

Finally, before leaving this discussion of continuous charge distributions let us note that a discrete set of charges also can be described by a continuous charge distribution $\rho(\boldsymbol{x})$ if the *Dirac delta function* $\delta(\boldsymbol{x}-\boldsymbol{X})$ described in Appendix A.12 is used to pick out the locations of each of the discrete charges q_i,

$$\rho(\boldsymbol{x}) = \sum_{i=1}^{n} q_1\,\delta(\boldsymbol{x}-\boldsymbol{X}), \tag{2.15}$$

since substitution of the charge density (2.15) in Eq. (2.9) and integrating using the properties of the delta function is equivalent to 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 way to solve for

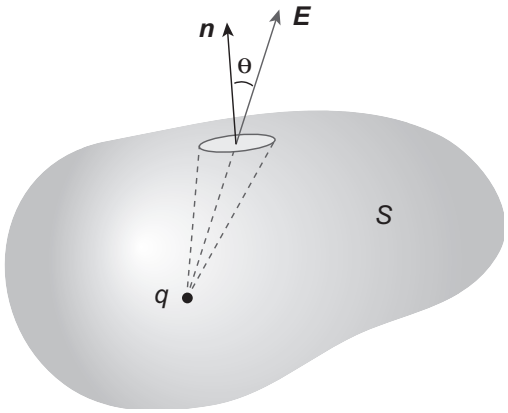


Fig. 2.4 Closed surface S illustrating Gauss's law. The electric field generated by the charge q is $\boldsymbol{E}$, the normal vector to the surface is $\boldsymbol{n}$, and da is an element of surface area.

the electric field. If a problem has some degree of symmetry, often another integral result called *Gauss's law* can lead to an easier solution. In Fig. 2.4 we indicate a charge q enclosed by a surface S. The normal component of $\boldsymbol{E}$ times a surface area element da is

$$\boldsymbol{E}\cdot\boldsymbol{n}\,da = \frac{q}{4\pi\varepsilon_0}\frac{\cos\theta}{r^2}\,da = \frac{q}{4\pi\varepsilon_0}\,d\Omega, \tag{2.16}$$

where the last step used $\cos\theta\,da = r^2\,d\Omega$, 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 S,

$$\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da = \begin{cases} q/\varepsilon_0 & (\text{for } q \text{ inside } S) \\ 0 & (\text{for } q \text{ outside } S) \end{cases} \qquad (\text{Gauss's law}), \tag{2.17}$$

where $\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da$ is called the *flux* through the surface S. Equation (2.17) is *Gauss's law in integral form* for a single charge. For a set of discrete charges Gauss's law can be generalized to the form,

$$\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da = \frac{1}{\varepsilon_0}\sum_i q_i, \tag{2.18}$$

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

$$\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da = \frac{1}{\varepsilon_0}\int_V \rho(\boldsymbol{x})\,d^3x, \qquad (\text{Gauss's law}) \tag{2.19}$$

where the integration is over the volume V contained within the surface S.[4] Gauss's law in the form of Eqs. (2.17)-(2.19) may be viewed as integral formulation of the law of

[4] Gauss's law (2.19) is a consequence of (1) inverse-square forces between charges, (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.)

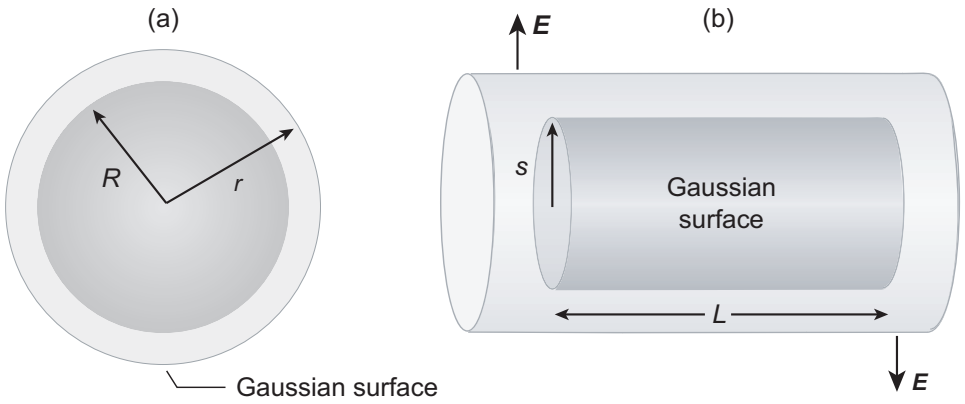


Fig. 2.5 Examples of applying Gauss's law to find the electric field. (a) A ball of uniformly distributed charge with radius R, used in Example 2.2. (b) A cylinder carrying a charge density proportional to the distance s from the cylindrical axis, used in Example 2.3. Gauss's law is often effective for examples like this that have a high degree of symmetry.

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,

$$\oint_S \boldsymbol{A}\cdot\boldsymbol{n}\,da = \int_V \boldsymbol{\nabla}\cdot\boldsymbol{A}\,d^3x, \tag{2.20}$$

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 $\boldsymbol{A}$.

The divergence theorem (2.20) allows Eq. (2.19) 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

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = \frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)}, \tag{2.21}$$

which is the *differential form of Gauss's law* (2.19) 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 [13].

Example 2.2 Figure 2.5(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 (2.18)

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

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

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

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 2.3 Consider Fig. 2.5(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 surface in the form of a cylinder of radius s and length L, as illustrated in Fig. 2.5(b). From Eq. (2.19), 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

[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).

[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. This similarity is of course because both Newtonian gravity and Coulomb's law correspond to inverse-square forces.

(Appendix A.7.2) with a cylindrical volume element $d\tau = s\,ds\,d\phi\,dz$

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

By symmetry $\boldsymbol{E}$ must point radially outward from the cylinder's central axis, as indicated in Fig. 2.5(b), so for the curved portion of the Gaussian cylinder

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

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

$$2\pi sL|\boldsymbol{E}| = \frac{1}{\varepsilon_0}\frac{2}{3}\pi kLs^3$$

and the electric field is given by

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

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

Quite generally, one sees from such examples that using Gauss's theorem to determine the electric field is a useful approach when there is a high degree of symmetry that can be exploited to simplify the evaluation of integrals. Typically, the goal in applying Gauss's law is to simplify the integral on the left side of Eqs. (2.18) or (2.19). 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, in the problems solved above the third condition was used in Example 2.2 and the third and fourth conditions were used in Example 2.3. If none of these conditions are satisfied, Gauss's law is still valid but there may be easier ways to determine the electric field.

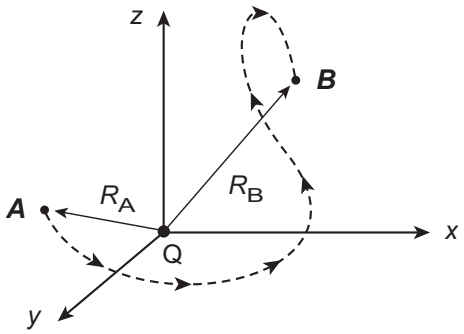


Fig. 2.6 Path for a line integral in an electric field that is generated by a charge Q located at the origin.

2.5 The Scalar Potential

Consider a line integral between two points $\boldsymbol{A}$ and $\boldsymbol{B}$ in a field generated by a single charge at the origin, as illustrated in Fig. 2.6. In spherical coordinates (Appendix A.7.1) the electric field $\boldsymbol{E}$ and line element $d\boldsymbol{l}$ are

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\boldsymbol{r}} \qquad d\boldsymbol{l} = dr\,\hat{\boldsymbol{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\phi\,\hat{\boldsymbol{\phi}}, \tag{2.22}$$

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

$$\begin{aligned}\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l} &= \frac{1}{4\pi\varepsilon_0}\int_{\boldsymbol{A}}^{\boldsymbol{B}}\frac{Q}{r^2}\,dr \\ &= \frac{-Q}{4\pi\varepsilon_0 r}\Bigg|_{R_A}^{R_B} \\ &= \frac{1}{4\pi\varepsilon_0}\left(\frac{Q}{R_A}-\frac{Q}{R_B}\right).\end{aligned} \tag{2.23}$$

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

$$\oint \boldsymbol{E}\cdot d\boldsymbol{l} = 0, \tag{2.24}$$

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

$$\int_S (\boldsymbol{\nabla}\times\boldsymbol{A})\cdot\boldsymbol{n}\,da = \oint_C \boldsymbol{A}\cdot d\boldsymbol{l} \qquad \text{(Stokes' theorem)}, \tag{2.25}$$

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.

Box 2.3 **Geometrical Interpretation of Stokes' Theorem**

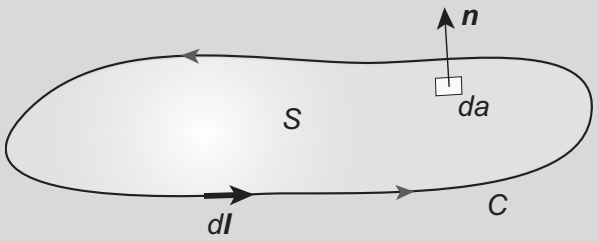
For a vector field $\boldsymbol{A}$, *Stokes' theorem* relates a surface integral of the curl vector field $\boldsymbol{\nabla}\times\boldsymbol{A}$ to a line integral around the (smooth) boundary of that surface:

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

where S is the 2D surface enclosed by the 1D boundary C, the outward normal to the surface is $\boldsymbol{n}$, and the curl $\boldsymbol{\nabla}\times\boldsymbol{A}$ may be expressed explicitly by

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

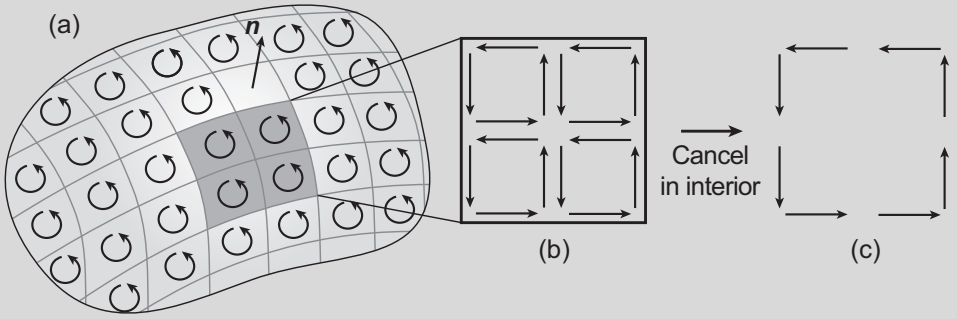
in cartesian coordinates. The orientation of the surface $\boldsymbol{n}$ and the direction of the integration path around the boundary curve that are illustrated in the following figure



are related by the following *right-hand rule*:

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

The physical content of Stokes' theorem may be understood geometrically if one divides the surface up into plaquettes:



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.

This implies that

$$\int_S (\nabla \times \boldsymbol{E}) \cdot \boldsymbol{n}\, da = \oint_P \boldsymbol{E} \cdot d\boldsymbol{l} = 0, \tag{2.26}$$

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

$$\nabla \times \boldsymbol{E} = \boldsymbol{0}. \tag{2.27}$$

Thus, we have shown that *the curl of an electric field* $\boldsymbol{E}$ *is zero*.[7]

Because of Eq. (2.27) 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. 2.6 has *the same value for all possible paths*. Thus we can define a function $\Phi(\boldsymbol{x})$ by

$$\Phi(\boldsymbol{x}) = -\int_{\mathscr{O}}^{\boldsymbol{x}} \boldsymbol{E}(\boldsymbol{x}) \cdot d\boldsymbol{l}, \tag{2.28}$$

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

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{r}\right), \tag{2.29}$$

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

$$\Phi(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{n}\frac{q_i}{r_i}. \tag{2.30}$$

For a continuous charge distribution in one, two, or three spatial dimensions the potential evaluates to

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x' \qquad \text{(3D volume charge)}, \tag{2.31a}$$

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, da' \qquad \text{(2D surface charge)}, \tag{2.31b}$$

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\lambda(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, dl' \qquad \text{(1D line charge)}, \tag{2.31c}$$

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 a special vector field because it has vanishing curl. We shall now use this to reduce finding the electric field, which is a vector problem, to a simpler scalar

[7] Note that this is true for *static electric fields*, but may no longer hold for time-dependent problems. For example, see Eq. (10.9).

problem. The *potential difference* between points $\boldsymbol{A}$ and $\boldsymbol{B}$ in Fig. 2.6 is given by

$$\begin{aligned}\Phi(\mathrm{B})-\Phi(\mathrm{A}) &= -\int_{\mathscr{O}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l}+\int_{\mathscr{O}}^{\boldsymbol{A}} \boldsymbol{E}\cdot d\boldsymbol{l} \\ &= -\int_{\mathscr{O}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l}-\int_{\boldsymbol{A}}^{\mathscr{O}} \boldsymbol{E}\cdot d\boldsymbol{l} \\ &= -\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l}. \end{aligned} \tag{2.32}$$

Applying the fundamental theorem for gradients (A.27) to the left side gives,

$$\Phi(\boldsymbol{B})-\Phi(\boldsymbol{A})=\int_{\boldsymbol{A}}^{\boldsymbol{B}} (\boldsymbol{\nabla}\Phi)\cdot d\boldsymbol{l}, \tag{2.33}$$

and therefore from Eqs. (2.32) and (2.33),

$$\int_{\boldsymbol{A}}^{\boldsymbol{B}} (\boldsymbol{\nabla}\Phi)\cdot d\boldsymbol{l}=-\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l}. \tag{2.34}$$

But since Eq. (2.34) must be true for arbitrary points $\boldsymbol{A}$ and $\boldsymbol{B}$, the integrands of the integrals on the two sides of Eq. (2.34) must be equal and we obtain

$$\boldsymbol{E}=-\boldsymbol{\nabla}\Phi, \tag{2.35}$$

which is a differential version of Eq. (2.28).

> The electric field vector $\boldsymbol{E}$ may be obtained by taking minus the 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. Then the potential Φ has units of newton-meters per coulomb, or joules per coulomb, where a joule per coulomb is defined to be a volt.

A major advantage of the potential formulation is that it is usually easier to construct the scalar potential than the electric field, but if one knows the potential Φ, the electric field can be obtained simply by taking the gradient, as in Eq. (2.35). 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 set of restrictions placed on the components $\boldsymbol{E}$ by the vanishing of the curl in Eq. (2.27), since expansion of $\boldsymbol{\nabla}\times\boldsymbol{E}=0$ leads to the constraint equations

$$\frac{\partial E_x}{\partial y}=\frac{\partial E_y}{\partial x} \qquad \frac{\partial E_z}{\partial y}=\frac{\partial E_y}{\partial z} \qquad \frac{\partial E_x}{\partial z}=\frac{\partial E_z}{\partial x}, \tag{2.36}$$

as you are asked to show in Problem 2.5.

Changing the arbitrary reference point $\mathscr{O}$ appearing in Eq. (2.28) shifts the potential by a constant amount, implying that there is an essential ambiguity in defining Φ. It follows that 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 $\mathscr{O}$, and

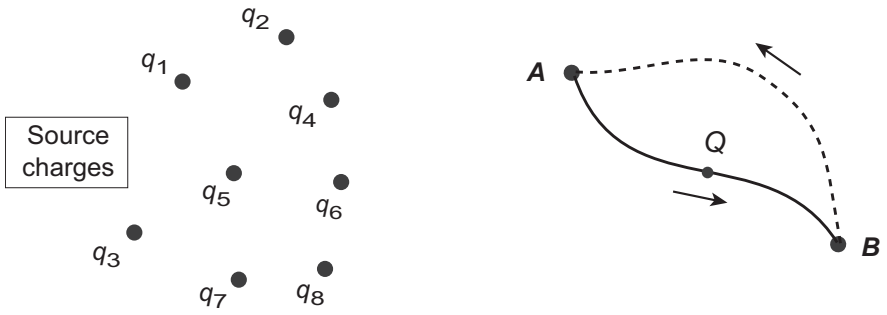


Fig. 2.7 Moving an electrical test charge Q on a path between endpoints $\boldsymbol{A}$ and $\boldsymbol{B}$ in an electric field generated by a static set of source charges. The work done is independent of the path, depending only on the endpoints $\boldsymbol{A}$ and $\boldsymbol{B}$, indicating that the electric field is *conservative*. Alternatively, if the path is closed by the dashed curve from $\boldsymbol{B}$ to $\boldsymbol{A}$, the work done by the electric field over the closed path $\boldsymbol{A}$ to $\boldsymbol{B}$ and back to $\boldsymbol{A}$ is identically zero.

2. *gradients of the potential* $\nabla\Phi$ are unaffected by shifting the reference point a constant amount, since the derivative of a constant is zero.

Thus, potential differences and potential gradients have physical meaning. The selection of the reference point $\mathscr{O}$ for the potential is in principle arbitrary, but the most common choice in electrostatics is to take $\mathscr{O}$ to be an infinite distance away from the (assumed localized) charges, where the potential is taken to drop to zero.[8]

2.7 Work and Energy for 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 on the test charge by the field produced by the source charges?*

2.7.1 Work Required 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

$$W = \int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{F}\cdot d\boldsymbol{l} = -Q\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l} = Q\left[\Phi(\boldsymbol{B}) - \Phi(\boldsymbol{A})\right], \tag{2.37}$$

[8] 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 in real-world problems charge distributions are typically bounded spatially.

where Eq. (2.32) has been used. Because the work was done against an electric field, it is *independent of path.* An alternative statement is that over a closed path

> Therefore the electrostatic force is *conservative,* meaning that the work depends only on the *difference in potentials between the endpoints of the path* and not on the details of the path followed. An alternative statements is that the work done on a closed path in a field is identically zero if the field is conservative.

If we wish to move the charge from an infinite distance away to a point $\boldsymbol{B}$,

$$W = Q\,[\Phi(\boldsymbol{B}) - \Phi(\infty)], \tag{2.38}$$

so the work done in moving a test charge from infinity to a point $\boldsymbol{x} \equiv \boldsymbol{B}$ is

$$W = Q\,\Phi(\boldsymbol{x}), \tag{2.39}$$

if the standard choice is made that the reference point for the potential is $\Phi(\infty) = 0$,

> Thus $\Phi(\boldsymbol{x}) = W/Q$ and the scalar potential $\Phi(\boldsymbol{x})$ is the *work per unit charge* required to move a test charge from infinity to the position $\boldsymbol{x}$ in an electric field. It is also the potential energy stored in the fields of the assembled charge configuration that could be released by moving all the assembled charges back to infinity.

Example 2.4 illustrates using Eq. (2.39) to calculate the total work done in assembling a local set of charges.

Example 2.4 Let's use Eq. (2.39) 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. (2.39) summed pairwise over contributions from all charges. Therefore the total work required to assemble the final distribution of n charges is

$$\begin{aligned} W &= \frac{1}{2}\frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{q_i q_j}{r_{ij}} \\ &= \frac{1}{2}\sum_{i=1}^{n} q_i\left(\sum_{j\neq i}^{n}\frac{1}{4\pi\varepsilon_0}\frac{q_j}{r_{ij}}\right) \\ &= \frac{1}{2}\sum_{i=1}^{n} q_i\,\Phi(\boldsymbol{r}_i), \end{aligned} \tag{2.40}$$

where r_{ij} is the distance between charges i and j, the restriction $i \neq j$ omits infinite self-energies of a charge interacting with itself (see the discussion in Section 1.6 of infinite self energies of point particles), 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. (2.30)].

Equation (2.40) represents the work required to assemble the n charges into some local configuration. Therefore, Eq. (2.40) also represents a total (potential) energy stored in the electric fields of the final assembly of charges, which could be released by moving all of the charges back to infinity.

2.7.2 Energy of a Continuous Charge Distribution

Previous sections have shown how to calculate the energy of a discrete set of charges. Let's now address determining the energy of a *continuous distribution of charge*. From Eq. (2.40) we infer that assembling a continuous volume charge requires an amount of work

$$W = \frac{1}{2}\int \rho\,\Phi\, d\tau, \tag{2.41}$$

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. (2.21) to express ρ in terms of $\boldsymbol{E}$,

$$W = \frac{\varepsilon_0}{2}\int (\boldsymbol{\nabla}\cdot\boldsymbol{E})\,\Phi\, d\tau, \tag{2.42}$$

and integrate this by parts (see Appendix A.11) to give

$$W = \frac{\varepsilon_0}{2}\left(-\int_V \boldsymbol{E}\cdot(\boldsymbol{\nabla}\Phi)\, d\tau + \oint_S \Phi\boldsymbol{E}\cdot d\boldsymbol{a}\right), \tag{2.43}$$

where $d\boldsymbol{a} \equiv \boldsymbol{n}\, da$. But from Eq. (2.35), $\boldsymbol{\nabla}\Phi = -\boldsymbol{E}$, so

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

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, giving

$$W = \frac{\varepsilon_0}{2}\int E^2 d\tau, \tag{2.45}$$

where it is understood that the integration is over all space. The use of Eq. (2.45) is illustrated in Example 2.5.

Example 2.5 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. (2.45),

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

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

2.8 Equipotential Surfaces and Field Lines

The equation $\Phi(\boldsymbol{x}) = \Phi_0$ defines a family of 2D surfaces having constant potential Φ_0 in 3D space that are called *equipotential surfaces*. By virtue of Eq. (2.37), it takes no work to move charges on such equipotential surfaces. It is often useful to visualize the potential in an electrostatics problem by displaying equipotential contours graphically. Likewise, it is useful to visualize the electric field associated with the potential. The potential is a scalar, having one value at a given coordinate. The electric field is a vector field; it can be represented by an arrow at a given point, indicating a magnitude by the length (or color in a color plot) and a direction by an orientation angle for the arrow.

2.8.1 Equipotential Contours and Vector Fields

Figure 2.8 illustrates a 2D projection of equipotential contours (the solid lines) superposed on vectors indicating the magnitude and direction of the corresponding electric field for a static dipole charge configuration (see Box 3.3). Note that in Fig. 2.8 the vectors representing the electric field are generally perpendicular to the potential contours where they cross. We can see that this must be so immediately by considering an equipotential contour. From Eq. (2.37), for any path lying entirely on an equipotential surface,

$$\int_{\boldsymbol{x}_1}^{\boldsymbol{x}_2} \boldsymbol{E}\cdot d\boldsymbol{l} = 0, \tag{2.46}$$

which indicates either that $\boldsymbol{E} = 0$ along the entire path, or that the direction of $\boldsymbol{E}$ is orthogonal to the equipotential surface at every point. Thus when any non-vanishing electric field lies on an equipotential surface the field vector must be oriented at right angles to the equipotential surface.

2.8.2 Electric Field Lines

Rather than drawing the full projection in 2D of the vector field, as has been done in Fig. 2.8, it is often useful to construct *electric field lines*, which are continuous curves drawn connecting points corresponding to the same magnitude of the electric field, drawn such that a differential element of arc length on the field line $d\boldsymbol{l}$ points in the direction of $\boldsymbol{E}(\boldsymbol{x})$ at each point; that is

$$d\boldsymbol{l} = \lambda\boldsymbol{E}, \tag{2.47}$$

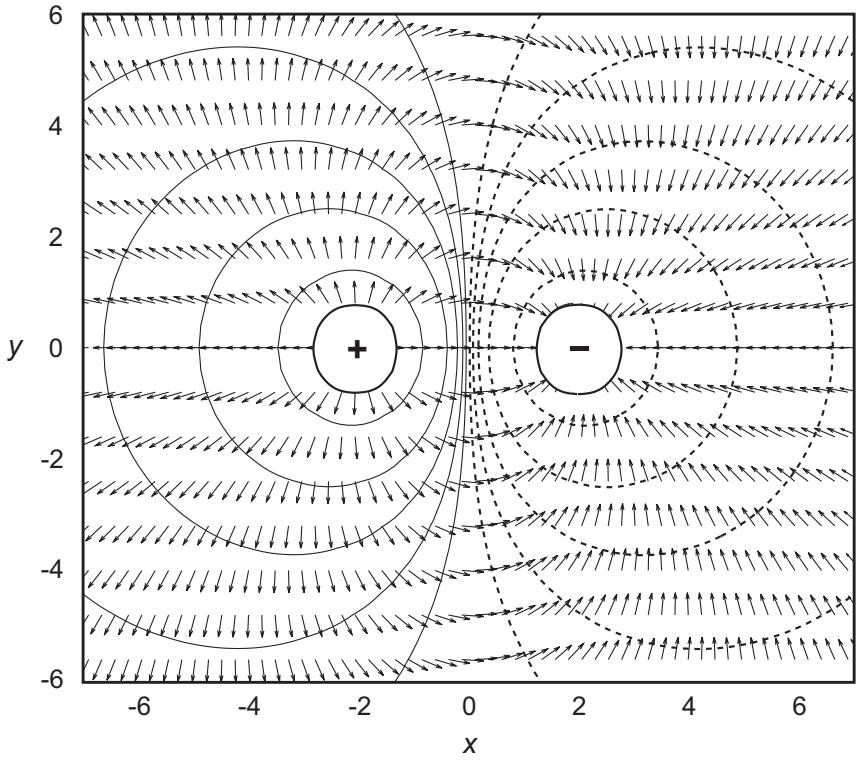


Fig. 2.8 Equipotential contours (solid curves) and the electric vector field (arrows) originating in a dipole charge distribution (see Box 3.3). For clarity, plotting of contours and vectors has been suppressed near the two charges. Notice that the field vectors cross equipotential contours at right angles, and must reverse from outward-going around the positive charge to inward-going around the negative charge. To avoid clutter the equipotentials are not labeled, but they are negative on the right side of the diagram ($x < 0$), as indicated by dashed contours, and positive on the left side $x > 0$, as indicated by solid contours.

where λ is a constant. Thus, from the discussion in Section 2.8.1, every electric field line is locally normal to an equipotential surface,[9] except at points where $\boldsymbol{E} = 0$. The differential equations defining the electric field lines correspond to writing out the components of Eq. (2.47); in cartesian coordinates they take the form

$$\frac{dx}{dE_x} = \frac{dy}{dE_y} = \frac{dz}{dE_z} = \lambda. \tag{2.48}$$

An example of electric field lines for an electric dipole charge distribution is given in Fig. 2.9. Each field line connects points having a constant electric field strength, with the direction of $\boldsymbol{E}$ given by the tangent to the curve at that point. Representative electric field vectors are shown at various points. For example, two vectors are drawn on the uppermost lobe, both of the same length since they are on the same field line, but with different directions corresponding to tangents to the field line where they are drawn. Figure 2.10

[9] Notice that this is consistent with the usual geometrical interpretation of the gradient in $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ being related to a direction of steepest descent in a potential surface.

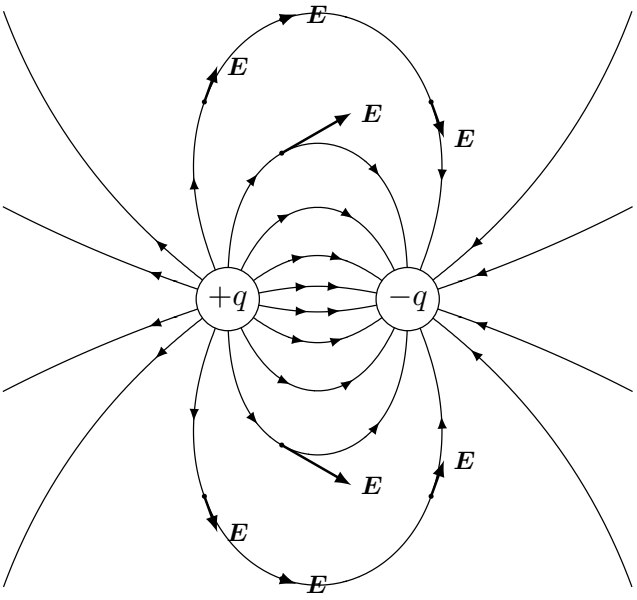


Fig. 2.9 Electric field lines for an electric dipole charge distribution with charges $\pm q$. Notice the convention reflected in the field lines that the electric field points away from positive charges and toward negative charges. Some representative vectors $\boldsymbol{E}$ are shown at several points, with length proportional to the field strength on that curve, and direction tangent to the field line at that point.

illustrates plotting electric field lines and electric potential contours on the same graph for an electric dipole. Figure 2.11 displays electric field lines and electric potential contours for a charge distribution consisting of two nearby charges of the same sign.

Notice in these plots that since $\boldsymbol{E}$ has a unique direction at every spatial point, electric field lines cannot cross (except at a null point where the field is tending to zero). The number density of field lines passing through a differential element of an equipotential surface is assumed to be proportional to the magnitude of $\boldsymbol{E}$ at that point. This implies that the total number of field lines passing through a surface S is proportional to the electric flux $\mathscr{F}_{\mathrm{E}}$,

$$\mathscr{F}_{\mathrm{E}} = \int_S \hat{\boldsymbol{n}} \cdot \boldsymbol{E} dS. \tag{2.49}$$

As you are asked to show in Problem 2.8, if the surface S is *closed*, charge conservation demands that the flux through the surface must be proportional to the charge enclosed by the surface,

$$\mathscr{F}_{\mathrm{E}} = \frac{Q_{\mathrm{V}}}{\varepsilon_0}, \tag{2.50}$$

which means that the net number of electric field lines crossing S is proportional to the net charge enclosed by S. As a corollary, if no charge is enclosed in the volume V, every field line that enters V must also leave V.

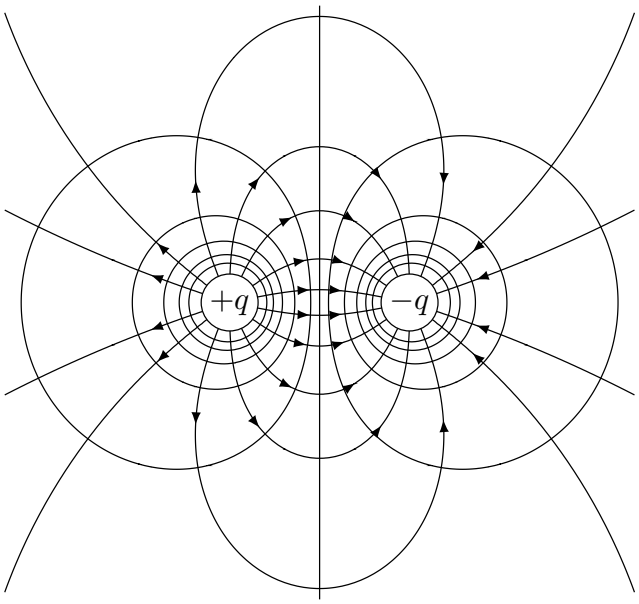


Fig. 2.10 Electric field lines (black with arrows) and equipotential contours (gray, no arrows) for an electric dipole with charges $\pm q$. Notice that the field lines are always locally perpendicular to the potential contours.

2.8.3 2D Projections of 3D Physics

Two-dimensional representations of (usually 3D) electric fields and potentials can be very useful in visualizing an electrostatics problem, but 2D projections can distort the actual physical situation in a 3D problem. This is particularly true in interpreting the density of field lines as indicating the local strength of the field, since projection from 3D to 2D can give misleading information about the actual density of field lines in 3D. On the other hand, more realistic 3D volume rendering is technically more demanding, and may present its own problems of visual interpretation when displayed in static 2D media.

2.9 Superposition of Scalar Potentials

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,

$$\boldsymbol{F} = \boldsymbol{F}_1 + \boldsymbol{F}_2 + \boldsymbol{F}_3 + \ldots \tag{2.51}$$

and since $\boldsymbol{F} = Q\boldsymbol{E}$, dividing the terms in Eq. (2.51) by Q implies linearity for the electric fields $\boldsymbol{E}$ also,

$$\boldsymbol{E} = \boldsymbol{E}_1 + \boldsymbol{E}_2 + \boldsymbol{E}_3 + \ldots \tag{2.52}$$

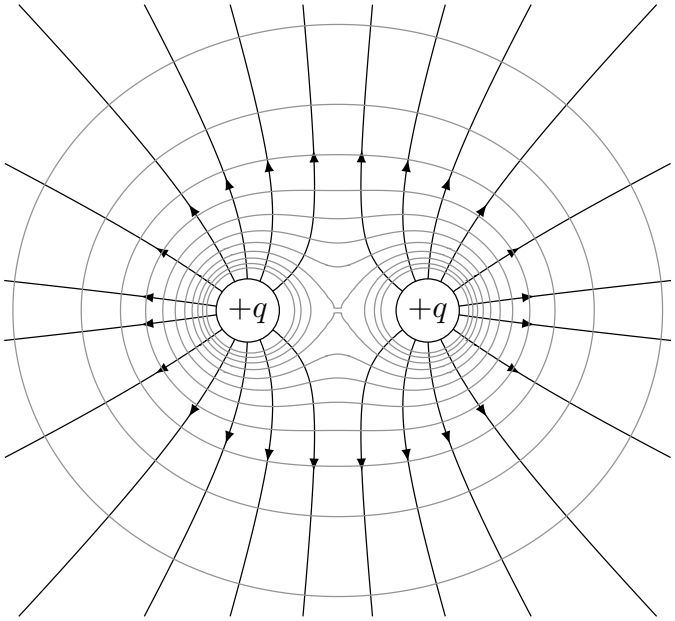


Fig. 2.11 Electric field lines (black with arrows) and equipotential contours (gray, no arrows) for two nearby charges of the same sign $+q$.

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

$$\Phi = \Phi_1 + \Phi_2 + \Phi_3 + \ldots, \tag{2.53}$$

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

Linear superposition of electrostatic forces and electric fields in Eqs. (2.51) and (2.52) corresponds to *vector sums*, but linear superposition of potentials in Eq. (2.53) entails an ordinary *arithmetic sum* over scalar quantities Φ_i, which is typically easier to deal with than a sum over vectors.

Examples 2.6 and 2.7 illustrate calculating the potential Φ of a charge distribution and then taking its gradient to give the electric field.

Example 2.6 Let's calculate the potential and corresponding electric field for the electric dipole charge configuration displayed in Fig. 2.12(a) at the point P, and also also approximate the result for a very large distance $x \gg a$ along the x axis from the charges. At

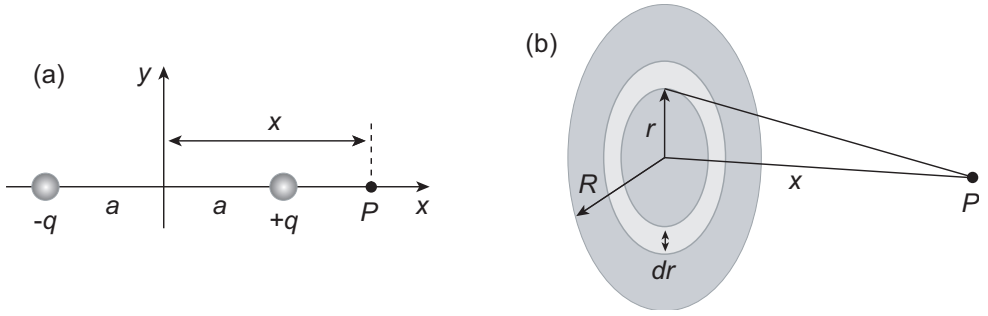


Fig. 2.12 (a) An electric dipole charge configuration used in Example 2.6. (b) Uniformly charged disk of radius R and surface charge density σ used in Example 2.7.

the point P the potential Φ is the sum of contributions from the two charges,

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

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

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

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 2.7 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. 2.12(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 since the potential obeys the superposition principle the total Φ at the point P is obtained by integration over the disk,

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

Then 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 2.6 and 2.7, the potentials can be worked out easily and then the electric field is determined by taking the gradient of the potential. However, more complex situations may require more powerful and systematic ways to determine potentials. In the next section we show that the problem of finding potentials can be cast in the form of solving a second-order partial differential equation. Although it can become mathematically involved, this approach may be preferred over the ones we have examined so far, particularly for problems with complicated boundary conditions, because it provides a systematic procedure that may be applied to a broad range of complex problems.

2.10 The Poisson and Laplace Equations

We have already found in Eq. (2.27) that the curl $\boldsymbol{\nabla}\times\boldsymbol{E}$ of the electric field vanishes. It is of interest then to ask, what is the divergence $\boldsymbol{\nabla}\cdot\boldsymbol{E}$ of the electric field? From Eq. (2.35),

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

and comparing this result with Gauss's law (2.21) gives *Poisson's equation,*

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon_0} \qquad \text{(Poisson's equation)}, \tag{2.54}$$

where the *Laplacian operator* ∇^2 in cartesian coordinates is given by [see Eq. (A.19)][10]

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \tag{2.55}$$

which operates on a scalar to return a scalar. We will need the Laplacian operator at times in other coordinate systems, so let's go ahead and write it in spherical coordinates (r,θ,ϕ),

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}, \tag{2.56}$$

[see Eq. (A.50)], and in cylindrical coordinates (ρ,ϕ,z),

$$\nabla^2 = \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2}{\partial\phi^2} + \frac{\partial^2}{\partial z^2}. \tag{2.57}$$

[10] Some author use the symbol Δ instead of ∇^2 for the Laplacian operator. The Poisson and Laplace equations are partial differential equations (PDEs), which are typically more difficult to deal with than ordinary differential equations (ODEs). We shall address methods of solution for the Poisson and Laplace equations shortly.

[see Eq. (A.57)].

For regions where there is no charge density $\rho = 0$ and Poisson's equation reduces to *Laplace's equation,*

$$\nabla^2\Phi = 0 \qquad \text{(Laplace's equation).} \tag{2.58}$$

By construction, solving Poisson's equation (or Laplace's equation if $\rho = 0$) is equivalent physically to solving for the potential Φ using Eq. (2.28), and once Φ has been determined by either means the electric field can be calculated from Eq. (2.35).

An important property of Laplace's equation (2.58) is that *it is linear:* if each of a set of potentials $\{\Phi_1, \Phi_2, \Phi_3, \ldots, \Phi_n\}$ satisfy (2.58), then a linear combination of them,

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

with arbitrary constants a_i is also a solution.

Example 2.8 Earlier it was asserted that Eq. (2.31a) gives the scalar potential $\Phi(\boldsymbol{x})$ for a 3D continuous charge distribution, which means that $\Phi(\boldsymbol{x})$ should be a solution of the 3D Poisson equation (2.54). Let's check that it is. From Eq. (2.31a),

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

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 as the integrand is singular as $\boldsymbol{x}' \to \boldsymbol{x}$, but this may be handled simply using that from Eq. (A.75) 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 (2.54).

A more involved solution for Example 2.8 that doesn't invoke the Dirac δ-function may be found in Ch. 1 of Jackson [19].

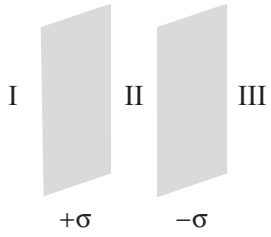
Background and Further Reading

Introductions to the material of this chapter may be found in Corson and Lorrain [7], Griffiths [13], Kamberaj [21], Lorrain and Corson [26], and Purcell and Morin [32]. More advanced treatments may be found in Jackson [19], Garg [11], Chaichian et al [5], Zangwill [42], and Wald [40].

Problems

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

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



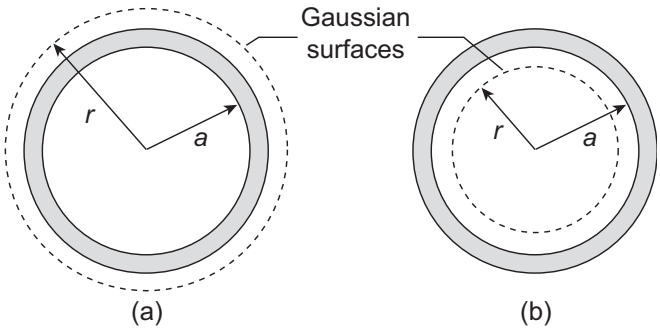
Find the electric fields in regions I, II, and III.

2.3 Prove that an isolated good conductor in electrostatic equilibrium (no net motion of charge carriers) 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).

Hint: Gauss's law will be extremely useful. ***

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.

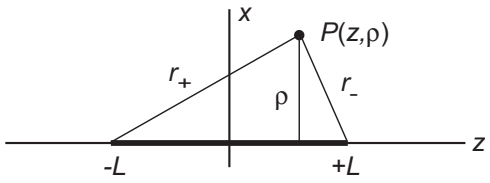


Find the electric field inside and outside the spherical shell using the Gaussian surfaces indicated by dashed circles.

2.5 (a) Use Stokes' theorem to prove that an electric field is conservative by examining the work done on a test charge moved over a closed path. (b) Prove that the components of an electric field $\boldsymbol{E}$ satisfy the constraints imposed by Eq. (2.36).

2.6 Two charge distributions $\rho_A(\boldsymbol{x})$ and $\rho_B(\boldsymbol{x})$ are related by a change of scale: $\rho_B(\alpha\boldsymbol{x}) = \rho_A(\boldsymbol{x})$, where α is a constant. What is the relationship of the scalar potentials and the electric fields for these two charge distributions?

2.7 Consider a 1D line segment of length $2L$, shown in the following figure,

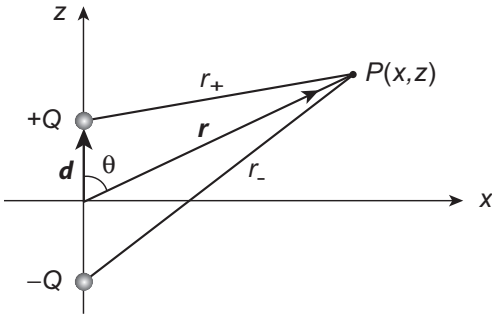


which carries a uniform charge density per unit length of λ. What is the electric potential at an arbitrary point $P(z,\rho)$? ***

2.8 Use Gauss's law to show that if an equipotential contour corresponds to a closed surface S, the electric flux (2.49) through the surface must satisfy $\mathscr{F}_E = Q_V/\varepsilon_0$, where Q_V is the charge contained within the volume V bounded by the surface S.

2.9 A uniform electric field $\boldsymbol{E}$ of strength $4\times10^4\,\mathrm{V\,m^{-1}}$ is directed along the positive x-axis. A proton ($Q = 1.6\times10^{-19}\,\mathrm{C}$) is released from rest at point A and is found to undergo a displacement of $d = 0.5\,\mathrm{m}$ in the direction of $\boldsymbol{E}$ to point B. (a) What is the difference in electrical potential between points A and B? (b) How much does the potential energy of the proton change over this displacement?

2.10 Consider the electric dipole configuration

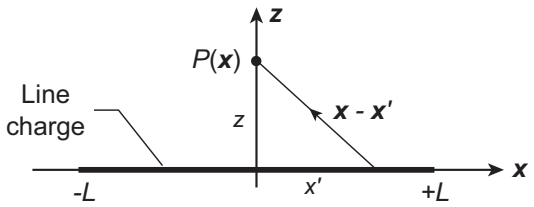


where the distance to the point $P(x,z)$ is much larger than the separation of the two charges. What is the electric potential Φ and what are the components of the electric field E_x and E_z evaluated at $P(x,z)$, expressed as functions of the distance $r = |\boldsymbol{r}|$ and the angle θ?

2.11 (a) Show that the general formula for the electric field produced by an infinite line charge of uniform density λ at a perpendicular distance z from the line charge is

$$E = \frac{\lambda}{2\pi\varepsilon_0 z}\hat{\boldsymbol{z}},$$

by finding the electric field at the point P in the following figure,



produced by a line charge λ of length $2L$, and then taking the limit $L \to \infty$. (b) Show that in the limit $z \gg L$, the electric field looks like that of a point charge of magnitude $2\lambda L$. ***

3 Electrostatic Boundary Value Problems

In Ch. 2 we introduced the Poisson equation (2.54), with solutions corresponding to scalar potentials $\Phi(\boldsymbol{x})$ in the presence of a charge density, and the Laplace equation (2.58), with solutions corresponding to scalar potentials $\Phi(\boldsymbol{x})$ in regions having no charge density, with the electric field then obtained in either case by taking a gradient of the potential. Those solutions result from solving the corresponding partial differential equations subject to boundary conditions, which can be a highly nontrivial undertaking. This chapter addresses the nature of the solutions of the Poisson and Laplace equations, and some means of obtaining those solutions without necessarily solving the differential equations directly. However, many problems are sufficiently complicated that the simplest approach is to grasp the nettle and solve the Poisson and Laplace partial differential equations directly. We shall consider such direct solutions in Ch. 4.

3.1 Electric Fields and Surface Charge Layers

Let us now demonstrate that an electric field necessarily exhibits a discontinuity if a surface charge layer is crossed. Consider Fig. 3.1, 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,

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

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 arising from the surface charge are perpendicular to the local plane and parallel to $\boldsymbol{n}$, so

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

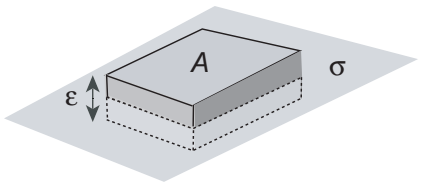


Fig. 3.1 Discontinuous normal of $\boldsymbol{E}$ at surface of a conductor. The surface charge density is σ and the rectangular Gaussian surface has a top area of A and height of ε.

and therefore

$$\boldsymbol{E} = \frac{\sigma}{2\varepsilon_0}\hat{\boldsymbol{n}}, \tag{3.1}$$

where $\hat{\boldsymbol{n}}$ is a unit vector perpendicular to the surface and pointing away from it in Fig. 3.1. Then the difference between electric fields above and below the plane is

$$\boldsymbol{E}_{\text{above}} - \boldsymbol{E}_{\text{below}} = \frac{\sigma}{\varepsilon_0}\hat{\boldsymbol{n}}, \tag{3.2}$$

An electric field is necessarily discontinuous at a boundary charge layer: it points in the same direction on either side of the boundary, but its magnitude changes by σ/ε_0 when the boundary is crossed.

On the other hand, the *scalar potential 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,

$$\Phi_{\text{above}} - \Phi_{\text{below}} = -\int_{\boldsymbol{A}}^{\boldsymbol{B}} \boldsymbol{E}\cdot d\boldsymbol{l}, \tag{3.3}$$

which tends to zero as the distance between $\boldsymbol{A}$ and $\boldsymbol{B}$ is decreased. The *gradient of* Φ and $\boldsymbol{E}$ are related by $\boldsymbol{\nabla}\Phi = -\boldsymbol{E}$, so the scalar potential is continuous but $\boldsymbol{\nabla}\Phi$ inherits the discontinuity in $\boldsymbol{E}$ at the boundary,

$$\boldsymbol{\nabla}\Phi_{\text{above}} - \boldsymbol{\nabla}\Phi_{\text{below}} = -\frac{\sigma}{\varepsilon_0}\hat{\boldsymbol{n}}. \tag{3.4}$$

This can also be rewritten as

$$\frac{\partial\Phi_{\text{above}}}{\partial n} - \frac{\partial\Phi_{\text{below}}}{\partial n} = -\frac{\sigma}{\varepsilon_0} \qquad \frac{\partial\Phi}{\partial n} = \boldsymbol{\nabla}\Phi\cdot\hat{\boldsymbol{n}}, \tag{3.5}$$

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 the surface in Fig. 3.1 is approached very closely from the top or bottom.

3.2 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 of the solutions that are generic. Let us consider one-dimensional equations first before tackling 2D and 3D versions.

3.2.1 The Laplace Equation in 1D

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

$$\nabla^2\Phi = 0 \quad \longrightarrow \quad \frac{\partial^2\Phi}{\partial x^2} = 0 \quad \longrightarrow \quad \frac{d^2\Phi}{dx^2} = 0, \tag{3.6}$$

the last step indicates explicitly that the usual partial differential equation (PDE) reduces to an ordinary differential equation (ODE), if there is only one variable. The ODE (3.6) has a general solution

$$\Phi(x) = ax + b, \tag{3.7}$$

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. 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. (3.6), and can be considerably more involved than this simple example. 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* $\Phi(x+c)$,

$$\Phi(x) = \frac{1}{2}\left[\Phi(x+c) + \Phi(x-c)\right], \tag{3.8}$$

   for arbitrary c.
2. There can be *no local maxima or minima for the solution* as a function of the parameter x, which is a corollary of the first point.

This latter point is sometimes called *Earnshaw's theorem,* which may be viewed as the mean value theorem applied to electrostatics.[1] An alternative statement of Point 2 is that the average of the scalar potential Φ over any sphere that lies entirely in a charge-free region is equal to the value of the potential at the center of the sphere (see Problem 4.2).

Point 2 implies that maxima or minima of a 1D Laplace solution can occur only at the endpoints of the plot, since if there were a local maximum or minimum of Φ at some value other of x not an endpoint, it could not be the average of points on either side of it, contradicting the requirement of Eq. (3.8). Illustrations of this absence of local maxima or minima will be given later in Figs. 4.2 and 4.4 for 2D Laplace solutions. One implication of

[1] *Mean Value Theorem (MVT):* For a continuous and differentiable function, the average rate of change over a closed interval is equal to the instantaneous rate of change at some point in the interval. An illustrative example of applying the MVT concerns detection of automobile speeds with police radar. Suppose the speed limit on a long stretch of road to be 60 miles per hour (MPH). A car is measured by radar at Point A to have a speed of 55 MPH, and a half hour later and 40 miles down the road at Point B the same car is measured to have a speed of 58 MPH. For both local measurements at A and B the car was traveling below the speed limit of 60 MPH, but globally the average speed between A and B was 40 miles/0.5 hours = 80 MPH. By the MVT, it is necessarily true that at some point between A and B the car had a speed of exactly 80 MPH, which exceeds the speed limit by 20 MPH.

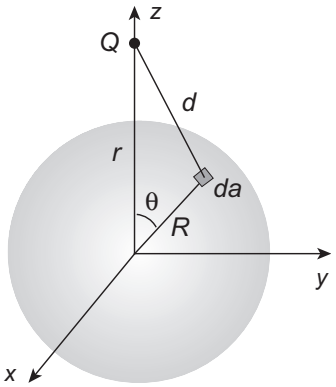


Fig. 3.2 Averaging Φ over a sphere of radius R with a single point charge q lying on the z-axis, external to the sphere and a distance r from its center (used in Example 3.1).

Earnshaw's theorem is that no configuration of electrical charges can be held in equilibrium by purely electrostatic forces, because there can be no local minima in energy (which varies as the square of the electric field).

3.2.2 2D and 3D Laplace Equations

Let us now move to 2D and 3D versions of Laplace's equation, where it is necessary to solve partial differential equations that take the form

$$\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}=0 \qquad \frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}+\frac{\partial^2\Phi}{\partial z^2}=0, \tag{3.9}$$

in cartesian coordinates. 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 with boundary conditions. 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 solution $\Phi(\boldsymbol{x})$ at a point $\boldsymbol{x}$ is an average of values over a sphere centered at $\boldsymbol{x}$.

$$\Phi(\boldsymbol{x})=\frac{1}{4\pi R^2}\oint \Phi\, da, \tag{3.10}$$

   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, so extrema can occur only on the boundaries.

(See illustrations of Point 2 in Figs. 4.2 and 4.4.)

Example 3.1 Let's check Point 1 for a single point charge Q outside a sphere that is a distance r from the center of the sphere, as illustrated in Fig. 3.2. From this figure 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. 3.2),

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

and from Eq. (3.10) the average over the sphere is

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

which is the potential that would be produced by placing Q at the center of the sphere (see Eq. (2.29)).

3.3 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. 3.3, where the task is to solve for the scalar potential in some volume V bounded by a surface S with parts held at constant potentials. 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 most common in electrostatics problems.[2]

1. *Dirichlet boundary conditions* correspond to specification 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.

[2] From a laboratory perspective, the Dirichlet criteria are typically the simplest and most natural way to set boundary conditions. For example, in the lab one often has conductors connected to batteries, which maintain a fixed finite potential on the conductor, or connected to ground, which corresponds effectively to maintaining $\Phi = 0$. But there may be other situations in which we do not know the potential on the boundaries, but we do know the total charge on each bounding surface. Then Neumann boundary conditions become more natural. Practically, Neumann boundary conditions are seldom natural when only perfect conductors and dielectric matter are present; they can arise in applications such as the theory of waveguides or when steady currents flow through an ohmic medium [42]. As will be mentioned later, in principle mixed conditions can result if Dirichlet constraints are applied to parts of a boundary and Neumann constraints to other parts, but that is typically mathematically complex. Hence we shall emphasize Dirichlet boundary conditions when examples are needed in our discussion.

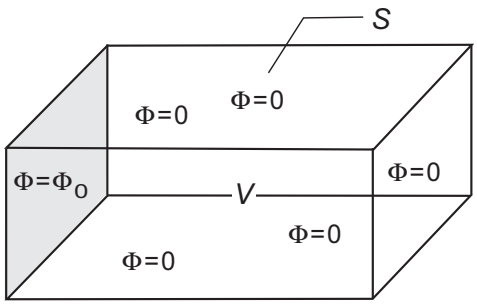


Fig. 3.3 Generic example of an electrostatics problem formulated in a 3D volume V that is bounded by a 2D surface S. In this case the (finite) volume V is surrounded by a rectangular conducting box S with boundary conditions corresponding to five of the six conducting box faces held at zero potential, $\Phi = 0$, 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 (which corresponds to 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 permits formulation of a second uniqueness theorem.

Uniqueness Theorem II: If a volume V is surrounded by conductors of having charge densities 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.

3.3.1 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. 3.3) can be obtained using *Green function methods* [19]. These may be

derived beginning with the divergence theorem,

$$\oint_S \boldsymbol{A}\cdot\boldsymbol{n}\,da = \int_V \boldsymbol{\nabla}\cdot\boldsymbol{A}\,d^3x \qquad \text{(Divergence theorem)}, \tag{3.11}$$

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

$$\boldsymbol{\nabla}\cdot(\phi\boldsymbol{\nabla}\psi) = \phi\nabla^2\psi + \boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\psi \tag{3.12}$$

$$\phi\boldsymbol{\nabla}\psi\cdot\boldsymbol{n} = \phi\frac{\partial\psi}{\partial n}, \tag{3.13}$$

where $\partial/\partial n$ is the normal derivative at the surface, directed from inside to outside the volume V. Substitution of Eqs. (3.12) and (3.13) into the divergence theorem (3.11) then gives *Green's first identity*

$$\int_V (\phi\nabla^2\psi + \boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\psi)\,d^3x = \oint_S \frac{\partial\psi}{\partial n}\,da \qquad \text{(Green's first identity)}. \tag{3.14}$$

If we then subtract from Eq. (3.14) the same expression but with ϕ and ψ interchanged, the $\boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\psi$ terms cancel, giving *Green's theorem* (also termed *Green's second identity*)

$$\int_V (\phi\nabla^2\psi - \psi\nabla^2\phi)\,d^3x = \oint_S \left[\phi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\phi}{\partial n}\right] da \qquad \text{(Green's theorem)}. \tag{3.15}$$

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.75)

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

so that Eq. (3.15) becomes

$$\int_V \left[-4\pi\Phi(\boldsymbol{x}')\delta(\boldsymbol{x}-\boldsymbol{x}') + \frac{1}{\varepsilon_0 R}\rho(\boldsymbol{x}')\right] d^3x' = \oint_S \left[\Phi\frac{\partial}{\partial n'}\left(\frac{1}{R}\right) - \frac{1}{R}\frac{\partial\Phi}{\partial n'}\,da'\right]. \tag{3.16}$$

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

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \frac{\rho(\boldsymbol{x}')}{R}\,d^3x' + \frac{1}{4\pi}\oint_S \left[\frac{1}{R}\frac{\partial\Phi}{\partial n'} - \Phi\frac{\partial}{\partial n'}\left(\frac{1}{R}\right)\right] da'. \tag{3.17}$$

Notice two important implications of this result.

1. If the surface S is at infinity and the electric field on S falls off faster than R^{-1}, the surface term in Eq. (3.17) vanishes and we recover
$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',$$
which is the usual result of Eq. (2.31a),
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. (3.17) is *overdetermined* because it isn't permitted to impose both types of boundary conditions on the same closed surface.[3] Since it is overdetermined, Eq. (3.17) is *not a valid solution*. In Section 3.4 we shall address how to correct this deficiency.

3.3.2 Proof of Uniqueness

Let us now demonstrate explicitly that the use of either (but not both) Dirichlet or Neumann boundary conditions defines a unique potential problem, following the presentation in Jackson [19]. Assume the contrary proposition that the solution is *not unique* and that there are two potentials, Φ_1 and Φ_2 that satisfy the Laplace equation with the same boundary conditions. Let the difference between the two hypothesized solutions be defined by $U = \Phi_2 - \Phi_1$. Then by Laplace
$$\nabla^2 U = \nabla^2\Phi_2 - \nabla^2\Phi_1 = 0$$
inside the volume V, while on the boundary S either

1. $U = 0$ (Dirichlet boundary conditions) or
2. $\partial U/\partial n = 0$ (Neumann boundary conditions).

Setting $\phi = \psi = U$ and applying Green's first identity (3.14) gives
$$\int_V (U\nabla^2 U + \boldsymbol{\nabla}U\cdot\boldsymbol{\nabla}U)\,d^3x = \oint_S U\frac{\partial U}{\partial n}\,da. \tag{3.18}$$
For either Dirichlet or Neumann of boundary conditions the surface integral on the right vanishes, as does the first term in the volume integral on the left, and this reduces to
$$\int_V |\boldsymbol{\nabla}U|^2\,d^3x = 0, \tag{3.19}$$
which implies that U is constant inside V, since $\boldsymbol{\nabla}U = 0$. For Dirichlet boundary conditions $U = 0$ on S (the two purported solutions must be equal on the boundary since they are assumed to have the same boundary conditions). Furthermore, solutions of the Laplace or Poisson equation must have any extrema on the boundaries (see Sections 3.2.1 and 3.2.2),

[3] 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 *contradictory* (its equations can be manipulated to obtain different incompatible results).

so since $U = 0$ on the boundary, U is also zero inside V, implying that $\Phi_1 = \Phi_2$ and the solution is unique. Likewise, for Neumann boundary conditions the solution is unique, up to an arbitrary additive constant. For simplicity this uniqueness proof was for the Laplace equation, but a similar proof of uniqueness follows for the Poisson equation (see Problem 3.9).

Thus we have shown that for either Dirichlet or Neumann boundary conditions the *solution of the Laplace 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).[4] However, since Dirichlet and Neumann boundary conditions each define a unique solution, imposing both Dirichlet and Neumann conditions on the same boundary surface will overdetermine the system and no reliable solution will exist. Let us conclude this section by summarizing some general statements about Dirichlet and Neumann boundary conditions in electrostatics problems gleaned from the preceding discussion.

1. One can specify either Dirichlet or Neumann constraints at each point on a boundary, *but not both*. If both are specified the system is *overdetermined*.
2. It is legitimate to specify parts of a boundary using Dirichlet conditions and other parts using Neumann conditions (as long as Dirichlet and Neumann conditions don't overlap on any part of the boundary).
3. The uniqueness property means that a solution obtained by any method 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 the *Helmholtz Theorem,* which is described in Box 3.1.

3.4 Boundary-Value Problems by Green Functions

In obtaining the result of Eq. (3.17) (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

$$\boldsymbol{\nabla}'^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -4\pi\delta^3(\boldsymbol{x}-\boldsymbol{x}'), \tag{3.20}$$

where $\boldsymbol{\nabla}'^2$ is the Laplace operator ∇^2 acting on $\boldsymbol{x}'$ rather than $\boldsymbol{x}$. The function $1/|\boldsymbol{x}-\boldsymbol{x}'|$ is a specific example of a class of functions called *Green functions* that satisfy Eq. (3.20). A Green function is a solution to an inhomogeneous differential equation with a δ-function

[4] Although non-overlapping mixed boundary conditions are legitimate, they can be difficult to deal with relative to problems with a single type of boundary condition. An example of solving a problem using mixed boundary conditions may be found in Section 3.13 of Ref. [19].

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:

> A vector field with its curl and divergence specified 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}.
>
> *EXAMPLE:* It isn't a coincidence that the Maxwell equations (1.1) in the absence of a magnetic field $\boldsymbol{B}$ satisfy the electrostatic equations
>
> $$\boldsymbol{\nabla}\times\boldsymbol{E}(\boldsymbol{x}) = 0 \qquad \boldsymbol{\nabla}\cdot\boldsymbol{E}(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x})}{\varepsilon_0},$$
>
> insuring by the Helmholtz theorem that the electric field $\boldsymbol{E}(\boldsymbol{x})$ is uniquely defined at all points because its curl and divergence are defined at all points.

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

source term.[5] It provides a convenient method for solving more complicated inhomogenous differential equations through superposition, because general solutions for inhomogeneous differential equations can be approximated by a superposition of δ-functions evaluated at different spacetime points. See the example of computing a Green function for a driven harmonic oscillator in Box 3.2, and the evaluation of the retarded Green function for electromagnetic wave propagation in Section 11.4.

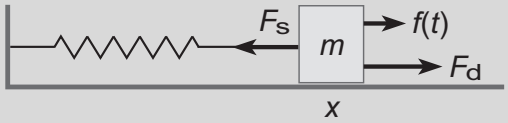
Generally a Green function $G(\boldsymbol{x},\boldsymbol{x}')$ satisfies

$$\boldsymbol{\nabla}'^2 G(\boldsymbol{x},\boldsymbol{x}') = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}') \tag{3.21}$$

[5] An *inhomogeneous differential equation* is one with a function on its right side, as opposed to a *homogeneous differential equation,* which has a zero on its right side. Green functions are applied to classical electromagnetism in the present discussion; they also find extensive use in other areas of mathematical physics such as quantum field theory, but that is a topic for another day.

Box 3.2 **Green Function for a Driven Harmonic Oscillator**

Before diving into the complexities of Green functions applied to electromagnetism, it is useful to see the method in action for a more transparent problem. We may illustrate by finding a solution for the mechanical problem of a 1D driven harmonic oscillator illustrated in the following figure.



For driving force $f(t)$, spring force $F_s = -m\omega_0^2 x$, and drag force $F_d = -2m\gamma\dot{x}$,

$$\left(\frac{d^2}{dt^2} + 2\gamma\frac{d}{dt} + \omega_0^2\right) x(t) = \frac{f(t)}{m} \qquad \text{(Equation of motion).}$$

First consider

$$\left(\frac{\partial^2}{\partial t^2} + 2\gamma\frac{\partial}{\partial t} + \omega_0^2\right) G(t,t') = \delta(t-t'),$$

where $G(t,t')$ is a Green function that describes an oscillator subject to a δ-function driving force. The solution of this equation is (see Problem 3.8),

$$x(t) = \int_{-\infty}^{\infty} G(t,t')\frac{f(t')}{m}\,dt',$$

The Green function method assumes that $f(t)$ can be decomposed into a superposition of δ-function pulses centered at different times. One way to find the Green function is by Fourier transforms (Section 11.4). The Fourier transform of $G(t,t')$ is

$$G(\omega,t') = \int_{-\infty}^{\infty} e^{i\omega t} G(t,t')\,dt.$$

Applying the Fourier transform to both sides of the Green function equation gives

$$\left(-\omega^2 - 2i\gamma\omega + \omega_0^2\right) G(\omega,t') = \int_{-\infty}^{\infty} e^{i\omega t}\delta(t-t')\,dt = e^{i\omega t'}$$

and thus,

$$G(\omega,t') = -\frac{e^{i\omega t'}}{\omega^2 + 2i\gamma\omega - \omega_0^2}.$$

The inverse transform is

$$G(t,t') = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}e^{-i\omega t}G(\omega,t') = -\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{e^{-i\omega(t-t')}}{(\omega-\omega_+)(\omega-\omega_-)},$$

where $\omega_\pm = -i\gamma \pm (\omega_0^2 - \gamma^2)^{1/2}$. There are two poles in the negative complex plane. An integration contour may be closed in the upper half-plane for $t < t'$ (enclosing no poles) and the lower half-plane for $t > t$ (enclosing both poles), giving solutions

$$G(t,t') = \Theta(t-t')e^{-\gamma(t-t')} \times \begin{cases} \left(\omega_0^2 - \gamma^2\right)^{-1/2} \sin\left[\left(\omega_0^2 - \gamma^2\right)^{1/2}(t-t')\right] & (\gamma < \omega_0), \\ \left(\gamma^2 - \omega_0^2\right)^{-1/2} \sinh\left[\left(\gamma^2 - \omega_0^2\right)^{1/2}(t-t')\right] & (\gamma > \omega_0), \end{cases}$$

where Θ is the unit step function.

where we define

$$G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} + F(\boldsymbol{x},\boldsymbol{x}'), \tag{3.22}$$

where $F(\boldsymbol{x},\boldsymbol{x}')$ satisfies the Laplace equation inside the volume V,

$$\boldsymbol{\nabla}'^2 F(\boldsymbol{x},\boldsymbol{x}') = 0. \tag{3.23}$$

The first term on the right side of Eq. (3.22) 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 $\boldsymbol{x}'$),

$$G_0(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}. \tag{3.24}$$

Recall that Eq. (3.17) was derived by substituting $G_0(\boldsymbol{x},\boldsymbol{x}')$ into Green's theorem (3.15), but we concluded that Eq. (3.17) 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 (3.22) raises the possibility that if we substitute Eq. (3.22) 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, substitution of $\phi = \Phi$ and $\psi = G(\boldsymbol{x},\boldsymbol{x}')$ into Green's theorem (3.15) and use of Eq. (3.21) generalizes Eq. (3.17) to

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

Now the freedom to choose $F(\boldsymbol{x},\boldsymbol{x}')$ in the definition (3.22) of the Green function means that we can make the surface integral in Eq. (3.25) depend on a specific type of boundary condition.

1. If *Dirichlet boundary conditions* are desired, we may require that

$$G(\boldsymbol{x},\boldsymbol{x}') = 0 \qquad \text{(Dirichlet)} \tag{3.26}$$

for $\boldsymbol{x}'$ on S. Then the first term in the surface integral of Eq. (3.25) vanishes and the *solution with Dirichlet boundary conditions* is

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' - \frac{1}{4\pi}\oint_S \left[\Phi(\boldsymbol{x}')\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'}\right] da'. \tag{3.27}$$

2. If instead *Neumann boundary conditions* are desired, the obvious choice $\partial G(\boldsymbol{x},\boldsymbol{x}')/\partial n' = 0$ for $\boldsymbol{x}'$ on S makes the second term in the surface integral of Eq. (3.25) vanish. But application of Gauss's theorem to Eq. (3.21) indicates that

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

implying that the simplest allowable boundary condition is

$$\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'} = -\frac{4\pi}{\Sigma} \qquad \text{(Neumann)}, \tag{3.28}$$

for $\boldsymbol{x}'$ on S, where Σ is the total area of the bounding surface S.[6] Then the *solution with Neumann boundary conditions* is

$$\Phi(\boldsymbol{x}) = \langle\Phi\rangle_S + \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' + \frac{1}{4\pi}\oint_S \frac{\partial\Phi}{\partial n'}G\,da', \tag{3.29}$$

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

Thus, we have shown formally in Eqs. (3.26) and (3.28) how to determine the scalar potential by imposing consistent boundary conditions in an electrostatic boundary-value problem.

The Green functions $G(\boldsymbol{x},\boldsymbol{x}')$ satisfy the rather simple boundary conditions (3.26) or (3.28). Nevertheless, they may be difficult to determine in many cases because of the dependence on the shape of the surface S.

3.5 The Method of Images

Various tricks and clever algorithms have been devised to find solutions for the Poisson or Laplace equations without actually solving the differential equations themselves. These methods are generally applicable when certain special features are present in a problem, and can often lead to much easier solutions than solving the relevant differential equation with boundary conditions. One such approach is the *method of images*, where the actual Poisson or Laplace problem is replaced with a different one (the *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 or II* given in Section 3.3 guarantees that the solution of the analog problem is also the solution of the original problem.

Let us illustrate this approach by solving a simple electrostatics problem using the image method. In Fig. 3.4(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 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 the field in the region above the conducting plane be

[6] A typical Neumann problem has a volume V bounded by two surfaces, one closed and finite, and the other at infinity [19]. Thus the surface area Σ is infinite, the term $\langle\Phi\rangle_S$ in Eq. (3.29) vanishes, and the right side of the boundary condition (3.28) is zero. Then both the Dirichlet boundary condition Eq. (3.26) and and the Neumann boundary condition Eq. (3.28) become *homogeneous* (zeros on the right side of the equations).

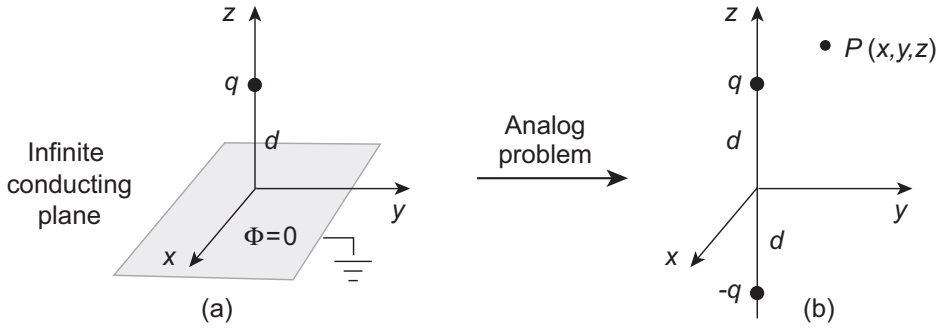


Fig. 3.4 (a) Charge q a distance d above an infinite, grounded, conducting plane at zero potential. (b) An analogous image-charge configuration with the charge $-q$ on the negative z axis, and no infinite conducting plane.

determined when the distribution of the polarized charge isn't known beforehand? 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, and
2. $\Phi \to 0$ at very large distances from the charge.

But before rushing off to try to solve the Poisson equation with these boundary conditions, recall Uniqueness Theorem I from Section 3.3, which indicates that there can be *only one independent solution of the Poisson equation* with precisely these boundary conditions. Thus, if we can find such a solution by any means (including guessing!), it must be correct.

This motivates us to consider the completely different problem illustrated in Fig. 3.4(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 (this is called an *image charge*, and the conducting plane of the original problem has been removed. The problem in Fig. 3.4(b) is simple compared with that of Fig. 3.4(a) and can be solved easily using Coulomb's law. For an arbitrary point $P(x,y,z)$ at $z > 0$, the distances from P to the two charges are, respectively,

$$r_+ = \sqrt{x^2+y^2+(z-d)^2} \qquad r_- = \sqrt{x^2+y^2+(z+d)^2},$$

and the potential at the point $P(x,y,z)$ generated by the two charges is

$$\begin{aligned}\Phi(x,y,z) &= \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{r_+} - \frac{q}{r_-}\right) \\ &= \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{\sqrt{x^2+y^2+(z-d)^2}} - \frac{q}{\sqrt{x^2+y^2+(z+d)^2}}\right).\end{aligned} \tag{3.30}$$

Now we notice that the *boundary conditions are the same for the two problems:* in the analog problem there is a single charge q in the region $z > 0$ (which is where we seek a solution), and from Eq. (3.30),

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

which are precisely the boundary conditions for the original problem in Fig. 3.4(a). To be sure there is an extra charge $-q$ in the $z<0$ region for the analog problem, but that is irrelevant because we *restricted the solution to the region* $z>0$ in the original problem. Thus, we conclude from Uniqueness Theorem I that the solution of our original problem in Fig. 3.4(a) is given by Eq. (3.30), which is the solution of the much simpler analog problem shown in Fig. 3.4(b). This approach can be a powerful one, but it is useful only if we can conjure a solvable analog problem that has the same boundary conditions as the original problem.

> In using the method of images the *image charges must be placed in a region that is not in the domain of the original problem*. In this example we accomplished that by putting the image charge in the region $z<0$, but required a solution only for $z>0$.

Now that the potential has been found in Eq. (3.30) for the upper half plane we can go further and determine the induced charge in the conducting plane caused by the positive charge at $z=d$. From Eq. (3.5) for the effect of charge layers, keeping only the contribution from above the layer,

$$\sigma = -\varepsilon_0 \frac{\partial \Phi}{\partial n}. \tag{3.31}$$

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

$$\sigma = -\varepsilon_0 \left.\frac{\partial \Phi}{\partial z}\right|_{z=0}. \tag{3.32}$$

The derivative may be evaluated from Eq. (3.30) as,

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

and from Eq. (3.32)

$$\sigma = -\varepsilon_0 \left.\frac{\partial \Phi}{\partial n}\right|_{z=0} = -\frac{qd}{2\pi[x^2+y^2+d^2]^{3/2}}. \tag{3.33}$$

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

$$Q = \int_0^{2\pi}\int_0^{\infty} \frac{-qd}{2\pi[x^2+y^2+d^2]^{3/2}}\, r\,dr\,d\phi = \left.\frac{qd}{\sqrt{r^2+d^2}}\right|_0^{\infty} = -q. \tag{3.34}$$

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 on the grounded

conducting plate. This force must be the same as the force in the analog problem between the charges q and $-q$ separated by a distance $2d$, because the charge q cannot tell whether it is seeing a point charge at a distance $2d$ or a grounded conducting plane at a distance d. Thus from the analog problem Coulomb's law gives for the force between the charge q and the conducting plane,

$$\boldsymbol{F} = -\frac{1}{4\pi\varepsilon_0}\frac{q^2}{(2d)^2}\,\hat{\boldsymbol{z}}, \tag{3.35}$$

where $\hat{\boldsymbol{z}}$ is a unit vector in the z direction. It will always be the case that in an image charge problem such as that in Fig. 3.4, the force between a point charge and the conductor is given by the force between the charge and image charge.

3.6 Green Function for the Conducting Sphere

As discussed in Section 3.4, 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. (3.27) solves for the potential Φ in terms of a Green function with Dirichlet boundary conditions and Eq. (3.29) 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.

The physical meaning of $F(\boldsymbol{x},\boldsymbol{x}')$ in Eq. (3.23) is that it is a *solution of the Laplace equation inside the volume V*, so it represents the *potential due to charges external to V*. It may be thought of as resulting from an external distribution of charges chosen to satisfy the Dirichlet boundary conditions (3.26) or the Neumann boundary conditions (3.28). But this arrangement of charges to satisfy boundary conditions suggests a connection to the method of images described in Section 3.5. As discussed by Jackson (see Section 1.10 of Ref. [19]):

> Determining the proper $F(\boldsymbol{x},\boldsymbol{x}')$ in Eq. (3.22) to satisfy the boundary conditions (3.26) or (3.28) is physically equivalent to using the method of images. Thus for image problems the potential due to a unit source and its image charge that satisfy homogeneous boundary conditions is just the Green function of Eq. (3.27) for Dirichlet boundary conditions, or Eq. (3.29) for Neumann boundary conditions.

As an example, consider the problem of a charge outside a conducting sphere that is solved by the image method in Problem 3.1 and is illustrated in Fig. 3.5(a). In the Green function $G(\boldsymbol{x},\boldsymbol{x}')$, our usual convention is that 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. 3.5(b). For Dirichlet boundary conditions on the sphere of radius a, the Green function defined by Eq. (3.21),

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

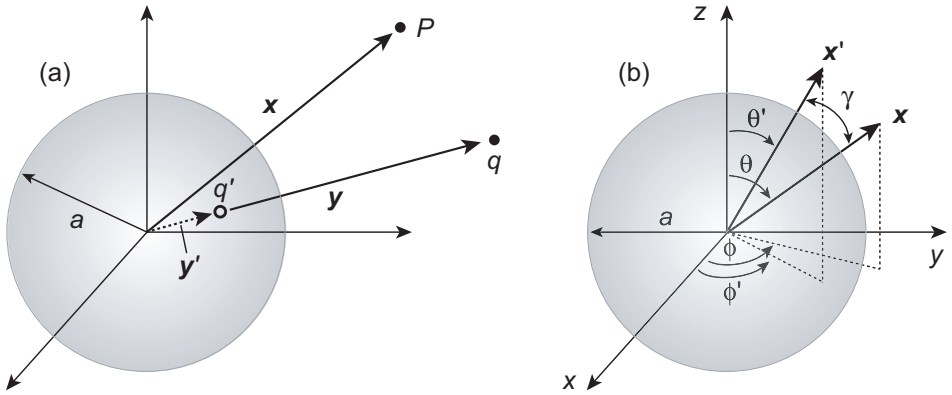


Fig. 3.5 (a) Geometry of image charge method for a charge q indicated by a solid dot outside a conducting sphere, with an image charge q' indicated by an open dot inside the sphere. (b) Coordinates associated with the Green function $G(\boldsymbol{x},\boldsymbol{x}')$ for a conducting sphere.

for a unit source and its image, the scalar potential is given by [see Fig. 3.5(a)]

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{|\boldsymbol{x}-\boldsymbol{y}|}+\frac{q'}{|\boldsymbol{x}-\boldsymbol{y}'|}\right), \tag{3.36}$$

with q' and $\boldsymbol{y}'$ chosen such that the potential vanishes on the surface of the sphere in Fig. 3.5(a) ($|\boldsymbol{x}| = a$), which requires that

$$q' = -\frac{a}{y}q \qquad y' = \frac{a^2}{y}. \tag{3.37}$$

Using (3.37), replacing q by $4\pi\varepsilon_0$ (to give unit source charge), and replacing y with x' in Eq. (3.36), the required Green function is then

$$G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} - \frac{a}{x'\,|\boldsymbol{x}-(a^2/x'^2)\boldsymbol{x}'|}, \tag{3.38}$$

which can be expressed in spherical coordinates as [19]

$$G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{(x^2+x'^2-2xx'\cos\gamma)^{1/2}} - \frac{1}{(x^2x'^2/a^2+a^2-2xx'\cos\gamma)^{1/2}}, \tag{3.39}$$

where γ is the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$ that is illustrated in Fig. 3.5(b). From Eq. (3.39) it is clear that

1. $G(\boldsymbol{x},\boldsymbol{x}')$ is symmetric in $\boldsymbol{x}$ and $\boldsymbol{x}$, as it should be,[7] and
2. $G(\boldsymbol{x},\boldsymbol{x}') = 0$ if either $\boldsymbol{x}$ or $\boldsymbol{x}'$ approaches the spherical radius a, as required by the Dirichlet boundary condition (3.26).

[7] For Dirichlet boundary conditions, the Green function $G(\boldsymbol{x},\boldsymbol{x}')$ viewed as a function of one of its variables represents a potential produced by a unit point source. Then symmetry in $\boldsymbol{x}$ and $\boldsymbol{x}'$ implies physical interchangeability of source point and observation point.

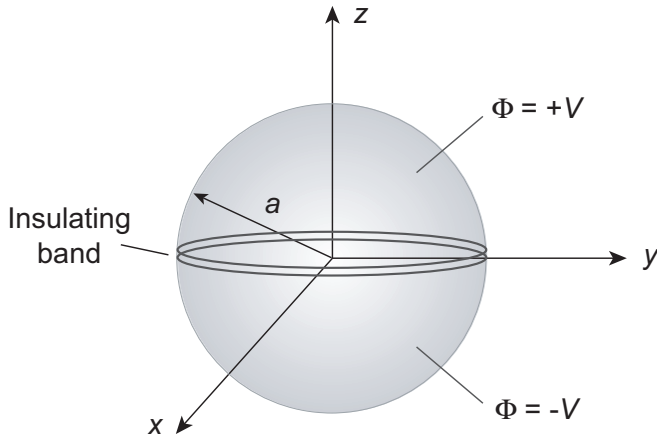


Fig. 3.6 Two hemispheres of radius a separated by an insulating band lying in the $z = 0$ plane, with the upper hemisphere held at potential $+V$ and the lower hemisphere held at potential $-V$. Used in Example 3.2.

For the solution (3.27), of the Poisson equation with Dirichlet boundary conditions,

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

the second term requires the normal derivative $\partial G/\partial n'$. Recall our standard convention that $\boldsymbol{n}'$ is the unit normal vector *outward from the volume of interest*. Thus, for the solution *outside the sphere* in Fig. 3.5(a), it is inward along $\boldsymbol{x}'$, toward the origin and

$$\left.\frac{\partial G}{\partial n'}\right|_{x'=a} = -\frac{x^2-a^2}{a(x^2+a^2-2ax\cos\gamma)^{3/2}}. \tag{3.40}$$

Therefore, from Eq. (3.27) 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

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi}\int \Phi(a,\theta',\phi')\frac{a(x^2-a^2)}{(x^2+a^2-2ax\cos\gamma)^{3/2}}\,d\Omega' \qquad \text{(Exterior)}, \tag{3.41}$$

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

$$\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi-\phi'). \tag{3.42}$$

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. (3.40) changes. Thus the interior solution is

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi}\int \Phi(a,\theta',\phi')\frac{a(a^2-x^2)}{(x^2+a^2-2ax\cos\gamma)^{3/2}}\,d\Omega' \qquad \text{(Interior)}. \tag{3.43}$$

If there is a charge distribution $\rho(\boldsymbol{x}')$, then one must add to Eqs. (3.41) and (3.43) the first term of Eq. (3.27) with the associated Green function (3.39).

Example 3.2 Let's illustrate use of the exterior solution (3.41) by considering a conducting sphere of radius a, divided into two hemispherical shells separated by an insulating ring, as illustrated in Fig. 3.6. From Eq. (3.41) 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

$$\Phi(x,\theta,\phi) = \frac{Va(x^2-a^2)}{4\pi}\int_0^{2\pi} d\phi' \times \int_0^1 d(\cos\theta') \left[(a^2+x^2-2ax\cos\gamma)^{-3/2} - (a^2+x^2+2ax\cos\gamma)^{-3/2}\right]. \quad (3.44)$$

Because of the complicated dependence among the angles in Eq. (3.42) this cannot be integrated easily in closed form. If one restricts to the positive z axis the integrals can be done, with the result [19]

$$\Phi(z) = V\left[1 - \frac{z^2-a^2}{z\sqrt{z^2+a^2}}\right] \qquad \text{(Valid on positive } z\text{-axis)}, \quad (3.45)$$

which correctly reduces to $\Phi = V$ at $z = a$. Of potentially more use is to expand the denominator of Eq. (3.44) in a power series and integrate term by term, which yields

$$\Phi(x,\theta,\phi) = \frac{3Va^2}{2x^2}\left[\cos\theta - \frac{7a^2}{12x^2}\left(\frac{5}{2}\cos^2\theta - \frac{3}{2}\cos\theta\right) + \cdots\right]. \quad (3.46)$$

This expansion has been shown to converge rapidly for large x/a, and agrees with the special solution (3.45) for $\cos\theta = 1$ [19].

3.7 Approximate Solutions

In a complicated electrostatics problem it may be sufficient to approximate the solution, as we have done in Eq. (3.46). Often this approximation takes the form of an expansion in some naturally small parameter, with the first few terms of the expansion giving sufficient accuracy for the problem at hand.

3.7.1 Spherical Harmonic Expansions of the Potential

The 3D volume charge density $\rho(\boldsymbol{x})$ appearing in the expression Eq. (2.31a) for the scalar potential,

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x', \quad (3.47)$$

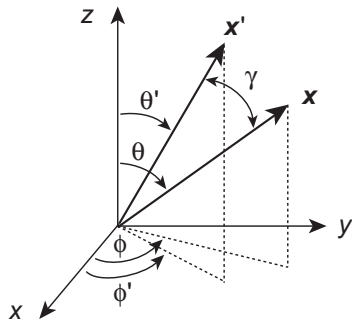


Fig. 3.7 Geometry for the Legendre polynomials and the spherical harmonics in Eqs. (3.48)-(3.51).

describes a distribution of charge localized in some finite volume V. In many situations we are interested in the potential $\Phi(\boldsymbol{x})$ resulting from this localized charge at distances sufficiently far away that $|\boldsymbol{x}| \gg |\boldsymbol{x}'|$ may be assumed, which suggests a series expansion of $1/|\boldsymbol{x}-\boldsymbol{x}'|$. Two types of expansions are commonly employed:

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

We will discuss here the multipole expansion in terms of Legendre polynomials or spherical harmonics. For a potential due to a unit point charge at $\boldsymbol{x}'$, we may expand $|\boldsymbol{x}-\boldsymbol{x}'|^{-1}$ as

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \sum_{l=0}^{\infty} \frac{r_<^l}{r_>^{l+1}} P_l(\cos\theta), \tag{3.48}$$

where $P_l(\cos\theta)$ is a Legendre polynomial [solution (4.36) of the Legendre equation (4.35)], $r_<$ is the smaller and $r_>$ the larger of $|\boldsymbol{x}|$ and $|\boldsymbol{x}'|$,[8] and θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$ (see Fig. 3.7). Equation (3.48) is termed a *multipole expansion*. In this expression

- 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.

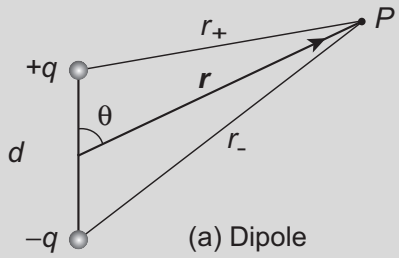
The reason for this terminology is suggested by the point-charge distributions discussed in Box 3.3. Comparison of these expressions with the terms in Eq. (3.48) using $r_> = r$ and $r_< = d$ (since $r \gg d$) suggests why Eq. (3.48) is called a multipole expansion: the $l = 1$ term is of the dipole 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 (3.48) can be expressed in terms of spherical harmonics using the *spherical*

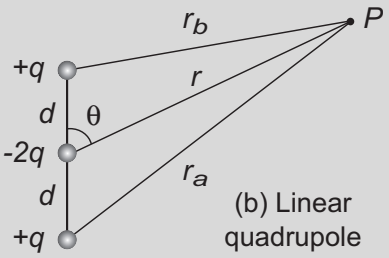
[8] The factor $r_<^l/r_>^{l+1}$ in Eq. (3.48) allows the expansion to be made in terms of either r/r' or r'/r, whichever is smaller.

Box 3.3 **Multipole moments**

What is the electric potential generated atthe point P by the following static charge distributions?



(a) Dipole



(b) Linear quadrupole

(a) 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.62), 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,

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

Hence the electric dipole potential is proportional to the Legendre polynomial $P_1(\cos\theta)$, and falls off at large distance as $1/r^2$. The potential contours and electric field vectors for a dipole are illustrated in Fig. 3.8.

(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 Legendre polynomial $P_2(\cos\theta)$ and varies as the cube of the inverse distance.

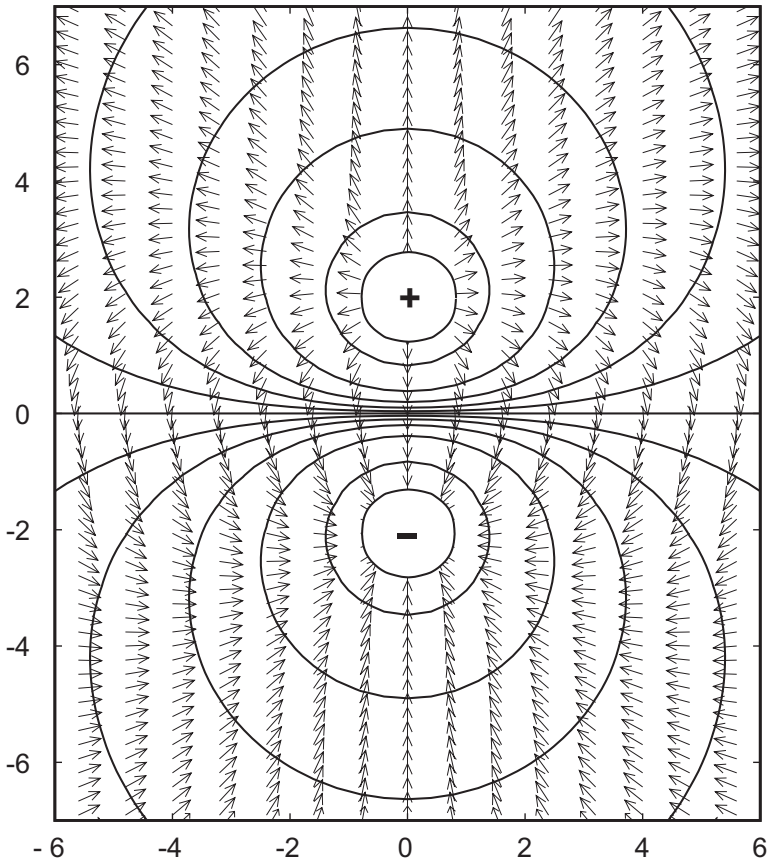


Fig. 3.8 Equipotential contours (curves) and the electric vector field (arrows) for the dipole charge distribution discussed in Box 3.3. For clarity, plotting of field lines and vectors has been suppressed very near the charges.

harmonic addition theorem

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1}\sum_{m=-l}^{l} Y_{lm}^*(\theta',\phi')\,Y_{lm}(\theta,\phi), \tag{3.49}$$

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

$$\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi-\phi'). \tag{3.50}$$

The spherical harmonic addition theorem (3.49) may then be used to express the expansion (3.48) of the potential at $\boldsymbol{x}$ due to a unit charge at $\boldsymbol{x}'$ as

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = 4\pi\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{1}{2l+1}\frac{r_<^l}{r_>^{l+1}}Y_{lm}^*(\theta',\phi')Y_{lm}(\theta,\phi), \tag{3.51}$$

where angles are defined in Fig. 3.7. 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. (3.47) is assumed to vanish outside of a small sphere of radius R centered on the origin, the potential generated by that charge outside the radius R can be expanded in spherical harmonics as

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{4\pi}{2l+1}\, q_{lm} \frac{Y_{lm}(\theta,\phi)}{r^{l+1}} \tag{3.52}$$

The expansion coefficients q_{lm} are given by

$$q_{lm} = \int Y^*_{lm}(\theta',\phi')(r')^l \rho(\boldsymbol{x}') d^3x' \tag{3.53}$$

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

$$Y_{lm}(\theta,\phi) = (-1)^m \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\phi}. \tag{3.54}$$

They behave under complex conjugation as

$$Y^*_{lm}(\theta,\phi) = (-1)^m Y_{l-m}(\theta,\phi), \tag{3.55}$$

and obey the *orthogonality condition*

$$\int_0^{2\pi} d\phi \int_0^{\pi} Y^*_{l'm'}(\theta,\phi) Y_{lm}(\theta,\phi) \sin\theta d\theta = \delta_{l'l}\delta_{m'm}, \tag{3.56}$$

and the *completeness relation*

$$\sum_{l=0}^{\infty} \sum_{m=-l}^{l} Y^*_{lm}(\theta',\phi') Y_{lm}(\theta,\phi) = \delta(\phi-\phi')\,\delta(\cos\theta - \cos\theta'). \tag{3.57}$$

The utility of a multipole expansion like Eqs. (3.52) or (3.48) is that the terms in the expansion of the source-charge distribution fall off with distance r as $r^{-(l+1)}$, justifying an approximation that retains only the first few terms in the multipole expansion.

If we define

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

$$\boldsymbol{p} \equiv \int \boldsymbol{x}' \rho(\boldsymbol{x}') d^3x', \tag{3.58}$$

3. and the *traceless quadrupole moment tensor* Q_{ij} by[9]

$$Q_{ij} \equiv \int (3x'_i x'_j - r'^2 \delta_{ij})\, \rho(\boldsymbol{x}') d^3x', \tag{3.59}$$

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

the spherical multipole moments in Eq. (3.53) may be expressed in cartesian coordinates as [19]

$$q_{00} = \frac{1}{\sqrt{4\pi}} \int \rho(\boldsymbol{x}')\, d^3x' = \frac{1}{\sqrt{4\pi}}\, Q, \tag{3.60a}$$

$$q_{11} = -\sqrt{\frac{3}{8\pi}} \int (x' - iy')\, \rho(\boldsymbol{x}')\, d^3x' = -\sqrt{\frac{3}{8\pi}}\, (p_x - ip_y), \tag{3.60b}$$

$$q_{10} = \sqrt{\frac{3}{4\pi}} \int z'\, \rho(\boldsymbol{x}') d^3x' = \sqrt{\frac{3}{4\pi}}\, p_z, \tag{3.60c}$$

$$q_{22} = \frac{1}{4}\sqrt{\frac{15}{2\pi}} \int (x' - iy')^2\, \rho(\boldsymbol{x}')\, d^3x' = \frac{1}{12}\sqrt{\frac{15}{2\pi}}\, (Q_{11} - 2iQ_{12} - Q_{22}), \tag{3.60d}$$

$$q_{21} = -\sqrt{\frac{15}{8\pi}} \int z'(x' - iy')\, \rho(\boldsymbol{x}')\, d^3x' = -\frac{1}{3}\sqrt{\frac{15}{8\pi}}\, (Q_{13} - iQ_{23}), \tag{3.60e}$$

$$q_{20} = \frac{1}{2}\sqrt{\frac{5}{4\pi}} \int (3z'^2 - r'^2)\, \rho(\boldsymbol{x}')\, d^3x' = \frac{1}{2}\sqrt{\frac{5}{4\pi}}\, Q_{33}, \tag{3.60f}$$

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 $1/|\boldsymbol{x} - \boldsymbol{x}'|$ (see Appendix A.8). The result

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \left[\frac{Q}{r} + \frac{\boldsymbol{p}\cdot\boldsymbol{x}}{r^3} + \frac{1}{2} \sum_{i,j} Q_{ij}\, \frac{x_i x_j}{r^5} + \cdots \right]. \tag{3.61}$$

is quoted without proof.[10]

The spherical multipole moments like Eq. (3.53) and the corresponding cartesian multipole moments like Eq. (3.58) generally have different numbers 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 level (a consequence of group theory; see Box 18.1) the source of 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

[9] With this definition Q_{ij} is traceless (see Problem 3.6). A quadrupole moment definition without the tracelessness property $Q_{ij} = \frac{1}{2}\int \rho(\vec{x}') x'_i x'_j \, \mathrm{d}^3x'$ 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.

[10] The mathematical difference between the cartesian multipole expansion and spherical multipole moment expansion is that the former is obtained by expanding $1/|\boldsymbol{x} - \boldsymbol{x}'|$ in a Taylor series (see Appendix A.8), while the latter is obtained by expanding $1/|\boldsymbol{x} - \boldsymbol{x}'|$ in a a binomial series (see Appendix A.9 and Problem 3.2). The spherical moment expansion is favored in fields like atomic or nuclear physics, where angular momentum conservation is important.

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. (3.53) 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 [19].

Example 3.3 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 symmetric, so only Y_{lm} with $m = 0$ are non-zero. From Eq. (3.53), 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,

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

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

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

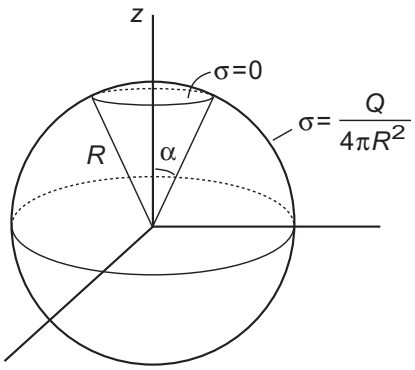


Fig. 3.9 Figure for Example 3.4: a spherical surface has a uniform surface charge of density $\sigma = Q/4\pi R^2$, except for a spherical cap at the north pole defined by a cone with opening $\theta = \alpha$ where $\sigma = 0$.

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. (3.60).

Example 3.4 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. 3.9. Use the jump condition at a charge layer of Eq. (3.2) for the electric field

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

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 3.1),

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

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. (3.47) and

(3.48),

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

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 [2]

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

(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,

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

for the inside solution, and for the outside solution,

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

The preceding example illustrates many techniques that can be used in solving electrostatic problems by multipole expansions.

3.7.2 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. (3.52) with definite (l,m) the spherical electric field components are [19]

$$E_r = \frac{l+1}{(2l+1)\varepsilon_0}\, q_{lm}\, \frac{1}{r^{l+2}}\, Y_{lm}(\theta,\phi), \tag{3.62a}$$

$$E_\theta = -\frac{1}{(2l+1)\varepsilon_0}\, q_{lm}\, \frac{1}{r^{l+2}}\, \frac{\partial}{\partial\theta}\, Y_{lm}(\theta,\phi), \tag{3.62b}$$

$$E_\phi = \frac{1}{(2l+1)\varepsilon_0}\, q_{lm}\, \frac{1}{r^{l+2}}\, \frac{im}{\sin\theta}\, Y_{lm}(\theta,\phi), \tag{3.62c}$$

with the multipole moments q_{lm} defined in Eq. (3.60).

Example 3.5 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 (3.62).

3.7.3 Energy of a Charge Distribution in an External Field

If a localized charge distribution $\rho(\boldsymbol{x})$ is subject to an *external potential* $\Phi(\boldsymbol{x})$,[11] the electrostatic energy is

$$W = \int \rho(\boldsymbol{x})\Phi(\boldsymbol{x})d^3x. \tag{3.63}$$

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

$$\begin{aligned}\Phi(\boldsymbol{x}) &= \Phi(0) + \boldsymbol{x}\cdot\boldsymbol{\nabla}\Phi(0) + \frac{1}{2}\sum_i\sum_j x_i x_j \frac{\partial^2\Phi}{\partial x_i \partial x_j}(0) + \cdots \\ &= \Phi(0) - \boldsymbol{x}\cdot\boldsymbol{E}(0) - \frac{1}{2}\sum_i\sum_j x_i x_j \frac{\partial E_j}{\partial x_i}(0) + \cdots \\ &= \Phi(0) - \boldsymbol{x}\cdot\boldsymbol{E}(0) - \frac{1}{6}\sum_i\sum_j \left(3x_i x_j - r^2\delta_{ij}\right)\frac{\partial E_j}{\partial x_i}(0) + \cdots,\end{aligned} \tag{3.64}$$

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 (3.64) in Eq. (3.63), the energy takes the form

$$W = q\Phi(0) - \boldsymbol{p}\cdot\boldsymbol{E}(0) - \frac{1}{6}\sum_i\sum_j Q_{ij}\frac{\partial E_j}{\partial x_i}(0) + \cdots, \tag{3.65}$$

where q is the total charge, the dipole moment $\boldsymbol{p}$ is defined in Eq. (3.58), and the quadrupole moment Q_{ij} is defined in Eq. (3.59).

Example 3.6 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 (3.65) 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.

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

(The energy shifts are small, corresponding typically to radiofrequency wavelengths.) Such methods allow the quadrupole moments of nuclei to be measured, which provide 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) is another way in which nuclear quadrupole deformations can be measured by analyzing their response to an applied electric field.

The expansion (3.65) 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.

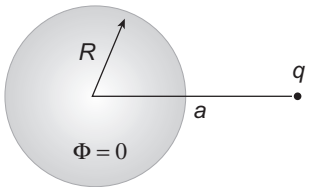
Thus multipole expansions serve a pedagogical as well as practical purpose in understanding electrostatic interactions.

Background and Further Reading

An introduction to the material of this chapter may be found in Griffiths [13] or Purcell and Morin [32]. More advanced treatments may be found in Jackson [19], Garg [11], Chaichian et al [5], and Zangwill [42].

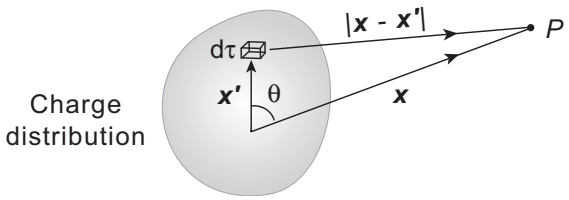
Problems

3.1 A charge q is placed outside a grounded ($\Phi = 0$) conducting sphere:



Use the method of images described in Section 3.5 to find the scalar potential Φ 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

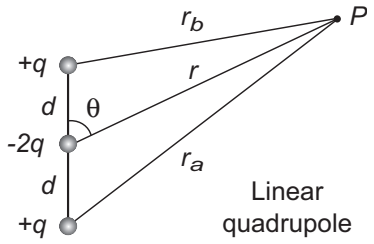
$$\varepsilon \equiv \frac{r'}{r}\left(\frac{r'}{r} - 2\cos\theta\right),$$

and that this is a series expansion in terms of Legendre polynomials that is equivalent to Eq. (3.51). Finally, show that

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

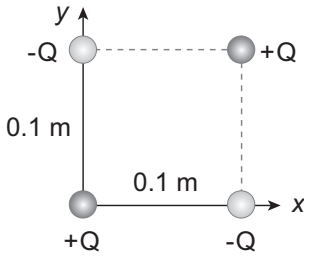
is the scalar potential at the point P.

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 charge distribution shown below,



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

3.5 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. *Hint*: See Eq. (A.76).

3.6 The quadrupole moment tensor is given by Eq. (3.59) 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.

3.7 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. (3.59). Assume the origin of the coordinate system to be at the lower left charge and that $|Q| = 3\,\mu$C.

3.8 Prove that in the Green function problem described in Box 3.2,

$$x(t) = \int_{-\infty}^{\infty} dt' G(t,t')\frac{f(t')}{m}$$

is a solution of the equation of motion

$$\left(\frac{d^2}{dt^2} + 2\gamma\frac{d}{dt} + \omega_0^2\right)x(t) = \frac{f(t)}{m},$$

for a driven mechanical oscillator. ***

3.9 Show that the uniqueness proof for solutions of the Laplace equation given in Section 3.3.2 carries through in similar fashion for the Poisson equation.

3.10 Prove Uniqueness Theorem II given in Section 3.3 by showing that if two different electric fields are assumed to correspond to solutions of the Poisson equation with the same boundary conditions, then the difference between them must be zero. *Hint*: Gauss's law in differential and integral form, general properties of conductors, and some vector calculus voodoo will do the job. ***

4 Solving the Poisson and Laplace Equations

For relatively simple electrostatics problems solutions may be found using Coulomb's law directly, Gauss's theorem, image methods, and so on, as we have shown in Chs. 2 and 3. 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. This chapter discusses some means to obtain solutions for those equations.

4.1 Solutions by Separation of Variables

It is more difficult to find systematic solutions for the Poisson and Laplace partial differential equations than for ordinary differential equations. 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 the Laplace or Poisson equation subject to boundary conditions by the *separation of variables method*.

4.1.1 Separation of Variables in Cartesian Coordinates

Let us introduce the separation of variables method using cartesian coordinates. The basic idea is to assume that the solution can be written in the product form

$$\Phi(x,y,z) = X(x)Y(y)Z(z), \tag{4.1}$$

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. 4.1 [7], where it is necessary to solve for the scalar potential Φ subject to boundary conditions given by specifying Φ on three conducting surfaces. 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,

$$\frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} = 0, \tag{4.2}$$

subject to the boundary conditions

$$\Phi = 0 \qquad (y = 0, y = a), \tag{4.3a}$$

$$\Phi = \Phi_0 \qquad (x = 0), \tag{4.3b}$$

$$\Phi \to 0 \qquad (x \to \infty). \tag{4.3c}$$

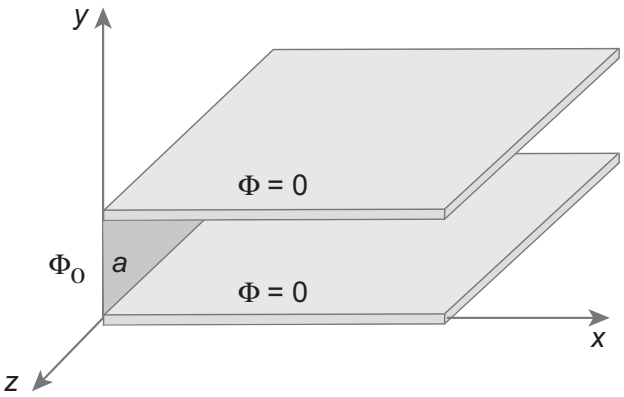


Fig. 4.1 Grounded ($\Phi = 0$) semi-infinite plane-parallel electrodes terminated by a planar electrode at potential Φ_0. The electrodes are assumed to extend infinitely to the right and to be infinite in the direction perpendicular to the page. (Maintaining the difference in potential between the plates at $\Phi = 0$ and $\Phi = \Phi_0$ requires a strip of insulator at the joints.) Adapted from Ref. [7].

Inserting the 2D product function

$$\Phi(x,y) = X(x)Y(y) \tag{4.4}$$

into Eq. (4.2) and dividing through by $\Phi = X(x)Y(y)$ gives

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} = 0, \tag{4.5}$$

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. (4.5) 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,

$$\frac{1}{X}\frac{d^2X}{dx^2} = C_1 \qquad \frac{1}{Y}\frac{d^2Y}{dy^2} = C_2. \tag{4.6}$$

Equation (4.5) 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*

$$\frac{d^2X}{dx^2} - k^2X = 0 \tag{4.7a}$$

$$\frac{d^2Y}{dy^2} + k^2Y = 0. \tag{4.7b}$$

As can be verified by substitution, the second equation (4.7b) has a solution

$$Y = A\sin(ky) + B\cos(ky), \tag{4.8}$$

where A and B are arbitrary constants. We may determine constants by imposing the boundary conditions (4.3). From Eq. (4.3a), $\Phi = 0$ at $y = 0$ requires that $B = 0$, and $\Phi = 0$ at $y = a$

requires that

$$k = \frac{n\pi}{a} \qquad (n = 1,2,3,\cdots). \tag{4.9}$$

Therefore,

$$Y = A\sin\left(\frac{n\pi y}{a}\right) \qquad (n = 1,2,3,\cdots). \tag{4.10}$$

The equation for X in Eq. (4.7a) now takes the form

$$\frac{d^2X}{dx^2} - \left(\frac{n\pi}{a}\right)^2 X = 0, \tag{4.11}$$

which has a solution

$$X = Ge^{n\pi x/a} + He^{-n\pi x/a}, \tag{4.12}$$

as may be verified by substitution. However, the boundary condition $\Phi \to 0$ as $x \to \infty$ of Eq. (4.3c) can be satisfied only if $G = 0$, so we discard the first term of Eq. (4.12) and insert Eqs. (4.10) and (4.12) into Eq. (4.4) to give

$$\Phi(x,y) = C\sin\left(\frac{n\pi y}{b}\right)e^{-n\pi x/a}, \tag{4.13}$$

with C another arbitrary constant.

The solution (4.13) satisfies the boundary conditions (4.3a) and (4.3c), but does not satisfy the boundary condition (4.3b). 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 (4.13) in the form

$$\Phi(x,y) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right)e^{-n\pi x/a}, \tag{4.14}$$

The boundary condition (4.3b) at $x = 0$ implies that

$$\Phi(0,y) = \Phi_0 = \sum_{n=0}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right), \tag{4.15}$$

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. (4.15) by $\sin[(pxy)/a]$, where p is an integer, and integrate from $y = 0$ to $y = a$,

$$\int_0^a \Phi_0 \sin\left(\frac{p\pi y}{a}\right)dy = \int_0^a \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right)\sin\left(\frac{p\pi y}{a}\right)dy. \tag{4.16}$$

For the integral on the left side of this equation

$$\int_0^a \Phi_0 \sin\left(\frac{p\pi y}{a}\right)dy = \begin{cases} \dfrac{2a\Phi_0}{p\pi} & (\text{if } p \text{ is odd}), \\ 0 & (\text{if } p \text{ is even}), \end{cases} \tag{4.17}$$

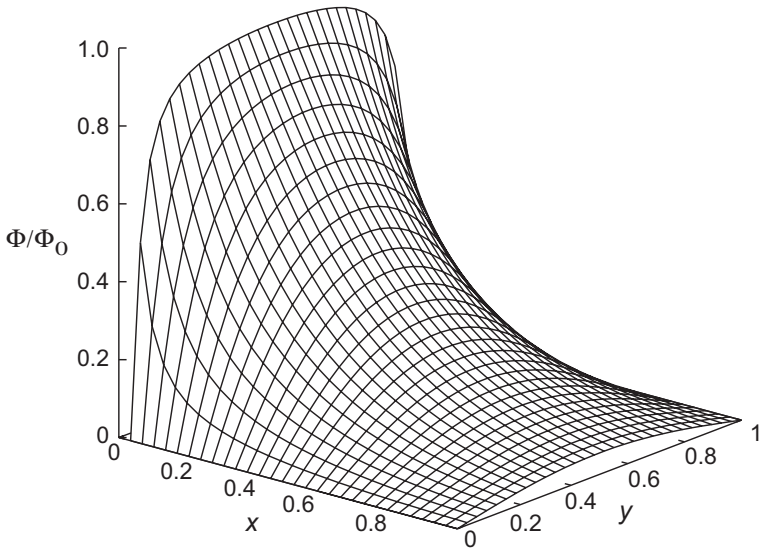


Fig. 4.2 Solution (4.21) of the 2D Laplace equation for the problem in Fig. 4.1, obtained by separation of variables.

while for the integral on the right side,

$$\int_0^a c_n \sin\left(\frac{n\pi y}{a}\right) \sin\left(\frac{p\pi y}{a}\right) dy = \begin{cases} 0 & (\text{if } p \neq n), \\ \dfrac{a}{2} c_n & (\text{if } p = n), \end{cases} \tag{4.18}$$

Comparing Eqs. (4.17) and (4.18), we conclude that the expansion coefficients are given by

$$c_n = \begin{cases} \dfrac{4\Phi_0}{n\pi} & (\text{if } n \text{ is odd}), \\ 0 & (\text{if } n \text{ is even}), \end{cases} \tag{4.19}$$

and the potential as a function of x and y is

$$\Phi(x,y) = \frac{4\Phi_0}{n\pi} \sum_{n=1,3,5,\cdots}^{\infty} \frac{1}{n} \sin\left(\frac{nxy}{a}\right) e^{-n\pi x/a}. \tag{4.20}$$

This series should converge quickly because of the rapid decrease of the factor $e^{-n\pi x/a}/n$ with n. A convergent power series is a perfectly adequate way to define a solution, but it happens that the series (4.20) can be summed exactly to give the more convenient closed form [13],

$$\Phi(x,y) = \frac{2\Phi_0}{\pi} \tan^{-1}\left(\frac{\sin(\pi y/a)}{\sinh(\pi x/a)}\right), \tag{4.21}$$

This solution is plotted in Fig. 4.2.

Example 4.1 Consider Fig. 4.3, where we wish to calculate the electrostatic potential in the region between the parallel-plane electrodes [7]. There is no z dependence so again

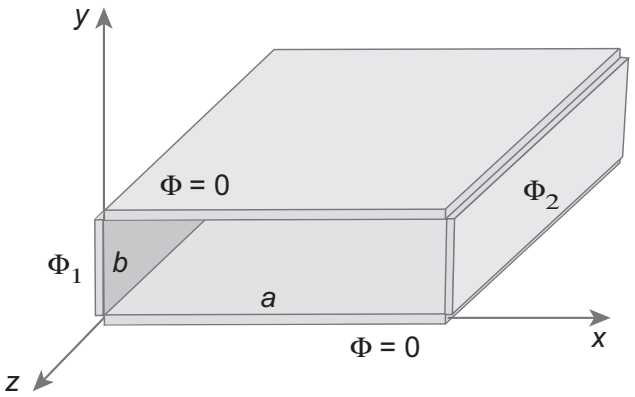


Fig. 4.3 Grounded ($\Phi = 0$) parallel-plane electrodes of width a and separated by a distance b are terminated on two sides with plane electrodes at potentials Φ_1 and Φ_2. The electrodes are assumed to be infinite in the direction perpendicular to the page. Adapted from Ref. [7].

let's solve the 2D Laplace equation by separation of variables. This problem has much in common with the example from Fig. 4.1 just worked out, but the boundary conditions are different.

$$\Phi = 0 \qquad (y = 0, y = b), \tag{4.22a}$$

$$\Phi = \Phi_1 \qquad (x = 0), \tag{4.22b}$$

$$\Phi = \Phi_2 \qquad (x = a), \tag{4.22c}$$

The most general solution is of the form

$$\Phi(x,y) = \sum_{n=1}^{\infty} \left(A_n e^{-n\pi x/b} + B_n e^{n\pi x/b} \right) \sin\left(\frac{n\pi y}{b}\right), \tag{4.23}$$

where the constants A_n and B_n may be determined using the boundary conditions. As you are asked to show in Problem 4.3, the coefficients A_n and B_n are given by

$$A_n = \frac{4}{n\pi}\left(\frac{\Phi_1 - \Phi_2 e^{-n\pi a/b}}{1 - e^{-2n\pi a/b}}\right) \qquad B_n = \frac{4e^{-n\pi a/b}}{n\pi}\left(\frac{\Phi_2 - \Phi_1 e^{-n\pi a/b}}{1 - e^{-2n\pi a/b}}\right), \tag{4.24}$$

where $n = 1, 3, 5, \cdots$. This gives $\Phi(x,y)$ when inserted in Eq. (4.23). This solution is plotted in Fig. 4.4.

Notice in Figs. 4.2 and 4.4 the property asserted in Section 3.2.2 that solutions of the Laplace or Poisson equations can have maxima or minima only on the boundaries of the domain, with no local maxima or minima permitted within the volume of the solution.

4.1.2 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

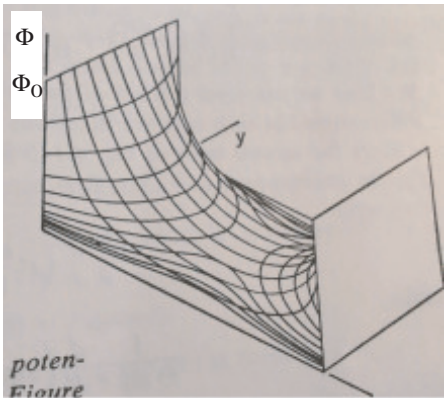


Fig. 4.4 Solution the 2D Laplace equation for the problem in Fig. 4.3 by separation of variables.(Excuse quality; temporary placeholder).

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. The Laplace equation in spherical coordinates (r,θ,ϕ) is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2\Phi}{\partial\phi^2}=0, \tag{4.25}$$

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

$$\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right)+\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right)=0. \tag{4.26}$$

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

$$\Phi(r,\theta)=R(r)\,\Theta(\theta), \tag{4.27}$$

where $R(r)$ is a function only of r and $\Theta(\theta)$ is a function only of θ. Substituting Eq. (4.27) into Eq. (4.26) and dividing through by $R\Theta$ gives,

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)=0, \tag{4.28}$$

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. (4.28) 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

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = k \tag{4.29a}$$

$$\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) = -k \tag{4.29b}$$

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. (4.29a). Multiplying both sides by R and carrying out the leftmost d/dr operation gives

$$r^2\frac{d^2R}{dr^2} + 2r\frac{dR}{dr} - kR = 0, \tag{4.30}$$

which has a solution

$$R = Ar^n + \frac{B}{r^{n+1}}, \tag{4.31}$$

that when inserted in Eq. (4.29a) requires n and k to be related by,

$$n(n+1) = k. \tag{4.32}$$

Now consider the Θ equation (4.29b). Multiplying by $\Theta\sin\theta$ and using Eq. (4.32), it may be written

$$\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + n(n+1)\sin\theta\,\Theta = 0 \tag{4.33}$$

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

$$\frac{df}{d\theta} = \frac{df}{d\mu}\frac{d\mu}{d\theta} = -\sin\theta\frac{df}{d\mu} = -\sqrt{1-\mu^2}\,\frac{df}{d\mu} \tag{4.34}$$

where we have used $d\mu/d\theta = -\sin\theta$ and $1-\mu^2 = \sin^2\theta$. Then, Eq. (4.33) becomes

$$\frac{d}{d\mu}\left[(1-\mu^2)\frac{d\Theta}{d\mu}\right] + n(n+1)\Theta = 0 \qquad \text{(Legendre's equation).} \tag{4.35}$$

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. [1] Letting $\cos\theta \equiv x$, normalized Legendre polynomials[2] are typically defined by *Rodrique's formula*

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n. \tag{4.36}$$

[1] Mathematically, Legendre's differential equation (4.35) arises naturally from solution of the Laplace and related partial differential equations using separation of variables in spherical coordinates. The eigenfunctions of the angular part of the Laplacian operator acting on Φ on the left side of Eq. (4.25) are spherical harmonics (see Section 3.7.1); modulo multiplicative constants, Legendre polynomials are the subset of spherical harmonics left invariant by rotations about the polar axis. Viewed in this way, the Legendre polynomials are intimately connected to rotational symmetry. Indeed, many of their properties are found most easily using symmetry and group theory methods, which in turn implies that they have deep physical and geometrical meaning.

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

Table 4.1 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	$\frac{1}{8}(63\cos^5\theta - 70\cos^3\theta + 15\cos\theta)$

Some low-order Legendre polynomials are tabulated in Table 4.1. A general solution of Laplace's equation in spherical coordinates assuming axial symmetry is then given by

$$\Phi(r,\theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\theta) + \sum_{n=0}^{\infty} B_n r^{-(n+1)} P_n(\cos\theta) \tag{4.37}$$

The Legendre polynomials form a complete set, so arbitrary boundary conditions for problems with axial symmetry can be satisfied using them. Furthermore, they satisfy the orthogonality condition

$$\int_{-1}^{+1} P_m(x) P_n(x) dx = \frac{2}{2n+1}\,\delta_{mn}, \tag{4.38}$$

or expressed explicitly in spherical coordinates,

$$\begin{aligned}\int_{-1}^{+1} P_m(\cos\theta) P_n(\cos\theta) d(\cos\theta) &= \int_0^{\pi} P_m(\cos\theta) P_n(\cos\theta)\sin\theta\, d\theta \\ &= \frac{2}{2n+1}\,\delta_{mb} = \begin{cases} \dfrac{2}{2n+1} & (\text{if } m = n,) \\ 0 & (\text{if } m \neq n), \end{cases}\end{aligned} \tag{4.39}$$

which are important in evaluating the coefficients in the expansion Eq. (4.37).

Example 4.2 Consider the conducting ball of radius a in an electric field illustrated in Fig. 4.5. 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 2D Laplace equation for the exterior field by separation of variables in spherical coordinates, assuming axial symmetry. We take as boundary conditions

$$\Phi = 0 \qquad (r = a) \tag{4.40}$$

$$\Phi = -E_0 r\cos\theta \qquad (r = \infty) \tag{4.41}$$

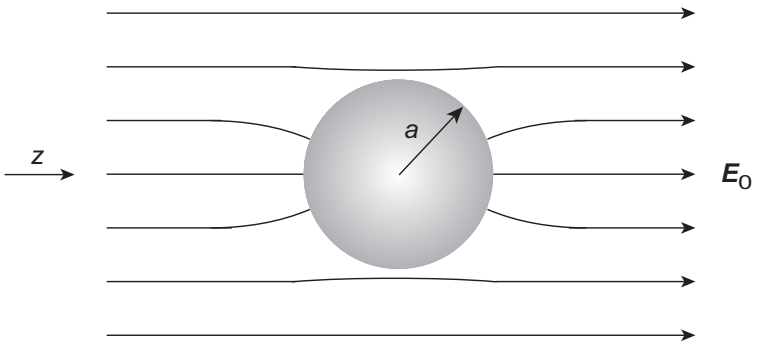


Fig. 4.5 A conducting ball of radius a in an external electric field $\boldsymbol{E}_0$ directed along the $\boldsymbol{z}$ axis, with axial symmetry assumed around the z axis.

At the radius of the ball $r = a$, from Eqs. (4.37) and (4.40)

$$\Phi(r,\theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\theta) + \sum_{n=0}^{\infty} B_n r^{-(n+1)} P_n(\cos\theta) = 0. \tag{4.42}$$

The coefficients A_n and B_n may be evaluated using the orthogonality relation (4.39) for the Legendre polynomials. Multiplying Eq. (4.42) by $P_m(\cos\theta)$ and integrating over $d(\cos\theta)$, from Eq. (4.39) the only non-vanishing terms are those for which $n = m$, implying that

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

which requires the coefficients to be related by

$$B_n = -A_n a^{2n+1}. \tag{4.43}$$

Therefore, from Eqs. (4.42) and (4.43),

$$\Phi(r,\theta) = \sum_{n=0}^{\infty} A_n \left(r^n - \frac{a^{2n+1}}{r^{n+1}}\right) P_n(\cos\theta). \tag{4.44}$$

Now as $r \to \infty$ the second term in parentheses in Eq. (4.44) becomes negligible compared with the first and the boundary condition (4.41) requires that

$$-E_0 r\cos\theta = -E_0 r P_1(\cos\theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\theta), \tag{4.45}$$

where we have used $P_1(\cos\theta) = \cos\theta$ from Table 4.1. 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. (4.43) all the B_n are zero except for

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

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

$$\begin{aligned}\Phi(r,\theta) &= -E_0 r\cos\theta + E_0\frac{a^3\cos\theta}{r^2} \\ &= -E_0\left(1-\frac{a^3}{r^3}\right)r\cos\theta, \end{aligned} \tag{4.46}$$

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. (4.46) by taking the gradient [see Eq. (A.47) for the gradient in spherical coordinates],

$$E_r = -\frac{\partial\Phi}{\partial r} = E_0\left(1+\frac{2a^3}{r^3}\right)\cos\theta, \tag{4.47}$$

$$E_\theta = -\frac{1}{r}\frac{\partial\Phi}{\partial\theta} = -E_0\left(1-\frac{a^3}{r^3}\right)\sin\theta. \tag{4.48}$$

At the surface of the conductor we know that

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

so solving for σ and using Eq. (4.47) with $r = a$ gives

$$\sigma = 3\varepsilon_0 E_0\cos\theta, \tag{4.49}$$

for the induced charge density.

4.1.3 Separation of Variables in Cylindrical Coordinates

Material to follow.

4.2 Variational Methods

Material to follow.

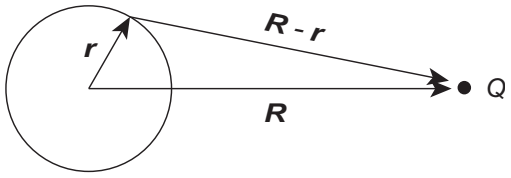
Background and Further Reading

An introduction to the material of this chapter may be found in Griffiths [13] or Purcell and Morin [32]. More advanced treatments may be found in Jackson [19], Garg [11], Chaichian et al [5], and Zangwill [42].

Problems

4.1 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.

4.2 One form of *Earnshaw's theorem* (see Section 3.2.1) states that, for any electric field $\boldsymbol{E}$, the average of the scalar potential Φ over any sphere that lies entirely in a charge-free region is equal to the value of the potential at the center of the sphere. For a charge Q outside a sphere of radius $r_0 = |\boldsymbol{r}|$ as in the following diagram,



the potential averaged over the sphere $\langle\Phi\rangle$ is given by

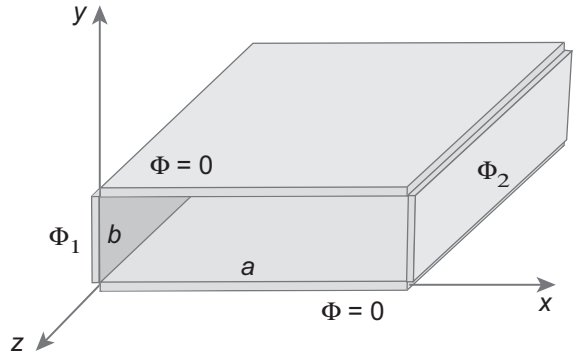
$$\langle\Phi\rangle = \frac{1}{4\pi r_0^2}\int \frac{1}{4\pi\varepsilon_0}\frac{Q}{|\boldsymbol{R}-\boldsymbol{r}|}\, r^2\delta(r-r_0)dr\,d\Omega.$$

This could be integrated directly, but instead prove Earnshaw's theorem by evaluating the integral using the spherical harmonic expansion

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

given in Eq. (3.51). ***

4.3 Consider the following figure ,



which illustrates grounded ($\Phi = 0$) parallel-plane electrodes of width a and separated by a distance b, which are terminated on the left and right sides with plane electrodes at potentials Φ_1 and Φ_2, and the top and bottom sides are held at potentials $\Phi = 0$.

The electrodes are assumed to be infinite in the direction perpendicular to the page. Calculate the potential in the region between the parallel-plane electrodes.

4.4 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 problem will be easy if you work Problem 4.1 first. ***

5 Electrostatics in Conducting Matter

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 having charge carriers that are tightly bound and do not transport electrical charge well, and
2. *conductors*, which have delocalized charge carriers that are free to transport electrical charge efficiently; conductors are also often called *metals*.

In atomic matter the charge carriers are typically electrons, electron holes, or ions, but for purposes of being definite in discussion, we shall normally assume electrons to be the charge carriers, unless otherwise stated.

5.1 Properties of Conductors

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

5.1.1 Electric Fields Inside a Conductor Vanish

Intuitively $\boldsymbol{E} = 0$ inside a conductor because if there were an electric field inside the conductor it would accelerate electrons, violating the assumption of electrostatic equilibrium. More precisely, consider Fig. 5.1(a) where a conducting slab is immersed in an external electric field $\boldsymbol{E}$ oriented horizontally. Initially the external field will attract negative charges to the left side of the slab, leaving a net positive charge on the right side. This *polarization of charge* within the slab will create an internal electric field $\boldsymbol{E}'$ that opposes the external field $\boldsymbol{E}$ (remember the convention that the electric field vector points from positive to negative charge). This piling up of negative charges on one side and positive charges on the opposite side in response to application of an external electric field is called an *induced*

[1] There are also things in between such as *semimetals* or *semiconductors* that transport charge poorly, or only under certain conditions, some materials like Mott insulators having different mechanisms than simple insulators, or superconductors that do a super job of carrying current without resistance. We will leave finer details of classification to materials science for now and concentrate on the electromagnetic behavior of simple and idealized conductors and insulators.

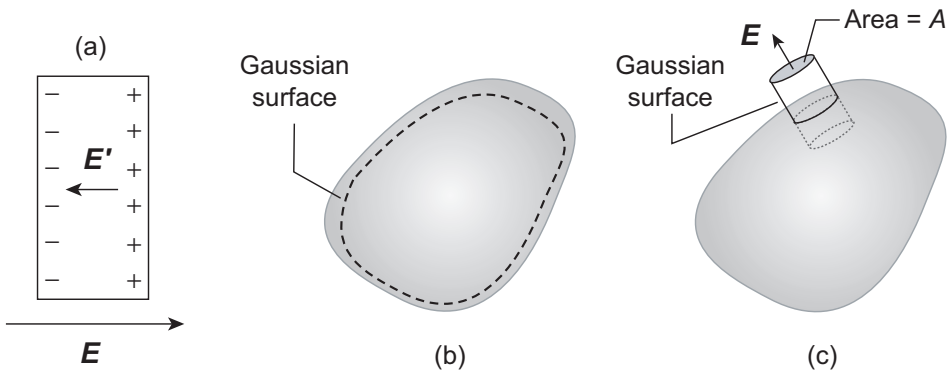


Fig. 5.1 (a) Conducting slab in a uniform electric field $\boldsymbol{E}$. (b) Gaussian surface (dashed curve) inside a conductor. (c) Cylindrical Gaussian surface perpendicular to the surface of a conductor.

charge. Charge will continue to flow within the slab until the induced internal electric field $\boldsymbol{E}'$ exactly cancels the external field $\boldsymbol{E}$, 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 oppose and cancel the external field typically occurs on a timescale sufficiently short that it can be assumed to be instantaneous for most considerations.

5.1.2 The Charge Density Is Zero Inside a Conductor

Just as the electric field vanishes inside a conductor, the charge density ρ is also zero. Consider the Gaussian surface shown in Fig. 5.1(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 charge enclosed by the surface. Hence the absence of an electric field means necessarily that the charge density $\rho = 0$ inside a conductor.

5.1.3 Excess Charge Resides on the Surface of a Conductor

That excess charge must reside at the surface of the conductor is a consequence of Gauss's law and follows immediately from the requirement that $\rho = 0$ inside the conductor. Since the Gaussian surface (dashed curve) in Fig. 5.1(b) can be place arbitrarily close to the surface, it follows that any excess charge can exist only at the surface of a conductor.

5.1.4 Any Exterior Electric Field is Normal to the Surface

No electric field exists inside a conductor; if any electric field exists outside the conductor it is necessarily

- generated by the excess surface charge, and is
- perpendicular to the surface of the conductor.

Consider Fig. 5.1(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 the electric field $\boldsymbol{E}$ 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 also is no flux through the flat face of the cylinder *inside the conductor*, because $\boldsymbol{E}$ is zero there. Hence any net flux passes only through the flat face of the Gaussian cylinder outside the conductor. Applying Gauss's law to this surface,

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

where σ is the surface charge density and A is the area of the endplate of the cylinder. It follows that any electric field must (1) be outside the conductor, (2) have a *magnitude proportional to the surface charge density*, and (3) have an *orientation normal to the surface*.

5.1.5 A Conductor is an Equipotential

A conductor necessarily has the same same potential Φ at each point in the conductor. This follows because for any two points either within 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 $\boldsymbol{E} = 0$ in the interior and on the surface.

5.1.6 Surface Charge Accumulates Where Curvature Is Large

If a conductor is of irregular shape, the surface charge density σ will be greatest where the surface has the *largest local curvature* (that is, the *smallest radius of curvature*). Consider an irregularly shaped conductor, as in Fig. 5.1(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 greater the local surface charge density σ.

5.2 Capacitance

From the equations of electrostatics the source charge Q is proportional to the scalar potential Φ, with the constant proportionality termed the *capacitance* C. For an isolated con-

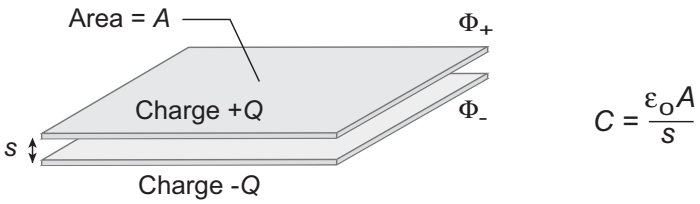


Fig. 5.2 A parallel plate capacitor.

ductor carrying a charge Q with potential Φ, the capacitance is

$$C = \frac{Q}{\Phi}. \tag{5.1}$$

In the SI system of units the charge is measured in coulombs, the potential in volts, and the capacitance in *farads* (F),[2] with $1\,\mathrm{F} \equiv 1\,\mathrm{coulomb/volt}$.

5.2.1 Parallel Plate Capacitors

Capacitance is also a useful concept in dealing with the charges and potentials associated with more than one conductor. The generic example is the *parallel plate capacitor* illustrated in Fig. 5.2. 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 the difference in potential for two conductors denoted by V,

$$V \equiv \Phi_+ - \Phi_-, \tag{5.2}$$

the capacitance for a parallel plate capacitor is given by

$$C = \frac{Q}{V} \qquad \text{(Two conductors)}. \tag{5.3}$$

For the capacitor in Fig. 5.2 the capacitance depends only on its geometry,

$$C = \frac{\varepsilon_0 A}{s} \qquad \text{(Parallel plate capacitor)}, \tag{5.4}$$

where A is the surface area of a plate and s is the separation between the plates (which we assume to be parallel to each other, and of equal area) .

As will be discussed in Section 6.1, the capacitance of a parallel plate capacitor like that shown in Fig. 5.2 can be increased significantly by replacing the gap between the electrodes with a layer of insulating (dielectric) material, such as teflon. This is a consequence of the electric field polarizing the charge distribution in the dielectric layer lying between the plates (see Box 6.1).

[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.

5.2.2 Energy Stored in a Capacitor

Charging a capacitor (for example, by connecting the two plates in Fig. 5.2 to the poles of a battery) stores energy in the electric field that is created between the oppositely-charged electrodes of the device. The amount of energy in the field may be determined by computing the work done to charge the capacitor to a particular level. Consider the parallelplate capacitor of Fig. 5.2. If at some point in the charging process the charge on the positive plate is q, from Eqs. (2.37) and (5.3) the work increment dW required to add the next charge increment dq is given by

$$dW = \left(\frac{q}{C}\right) dq, \tag{5.5}$$

and the total work that must be done to charge the plate from $q = 0$ to $q = Q$ is given by integration,

$$W = \int dW = \int_0^Q \left(\frac{q}{C}\right) dq = \frac{Q^2}{2C}. \tag{5.6}$$

Therefore, using $Q = CV$ from Eq. (5.3), the work required to charge to a potential difference V between the electrodes is

$$W = \frac{1}{2} CV^2, \tag{5.7}$$

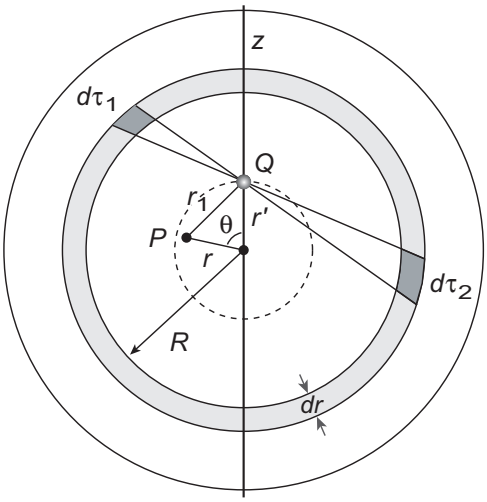
From Eq. (5.4), this takes the specific form

$$W = \frac{1}{2} \frac{A\varepsilon_0}{d} V^2, \tag{5.8}$$

for a parallel plate capacitor, where A is the area of a plate and d is the separation of the plates.

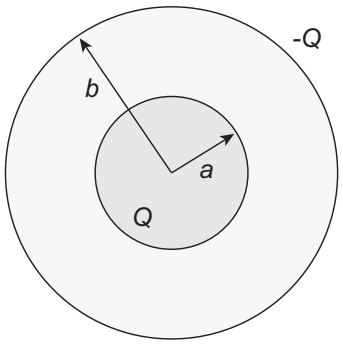
Problems

5.1 Consider a sphere containing a single point charge Q located a distance r' from the origin along the z-axis.



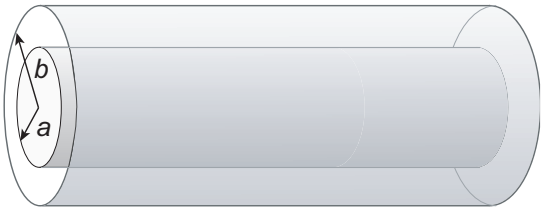
Calculate the average electric field inside the sphere. *Hint*: Separate the volume of the sphere τ into the part outside Q plus the part inside Q, as indicated by the dashed circle in the figure, and show by qualitative argument that the contribution from outside the dashed circle is zero.

5.2 Consider the spherical capacitor shown in the following figure,



where the outer spherical conducting shell has radius b and charge $-Q$, and the innner spherical conducting shell has radius a and charge $+Q$. What is the capacitance of this device? ***

5.3 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.

6 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, this chapter begins 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.

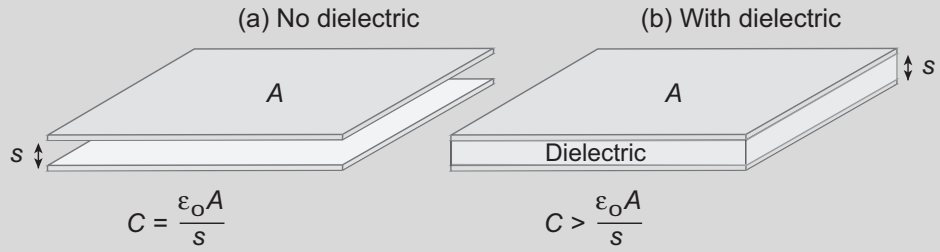
6.1 Dielectrics

As illustrated in Box 6.1, 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

$$C = \frac{Q}{\Phi_{12}} = \frac{\varepsilon_0 A}{s}, \tag{6.1}$$

Box 6.1 Capacitors and Dielectrics

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



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 of an electric field in a dielectric, as described in Section 6.1.1.

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.

6.1.1 Capacitors and Dielectrics

Now suppose that a layer of dielectric material is placed between the capacitor plates, as illustrated in Box 6.1. The capacitance may still be defined by the ratio of the charge to the potential difference between the plates, $C = Q/\Phi_{12}$, but now the *actual value of C will be increased* over that found for no dielectric between the plates. The presence of the dielectric between the plates allows more charge Q to accumulate 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. Assuming the charge on the upper plate to be positive, negative charges in the dielectric will be pulled upward and positive charges pushed down, as illustrated in Fig. 6.1.
3. This exposes a layer of uncompensated negative charge near the top and a layer of uncompensated positive charge near the bottom of the dielectric.[1]

[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.

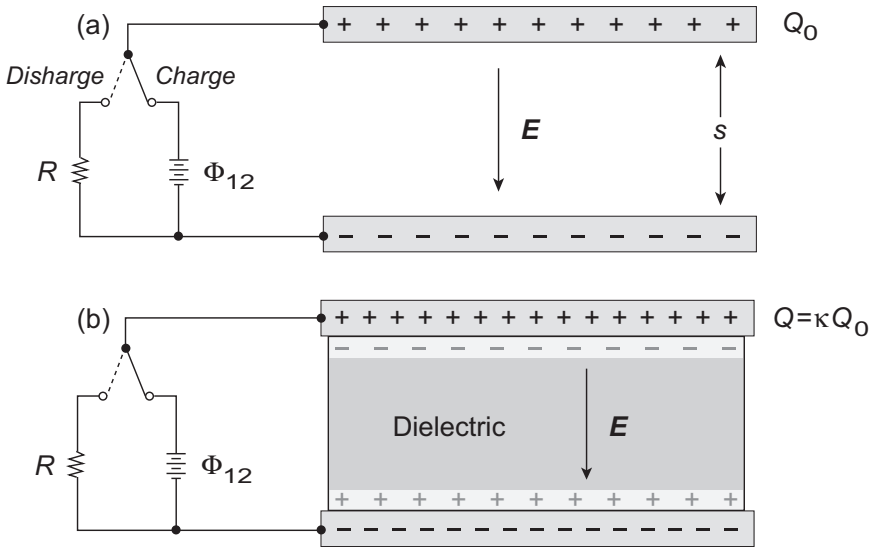


Fig. 6.1 Effect on capacitor of a dielectric between the plates. (a) No dielectric. (b) With dielectric. The switch allows charging and discharging the capacitor. Adapted from Ref. [32].

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 increases until the algebraic sum of Q and the induced charge in the upper layer is equal to the total charge on the top plate Q_0 before the dielectric was inserted.
6. It follows that the total charge Q in the top layer is *larger than the charge* Q_0 of the top plate in Fig. 6.1(a) before the dielectric was inserted.
7. Thus, the charge is the Q that appears in Eq. (6.1) and is, in the circuit of Fig. 6.1, the charge supplied by the battery to charge the capacitor.

The induced charge layer is not part of the charge Q appearing in Eq. (6.1). If the switch in the circuit of Fig. 6.1 is used to discharge the fully-charged capacitor though the resistance R, the charge Q will be dissipated.

Thus the induced charge present in the charged capacitor is absorbed back into the normal structure of the dielectric in the discharged capacitor.

6.1.2 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 in the dielectric matter 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

Table 6.1 Dielectric constants κ for some substances

Substance	Conditions	κ
Vacuum	(Definition)	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

that material:

$$Q = \kappa Q_0 \quad \longleftrightarrow \quad C = \kappa C_0. \tag{6.2}$$

Dielectric constants are ratios of charges and thus dimensionless; a few values of κ are given in Table 6.1 for some representative substances. The dielectric constant of the vacuum is defined to be $\kappa = 1$, typical gases have κ slightly larger than one, and liquids and solids can have dielectric constants varying widely in the range $\kappa \sim 1 - 100$. The remarkably large value $\kappa = 80.4$ for water merits an explanation that will be given later.

6.1.3 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 intrinsically with the dielectric are termed *bound charges*; those charges that are mobile and not bound in the dielectric are termed *free charges.* The *bound charges are not mobile* because they are attached physically to the atoms and molecules making up the dielectric.[2] *Bound charges can be polarized by electric fields* causing tiny charge displacements in the charged capacitor, 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); *the bound charges are not part of the charge Q on the capacitor plates*. Conversely, the free charges may be viewed as those charges that we have some agency over in an experiment (for example, through manipulation of the electrical circuit in Fig. 6.1 by using the switch to charge or discharge the capacitor).

[2] We assume that fields are not large enough to cause ionization and breakdown of the dielectric (possibly accompanied by smoke and fire, which would be unseemly in a theoretical discusssion).

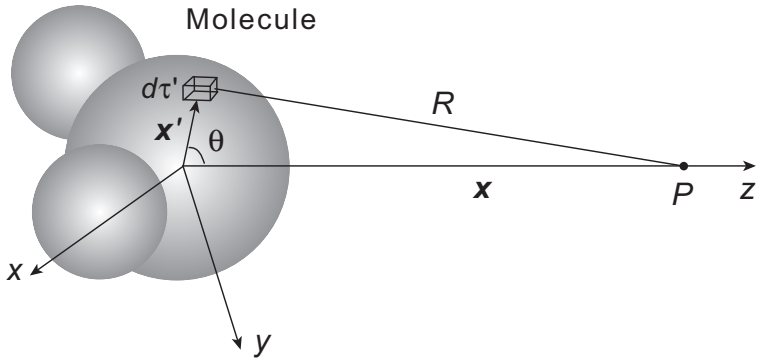


Fig. 6.2 Diagram for calculation of the potential at point P produced by a molecular charge distribution.

6.2 Moments of a Molecular Charge Distribution

Let's consider the potential Φ_P at a distant point P produced by the molecular electrical charge distribution of Fig. 6.2,

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}')}{R}\,d\tau', \tag{6.3}$$

where the integral is over all of the charge distribution. From the law of cosines (A.62) the distance R is given by

$$R = (r^2 + r'^2 - 2rr'\cos\theta)^{1/2}, \tag{6.4}$$

where we define $r = |\boldsymbol{x}|$ and $r' = |\boldsymbol{x}'|$, and Eq. (6.3) becomes

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\int (r^2 + r'^2 - 2rr'\cos\theta)^{-1/2}\rho(\boldsymbol{x}')\,d\tau'. \tag{6.5}$$

Let us now expand this expression for the potential in multipole moments.

6.2.1 Multipole Expansion of the Potential

For a distant point P we have $r \gg r'$ and a binomial expansion gives (see Problem 3.2)

$$\frac{1}{R} = (r^2 + r'^2 - 2rr'\cos\theta)^{-1/2} = \frac{1}{r}\left[1 + \frac{r'}{r}\cos\theta + \left(\frac{r'}{r}\right)^2 \frac{3\cos^2\theta - 1}{2} + \mathcal{O}\left(\left[\frac{r'}{r}\right]^3\right)\right]. \tag{6.6}$$

Then from Eqs. (6.3) and (6.6) the potential can be written as

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\left[\frac{1}{r}\int \rho\,d\tau' + \frac{1}{r^2}\int r'\cos\theta\rho\,d\tau' + \frac{1}{r^3}\int r'^2\frac{3\cos^2\theta - 1}{2}\rho\,d\tau' + \cdots\right], \tag{6.7}$$

where the constant $r = |\boldsymbol{x}|$ 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 the potential can be written as a power series (multipole expansion) with constant coefficients

$$\Phi_P = \frac{1}{4\pi\varepsilon_0}\left[\frac{C_0}{r} + \frac{C_1}{r^2} + \frac{C_2}{r^3} + \cdots\right], \tag{6.8}$$

where the coefficients C_i are the integrals over the internal charge distribution in Eq. (6.7). The electric field then follows from $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi_P$.

6.2.2 Dominence of Monopole and Dipole Terms

At large distance from the source, the rapid convergence of terms in Eq. (6.8) implies that only low-order multipoles dominate. Behavior of the electric potential Φ at large distance from the source typically will be *dominated by the first term that has a non-zero coefficient* in the multipole expansion (6.8). Let us examine the coefficients in Eq. (6.8). The monopole coefficient $C_0 = \int \rho(\boldsymbol{x}')d\tau'$ is the total charge. For a neutral atom or molecule, $C_0 = 0$. If the atom or molecules is ionized so that $\rho \neq 0$, the monopole term 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 since $C_0 = 0$. Furthermore, if the charge distribution is charge-neutral the value of C_1 is independent of the choice of origin.[3]

As we shall see, for our main task here of understanding the behavior of dielectrics, in a multipole expansion 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 all multipole moments of the charge distribution of order beyond 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 shall now consider induced moments in Section 6.3 and permanent dipole moments in Section 6.4.

6.3 Induced Dipole Moments

The simplest atom is hydrogen, consisting of one nucleus (a proton) and one electron. The nucleus is so small compared with the electron cloud that in hydrogen and in more complicated atoms and molecules the nuclei can be approximated as point charges. Quantum

[3] Generally, the value of a multipole moment may depend on the origin chosen for the coordinate system. But for the special case of a dipole moment in a charge-neutral dielectric, the value of the dipole moment is always independent of the choice of origin.

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 classical radii and shapes.

6.3.1 Polarization of the Electron Cloud

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 atomic charge cloud will be polarized, with the negative electrons pulled down and the positive nucleus pushed 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 6.1 estimates the size of this distortion in actual atoms.

Example 6.1 A very rough estimate of how much distortion will be caused by a field of strength E may be made 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 it is assumed that a field of the same order of magnitude would be required to produce significant distortion, the distortion may be estimated to be

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

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

The *dipole moment vector* $\boldsymbol{p}$ induced by an electric field $\boldsymbol{E}$ will point in the direction of the field and the magnitude of the induced dipole moment is generally proportional to the electric field (at least if the field is not too strong, which we assume to be the case).

6.3.2 Atomic Polarizabilities

The factor relating the dipole moment vector $\boldsymbol{p}$ to $\boldsymbol{E}$ is termed the *atomic polarizability* α,

$$\boldsymbol{p} = \alpha \boldsymbol{E}. \tag{6.9}$$

More generally, in a molecules the polarization may be asymmetric with respect to different axes and the scalar coefficient α in this equation must be replaced by a *polarizability tensor*, as described in Box 6.2. 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.[4] Some atomic polarizabilities are given in Table 6.2, listed in order of atoms with increasing total electron number. The large variations in polarizabilities that are evident in Table 6.2 may be attributed to differences in total electron number and in valence electronic structure. For example, the noble gas elements He, Ne, and Ar have 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 of an increase in the total electron number. As another example, the elements Na and K each have one loosely bound electron outside a closed electronic shell and that electron is very susceptible to perturbation by an electric field. As a result, Na and K have very large polarizabilities ($\kappa = 27$ and $\kappa = 34$, respectively).

Example 6.2 Let's estimate the amount of charge displacement in an atom induced by a typical electric field. From the polarizability of atomic hydrogen given in Table 6.2, the magnitude of the dipole moment induced by an electric field of strength one megavolt / meter is $p < 10^{-34}$ coulomb-meters.

Thus the characteristic atomic and molecular polarizations induced by applied electric fields are extremely small.

6.4 Permanent Dipole Moments of Polar Molecules

Some molecules have asymmetric shapes and exhibit permanent dipole moments in their normal ground state, even in the absence of an external field; a few examples are shown in Fig. 6.3. Molecules that have a permanent dipole moment are termed *polar molecules*. As a rule, the intrinsic dipole moments of polar molecules are much larger than the dipole moments induced by laboratory electric fields for either non-polar or polar molecules.

[4] Beware: both α and $\alpha/4\pi\varepsilon_0$ are sometimes called the "atomic polarizability" in the literature.

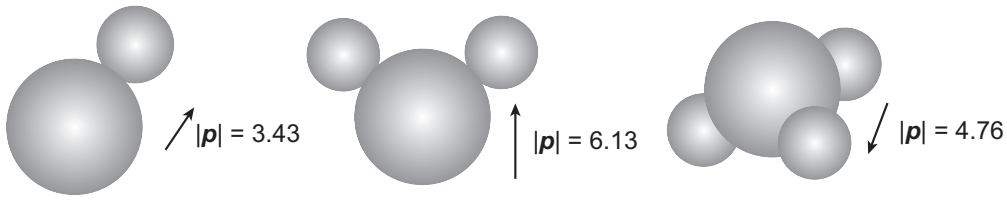
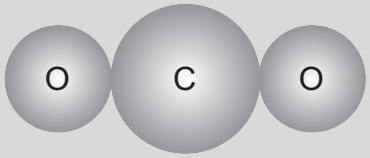


Fig. 6.3 Approximate geometry of the electron cloud and the observed dipole moment vector $\boldsymbol{p}$ for some polar molecules. Magnitudes $p = |\boldsymbol{p}|$ of the dipole moment vector are in units of 10^{-30} coulomb-meters (adapted from Ref. [32]).

Box 6.2 **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 (6.9) 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. (6.9) 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.

Example 6.3 As shown in Fig. 6.3, the magnitude of the (permanent) dipole moment for the polar molecule HCl is $p = |\boldsymbol{p}| = 3.43 \times 10^{-30}$ coulomb-meters (which is equivalent to shifting the charge of one electron by a distance of 2×10^{-9} cm [32]). From Example 6.2, 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 are typically multiple orders of magnitude larger than those induced by laboratory electric fields.

Table 6.2 Atomic polarizabilities $(\alpha/4\pi\varepsilon_0)$ in units of $10^{-30}\,\mathrm{m}^3$

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

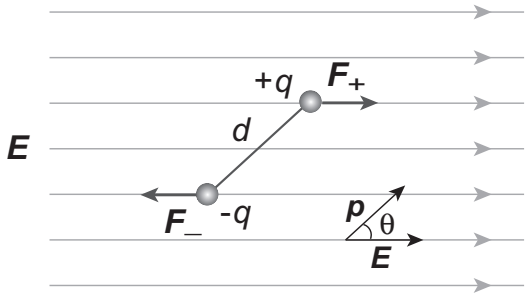


Fig. 6.4 Torque applied to a molecule with an intrinsic dipole moment in a uniform electric field.

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. 6.4), but the torque $\boldsymbol{N}$ acting on the dipole is

$$\boldsymbol{N} = \boldsymbol{p} \times \boldsymbol{E}, \tag{6.10}$$

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 with the field.

6.5 Polarization Density

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.[5] The net effect of either is to produce a set of dipoles aligned or partially aligned with the electric field. Our primary interest here is in the effect of this polarization, without regard to the mechanism by which it was produced. Therefore, a measure of polarization for matter can be formed by asking how many dipoles N of average dipole moment $\boldsymbol{p}$ there are in a unit volume, without regard to their

[5] An applied field induces polarization for polar and non-polar molecules. However, any induced polarization for a polar molecule is usually very small compared with its intrinsic polarization and can be neglected.

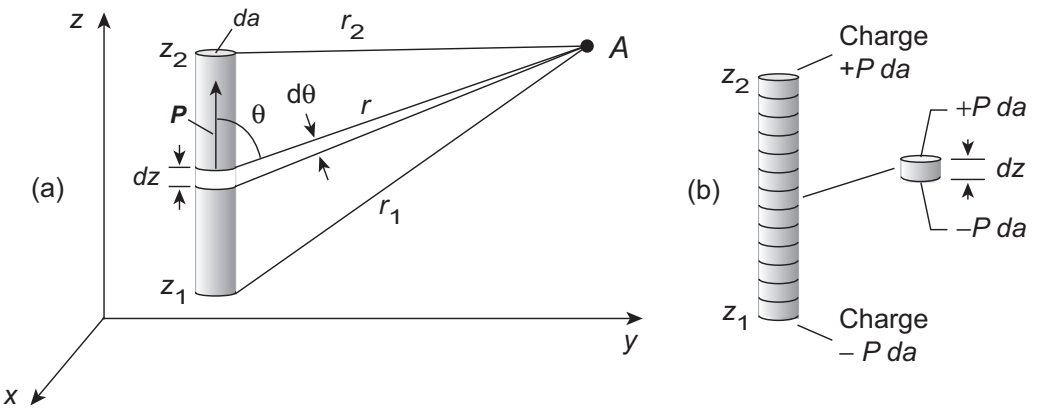


Fig. 6.5 (a) A column of polarized material with cross section da observed at a distant point A produces the same field as two charges, one at each end of (b). Adapted from Ref. [32].

source. 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

$$\boldsymbol{P} \equiv \boldsymbol{p}N = \left(\frac{\text{Number of dipole moments}}{\text{Unit volume}}\right), \tag{6.11}$$

which has units of C-m/m^3 = C/m^2. The *polarization-charge density* ρ_{pol} is given by minus the divergence of the polarization density,

$$\rho_{\text{pol}}(\boldsymbol{x}) = -\boldsymbol{\nabla}\cdot\boldsymbol{P}(\boldsymbol{x}). \tag{6.12}$$

The electric field is discontinuous across the surface of the dielectric (see Sections 6.9 and 6.11), so there must be a surface charge density given by the difference between the density (6.12) evaluated for the two media at the boundary that is given by

$$\sigma_{\text{pol}} = -(\boldsymbol{P}_2 - \boldsymbol{P}_1)\cdot\hat{\boldsymbol{n}} \tag{6.13}$$

where $\boldsymbol{P}_i$ is the polarization in Medium i and $\hat{\boldsymbol{n}}$ is a unit vector that is normal to the boundary and points from Medium 1 to Medium 2.

6.6 Field Outside Polarized Dielectric Matter

Let's now estimate the electric potential associated with the polarized dielectric material, considering separately the regions outside and inside of the dielectric matter. If it is assumed that the dielectric was assembled from neutral matter so that there is no net charge, there will be no monopole term in the multipole expansion (6.8). Therefore to 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. 6.5. The polarization $\boldsymbol{P}$ is uniform and points in the positive z direction, and we wish to calculate

the electric potential at the point A. An element of the cylinder of height dz in Fig. 6.5(a) has a dipole moment

$$\boldsymbol{p} = \boldsymbol{P} d\tau = \boldsymbol{P}\, da\, dz, \tag{6.14}$$

and its contribution to the potential at point A is

$$d\Phi_A = \frac{P\cos\theta\, da\, dz}{4\pi\varepsilon_0 r^2}. \tag{6.15}$$

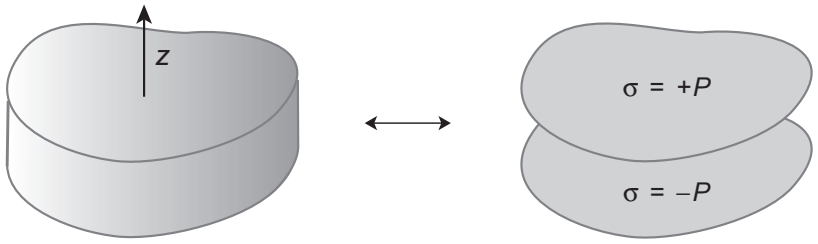
Then the potential at the point A produced by the entire column of polarized matter is obtained by integration,

$$\begin{aligned}\Phi_A &= \frac{P\,da}{4\pi\varepsilon_0}\int_{z_1}^{z_2}\frac{dz\cos\theta}{r^2} \\ &= \frac{-P\,da}{4\pi\varepsilon_0}\int_{z_1}^{z_2}\frac{dr}{r^2} \\ &= \frac{Pda}{4\pi\varepsilon_0}\left(\frac{1}{r_2}-\frac{1}{r_1}\right).\end{aligned} \tag{6.16}$$

But this the same as a potential produced by a positive point charge Pda at the top of the column at a distance r_2 from A and a negative point charge $-Pda$ placed at the bottom of the column a distance r_1 from A (see the dipole potential example in Box 3.3). 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. (6.16) plausible by the heuristic argument illustrated in Fig. 6.5(b). Consider making the cylinder shown in Fig. 6.5(a) by stacking on top of each other small cylinder segments of height dz, with a charge of $+Pda$ on its top face and $-Pda$ 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 uncompensated charge on its top face of $+Pda$, and the bottom-most segment, which will have an uncompensated charge on its bottom face of $-Pda$. Thus the column of tiny dipoles will appear to be a single large dipole with end charges $+Pda$ and $-Pda$, 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. 6.5(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 were located at the positions of the top and bottom of the slab, carrying constant surface charge density $\sigma = +P$ and $\sigma = -P$, respectively. See Fig. 6.6.



(a) Block of polarized dielectric (b) Top and bottom layers of surface charge

Fig. 6.6 (a) A block of dielectric polarized by an electric field in the z direction. (b) Two sheets of charge $\pm P$ at the positions of the top and bottom surfaces of the block in (a).

6.7 Field Inside Polarized Dielectric Matter

It may be expected that 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*.[6] The spatial average of $\boldsymbol{E}$ over a volume V inside the polarized matter is given by

$$\langle \boldsymbol{E} \rangle_V = \frac{\int \boldsymbol{E}\, d\tau}{\int d\tau} = \frac{1}{V} \int \boldsymbol{E}\, d\tau, \tag{6.17}$$

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. (6.17). Let us summarize some important properties of the macroscopic (that is, averaged) electric field.

1. The fundamental relation (2.27) that is obeyed by the microscopic electric field,

$$\boldsymbol{\nabla} \times \boldsymbol{E} = 0 \tag{6.18}$$

   remains valid for the macroscopic electric field.
2. This implies that the macroscopic electric field is still *derivable from a scalar potential*, through $\boldsymbol{E} = -\boldsymbol{\nabla}\boldsymbol{\Phi}$ (see Section 2.5).
3. 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 again dipole,

[6] 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 basic 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 operations. A more detailed discussion of the procedure of averaging microscopic quantities to produce macroscopic quantities may be found in Section 6.6 of Jackson [19] and Section 3.1 of Wald [40].

with an electric *polarization* $\boldsymbol{P}$ corresponding to the density of dipoles in Eq. (6.11),

$$\boldsymbol{P}(\boldsymbol{x}) = \sum_i N_i \langle \boldsymbol{p}_i \rangle,$$

where $\boldsymbol{p}_i$ is the dipole moment of the ith species (atoms or molecules), the average $\langle \cdots \rangle$ 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 $\boldsymbol{x}$.

6.8 The Electric Displacement $\boldsymbol{D}$

The macroscopic potential or field can be built up 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. (3.61), the macroscopically averaged potential is

$$\Delta\Phi(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{4\pi\varepsilon_0}\left[\frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\Delta V + \frac{\boldsymbol{P}(\boldsymbol{x}')\cdot(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\Delta V\right], \tag{6.19}$$

assuming $\boldsymbol{x}$ to lie outside ΔV. Setting $\Delta V \to d^3x'$ and integrating over all space gives the potential

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int d^3x' \left[\frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} + \boldsymbol{P}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)\right], \tag{6.20}$$

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

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int d^3x' \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\left[\rho(\boldsymbol{x}') - \boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')\right], \tag{6.21}$$

which represents the potential generated by an effective charge distribution $\tilde{\rho}(\boldsymbol{x}') = \rho(\boldsymbol{x}') - \boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')$.[7] Then Gauss's law (1.1a) reads

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = \frac{1}{\varepsilon_0}\left[\rho - \boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')\right], \tag{6.22}$$

which reduces to Eq. (1.1a) in the absence of polarization. If we define the *electric displacement* $\boldsymbol{D}$ by

$$\boldsymbol{D} \equiv \varepsilon_0\boldsymbol{E} + \boldsymbol{P}, \tag{6.23}$$

Eq. (6.22) can be written as

$$\boldsymbol{\nabla}\cdot\boldsymbol{D} = \rho. \tag{6.24}$$

Solution for potentials or fields requires that the relations that connect the displacement $\boldsymbol{D}$ and the electric field $\boldsymbol{E}$ be specified.

[7] 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 6.9. As was noted in Eq. (6.12), the polarization-charge density at a point $\boldsymbol{x}$ in polarized media is given by $\rho_{\rm pol}(\boldsymbol{x}) = -\boldsymbol{\nabla}\cdot\boldsymbol{P}(\boldsymbol{x})$.

6.8.1 Constituitive Relationships

The relationships that connect D and E are called *constitutive relations*. It is common to make two assumptions in these relationships.

1. The response to the applied field is linear, $\boldsymbol{P} \propto \boldsymbol{E}$.
2. The medium is isotropic, so that $\boldsymbol{P}$ is parallel to $\boldsymbol{E}$ and the coefficient of proportionality has no angular dependence,

$$\boldsymbol{P} = \varepsilon_0 \chi_e \boldsymbol{E}, \tag{6.25}$$

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

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} \qquad \varepsilon \equiv \varepsilon_0(1+\chi_e), \tag{6.26}$$

where ε is the *electric permittivity* and

$$\kappa \equiv \frac{\varepsilon}{\varepsilon_0} = 1 + \chi_e \tag{6.27}$$

is called the *dielectric constant*. Then the polarization can be written

$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E}. \tag{6.28}$$

If the dielectric medium is *uniform in addition to being isotropic*, then ε is independent of position and the divergence equation (6.24) can be written

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon}, \tag{6.29}$$

which is Gauss's law (1.1a) with ε_0 replaced by $\varepsilon = \varepsilon_0(1+\chi_e)$. Equations (6.24) and (6.18) are the macroscopic counterparts of the microscopic equations (1.1a) and (2.27), respectively.

For a medium fulfilling all the conditions specified 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 in the medium that oppose the applied field.

One consequence of immediate relevance to our prior discussion is that the capacitance of a parallel-plate capacitor is increased by a factor $\varepsilon/\varepsilon_0$ if the empty space between the electrodes is filled with a material having dielectric constant $\varepsilon/\varepsilon_0$ (neglecting the effect of fringing fields). Example 6.4 illustrates.

Example 6.4 From Eq. (6.26), inserting a dielectric layer in a parallel-plate capacitor alters the electric field, changing the capacitance from C_0 to

$$C = \kappa C_0 = \frac{\varepsilon}{\varepsilon_0} C_0 = (1-\chi_e)C_0. \tag{6.30}$$

For the capacitor described in Section 5.2.1, the capacitance with a dielectric layer between the plates is

$$C = \frac{\varepsilon}{\varepsilon_0} C_0 = \frac{\varepsilon}{\varepsilon_0} \frac{A\varepsilon_0}{d} = \frac{A\varepsilon}{d}. \tag{6.31}$$

The corresponding work required to charge a parallel-plate capacitor that was given in Section 5.2.2 is modified to

$$W = \frac{Q^2}{2C} = \frac{1}{2} CV^2 = \frac{1}{2} \frac{A\varepsilon}{d} V^2, \tag{6.32}$$

which is $\varepsilon/\varepsilon_0$ times the work required to charge the capacitor without the dielectric layer that was given in Eq. (5.8).

6.8.2 Boundary Conditions

If a system contains different media, boundary conditions must be considered for both $\boldsymbol{D}$ and $\boldsymbol{E}$ at all interfaces between different media. This will be addressed in more depth in Section 6.11, but 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 either static or time-varying fields)

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n} = \sigma, \tag{6.33a}$$

$$(\boldsymbol{E}_2 - \boldsymbol{E}_1) \times \boldsymbol{n} = 0, \tag{6.33b}$$

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).

6.9 Surface and Volume Bound Charges

The potential created by a single dipole can be written

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \frac{\boldsymbol{p} \cdot \hat{\boldsymbol{r}}}{r^2}, \tag{6.34}$$

where from Fig. 6.7, $r = |\boldsymbol{r}|$ where $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$, and $\boldsymbol{p}$ is the dipole moment vector. The total potential contributed by dipoles then follows by integration,

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \int_V \frac{\boldsymbol{P}(\boldsymbol{x}') \cdot \hat{\boldsymbol{r}}}{r^2} \, d\tau', \tag{6.35}$$

where $d\tau' \equiv d^3x'$ and $\boldsymbol{P}$ is the dipole moment density. This can be rewritten using

$$\boldsymbol{\nabla}' \left(\frac{1}{r} \right) = \frac{\hat{\boldsymbol{r}}}{r^2} \tag{6.36}$$

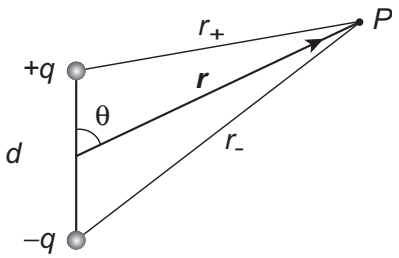


Fig. 6.7 Coordinate system for an electric dipole. The distance from the dipole to P is $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$ and $r \equiv |\boldsymbol{r}|$.

in the form

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \boldsymbol{P}\cdot\boldsymbol{\nabla}'\left(\frac{1}{r}\right)d\tau', \tag{6.37}$$

which can be integrated 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

$$\int_V \boldsymbol{\nabla}'\cdot\left(\frac{\boldsymbol{P}}{r}\right)d\tau' = \int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau' + \int_V (\boldsymbol{P}\cdot\boldsymbol{\nabla}')\frac{1}{r}d\tau'. \tag{6.38}$$

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. (6.38) to give

$$\oint_S \frac{1}{r}\boldsymbol{P}\cdot\boldsymbol{n}\,da' = \int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau' + \int_V (\boldsymbol{P}\cdot\boldsymbol{\nabla}')\frac{1}{r}d\tau', \tag{6.39}$$

multiply both sides by $1/4\pi\varepsilon_0$ and rearrange to give

$$\frac{1}{4\pi\varepsilon_0}\int_V (\boldsymbol{P}\cdot\boldsymbol{\nabla}')\frac{1}{r}d\tau' = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{r}\boldsymbol{P}\cdot\boldsymbol{n}\,da' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau', \tag{6.40}$$

and finally, upon comparing with Eq. (6.37),

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{r}\boldsymbol{P}\cdot\boldsymbol{n}\,da' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau'. \tag{6.41}$$

Thus we see that

1. The first term of Eq. (6.41) looks like the potential generated by a surface charge

$$\sigma_{\rm b} \equiv \boldsymbol{P}\cdot\hat{\boldsymbol{n}}, \tag{6.42}$$

with $\hat{\boldsymbol{n}}$ the unit vector normal to the surface and the subscript "b" indicating that it originates in the bound charges.

2. The second term of Eq. (6.41) looks like the potential generated by a volume charge of magnitude

$$\rho_b = -\boldsymbol{\nabla}\cdot\boldsymbol{P}. \tag{6.43}$$

Hence Eq. (6.41) can be written

$$\Phi = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{\sigma_b}{r}\,da' + \frac{1}{4\pi\varepsilon_0}\int_V \frac{\rho_b}{r}\,d\tau', \tag{6.44}$$

and the potential Φ and electric field $\boldsymbol{E}$ of a polarized object are equivalent to that produced by a volume charge density $\rho_b = -\boldsymbol{\nabla}\cdot\boldsymbol{P}$ plus a surface charge density $\sigma_b = \boldsymbol{P}\cdot\hat{\boldsymbol{n}}$. 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. (6.21)].

Example 6.5 Let's use the result of Eq. (6.44) 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. (6.42),

$$\sigma_b = \boldsymbol{P}\cdot\hat{\boldsymbol{n}} = P\cos\theta, \tag{6.45}$$

where the z-axis has been chosen as the direction of polarization. Thus, we need to evaluate the potential for a sphere with the surface charge (6.45) painted on it. This problem was already solved in Problems 4.1 and 4.4, with the result that

$$\Phi(r,\theta) = \begin{cases} \dfrac{Pr}{3\varepsilon_0}\cos\theta & (r \le R), \\ \dfrac{PR^3}{3\varepsilon_0 r^2}\cos\theta & (r \ge R). \end{cases} \tag{6.46}$$

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:

$$\boldsymbol{E} = -\boldsymbol{\nabla}\Phi = -\frac{P}{3\varepsilon_0}\hat{\boldsymbol{z}} = -\frac{1}{3\varepsilon_0}\boldsymbol{P} \qquad (r < R). \tag{6.47}$$

Outside the sphere the field has a dipole form

$$\Phi = \frac{1}{4\pi\varepsilon_0}\frac{\boldsymbol{p}\cdot\hat{\boldsymbol{r}}}{r^2} \qquad (r \ge R), \tag{6.48}$$

where the dipole moment is equal to the total dipole moment of the sphere,

$$\boldsymbol{p} = \frac{4}{3}\pi R^3\boldsymbol{P}, \tag{6.49}$$

since $\boldsymbol{P}$ is the dipole moment density.

Curves of constant electric field strength for the electric field of the uniformly polarized sphere computed in Example 6.5 are plotted in Fig. 6.8.

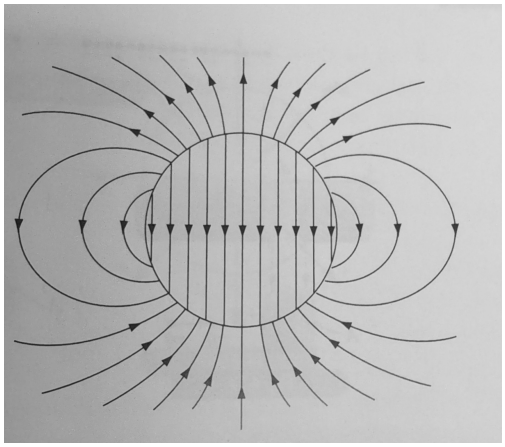


Fig. 6.8 Electric field for a uniformly polarized sphere determined in Example 6.5.

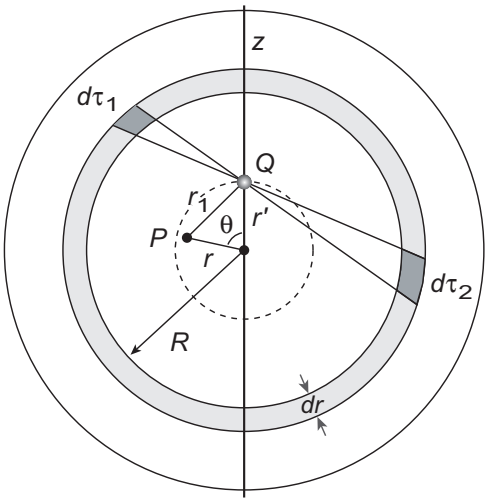


Fig. 6.9 A point charge Q located at a distance r' from the center of a sphere of radius R.

6.10 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. Hence one fundamental task will be to compute the electric field averaged over 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. 6.9 [7]. By symmetry the average field over the entire volume must be along the z-axis The average field is then

$$\langle E_z \rangle = \frac{\int_\tau E_z \, d\tau}{\int_\tau d\tau} = \frac{1}{\tau} \int_\tau E_z \, d\tau,$$

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. 6.9. 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. 6.9 that $r_1 = |\boldsymbol{x}-\boldsymbol{x}'|$. Thus, we may expand in the multipole expansion given in Eq. (3.48),

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

where we've used from Fig. 6.9 that $r < r'$. Writing this expansion out and substituting the explicit values for the Legendre polynomials from Table 4.1,

$$\Phi = \frac{Q}{4\pi\varepsilon_0}\frac{1}{r'}\left[1+\frac{r}{r'}\cos\theta+\frac{1}{2}\frac{r^2}{r'^2}(3\cos^2\theta-1)+\frac{1}{2}\frac{r^3}{r'^3}(5\cos^3\theta-3\cos\theta+\cdots\right].$$

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.

6.11 Boundary Conditions at Interfaces

In advanced applications one often must consider problems in which media with different properties are spatially adjacent and boundary conditions at all interfaces must be accounted for. We may do so by using the vacuum Maxwell equations modified to reflect the influence of the medium on classical electrodynamics. As we now discuss, the in-medium Maxwell equations may be expressed in either differential or integral form.

6.11.1 Differential Form of Maxwell Equations in Medium

In a medium that may be polarized by electric or magnetic fields, the *vacuum Maxwell equations* (1.1) are modified to the *Maxwell equations in medium*, expressed in differential form,

$$\boldsymbol{\nabla}\cdot\boldsymbol{D} = \rho \qquad \text{(Gauss's law)}, \tag{6.50a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{6.50b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{6.50c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{6.50d}$$

where $\boldsymbol{D}$ is defined in Eq. (6.23) and $\boldsymbol{H}$ is defined in Eq. (9.11).[8] The Maxwell equations (6.50) in differential form can be cast in integral form using the divergence theorem and Stokes' theorem.

6.11.2 Integral Form of Maxwell Equations in Medium

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 (2.20)

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

[8] The vacuum Maxwell equations (1.1) can be recovered from the Maxwell equations in medium (6.50) by substituting $\boldsymbol{D} = \varepsilon_0\boldsymbol{E}$ and $\boldsymbol{H} = \boldsymbol{B}/\mu_0$ into Eqs. (6.50), remembering from Eq. (2.4) that $\mu_0\varepsilon_0 = 1/c^2$.

applied to Eq. (6.50a) yields

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\, d^3x, \tag{6.51}$$

with Eq. (6.51) 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. (6.50c) yields the integral equation

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0, \tag{6.52}$$

where Eq. (6.52) is the magnetic analog of Eq. (6.51), requiring that there be no net flux B through the closed surface S because no magnetic charges are known to exist.

In a similar manner, suppose that C is a closed contour spanned by an open surface S', $d\boldsymbol{l}$ 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 (2.25),

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

to Eq. (6.50b) gives

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}'\,da, \tag{6.53}$$

which is an integral form of Faraday's law of magnetic induction. Likewise, applying Stokes' theorem to Eq. (6.50d) gives

$$\oint_C \boldsymbol{H}\cdot d\boldsymbol{l} = \oint_{S'} \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}\right)\cdot\boldsymbol{n}'\,da \tag{6.54}$$

which is an integral form of the Ampère–Maxwell law of magnetic fields. Summarizing equations (6.51)-(6.54), the integral forms of Maxwell's equations in medium are

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\, d^3x \qquad \text{(Gauss's law)} \tag{6.55a}$$

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}'\,da, \qquad \text{(Faraday's law)} \tag{6.55b}$$

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0, \qquad \text{(No magnetic charges)} \tag{6.55c}$$

$$\oint_C \boldsymbol{H}\cdot d\boldsymbol{l} = \oint_{S'} \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}\right)\cdot\boldsymbol{n}'\,da \qquad \text{(Ampère–Maxwell law)} \tag{6.55d}$$

Equations (6.55) 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 shall now illustrate following the discussion in Jackson [19].

6.11.3 Matching Conditions at Boundaries

Consider Fig. 6.10, where there is a boundary between Medium 1 and Medium 2 and we allow the most general possibility that the interface at the boundary can have both a surface charge σ and a surface current density $\boldsymbol{K}$. To facilitate the analysis, an infinitesimal

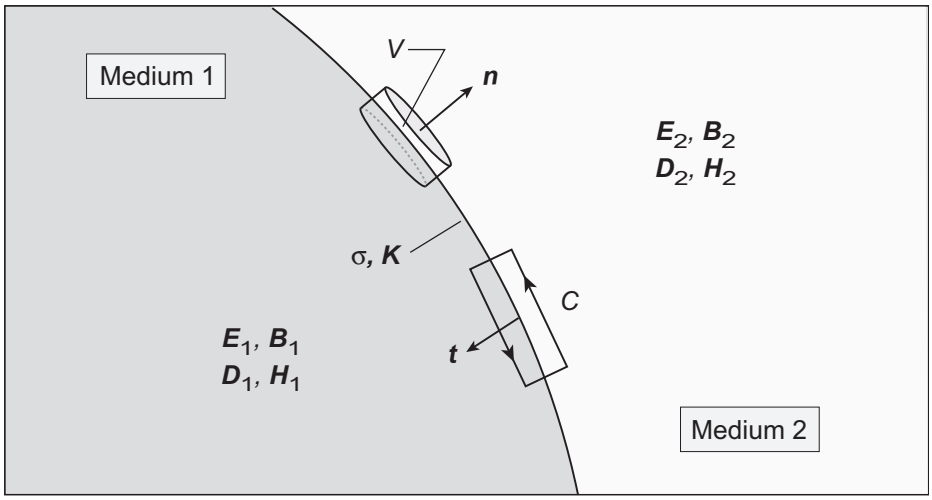


Fig. 6.10 Schematic illustration of a boundary surface between two different media that is assumed to carry surface charge σ and surface current density $\boldsymbol{K}$. The infinitesimal Gaussian cylinder of volume V is assumed to lie half in one medium and half in the other, with the normal to its surface $\boldsymbol{n}$ pointing from Medium 1 into Medium 2. The infinitesimal rectangular contour C is assumed to lie 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 in the figure) is tangent to the surface. Adapted from Ref. [19].

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 (6.55a) and (6.55c) to the cylindrical Gaussian pillbox in Fig. 6.10. In the limit that the pillbox is assumed to be very shallow the side of the cylinder does not contribute to the integrals on the left side of Eqs. (6.55a) and (6.55c). If the top and bottom of the cylinder are are assumed to be parallel to the interface and to each have area Δa, then the integral on the left side of Eq. (6.55a) is given by

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = (\boldsymbol{D}_2 - \boldsymbol{D}_1)\cdot\boldsymbol{n}\,\Delta a.$$

Now, applying a similar argument to the left side of Eq. (6.55c), the integral is given by

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = (\boldsymbol{B}_2 - \boldsymbol{B}_1)\cdot\boldsymbol{n}\,\Delta a.$$

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. (6.55a) 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 to each

other by

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n} = \sigma, \tag{6.56a}$$

$$(\boldsymbol{B}_2 - \boldsymbol{B}_1) \cdot \boldsymbol{n} = 0. \tag{6.56b}$$

Thus, stating Eqs. (6.56) in words:

1. the normal component of the magnetic field $\boldsymbol{B}$ is continuous across the interface, but
2. the discontinuity of the normal component of the electric displacement $\boldsymbol{D}$ at any point on the interface is equal to the surface charge density σ at that point.

In a similar manner, the infinitesimal rectangular contour C in Fig. 6.10 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. (6.55b) 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. (6.55d) 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. (6.55b) 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. (6.55d) 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 that given above for the right side of Eq. (6.55b). 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

$$\boldsymbol{n} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1) = 0, \tag{6.57a}$$

$$\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{K}, \tag{6.57b}$$

where in Eq. (6.57b) it is understood that the surface current only has components parallel to the interface at each point. Thus, stating Eqs. (6.57) in words:

1. the tangential component of the electric field $\boldsymbol{E}$ is continuous across an interface, but
2. the tangential component of $\boldsymbol{H}$ is discontinuous across the interface by an amount having magnitude $|\boldsymbol{K}|$ and direction $\boldsymbol{K} \times \boldsymbol{n}$.

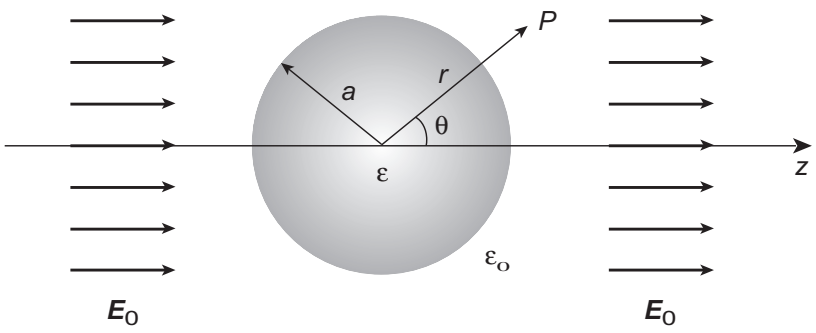


Fig. 6.11 A dielectric ball of radius a and dielectric constant $\kappa = \varepsilon/\varepsilon_0$ placed in an initially uniform external electric field $\boldsymbol{E}_0$ directed along the $\boldsymbol{z}$ axis. Used in the example discussed in Section 6.12.

The discontinuity equations (6.56) and (6.57) 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 evaluated over all of the space.

6.12 Example: A Dielectric Boundary Value Problem

Methods developed in previous chapters may be adapted to handle the presence of dielectrics in boundary value problems. Let us illustrate, in the process displaying techniques that are applicable to a variety of problems.

6.12.1 Dielectric Ball in External Electric Field

A 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. 6.11, with the assumption that there are no free charges inside or outside of the ball. Let's determine the scalar potential Φ and the electric field $\boldsymbol{E}$ using the Laplace equation with appropriate boundary conditions. Exploiting axial symmetry about the z axis and solving the Laplace equation in spherical coordinates by separation of variables (see Section 4.1.2), general solutions are of the form given by Eq. (4.37),

$$\Phi(r,\theta) = \sum_{l=0}^{\infty} (A_l r^l + B_l r^{-(l+1)}) P_l(\cos\theta).$$

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

$$\Phi_{\text{in}}(r,\theta) = \sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta), \tag{6.58}$$

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. (4.41)],

$$\Phi(r\to\infty) = -E_0 z = -E_0 r\cos\theta = -E_0 r P_1(\cos\theta).$$

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 $r\to\infty$. Multiplying both sides by $P_1(\cos\theta)$ and integrating,

$$-E_0 r\int P_1(\cos\theta)P_1(\cos\theta)d(\cos\theta) = B_l r^l \int P_1(\cos\theta)P_l(\cos\theta)d(\cos\theta).$$

Using the orthogonality relation (4.38),

$$\int_{-1}^{+1} P_k(\cos\theta)P_l(\cos\theta)d(\cos\theta) = \frac{2}{2l+1}\,\delta_{kl},$$

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,

$$\Phi_{\text{out}}(r,\theta) = -E_0 r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)} P_l(\cos\theta). \tag{6.59}$$

Now let's use the boundary conditions at the edge of the sphere to fix the other constants, by requiring the matching conditions at the surface $r=a$ for tangential and normal components of Φ,

$$-\frac{1}{a}\frac{\partial\Phi_{\text{in}}}{\partial\theta}\bigg|_{r=a} = -\frac{1}{a}\frac{\partial\Phi_{\text{out}}}{\partial\theta}\bigg|_{r=a} \qquad (\text{Tangential}), \tag{6.60}$$

$$-\varepsilon\frac{\partial\Phi_{\text{in}}}{\partial r}\bigg|_{r=a} = -\varepsilon_0\frac{\partial\Phi_{\text{out}}}{\partial r}\bigg|_{r=a} \qquad (\text{Normal}). \tag{6.61}$$

We must substitute the expansions (6.58) and (6.59) into these matching equations and solve to determine the constants. First consider the tangential matching. Substituting (6.58) and (6.59) into Eq. (6.60) 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

$$\sum_{l=0}^{\infty} a^l A_l \frac{\partial}{\partial\theta}P_l(\cos\theta) = -aE_0\frac{\partial}{\partial\theta}P_1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)}C_l\frac{\partial}{\partial\theta}P_l(\cos\theta). \tag{6.62}$$

Let's convert the derivatives of Legendre polynomials $P_n(x)$ to associated Legendre polynomials $P_n^m(x)$ using the general relationship [2],

$$P_n^m(x) = (1-x^2)^{m/2}\frac{d^m}{dx^m}P_n(x), \tag{6.63}$$

which, upon specializing to $m = 1$, yields

$$\frac{dP_n(x)}{dx} = (1-x^2)^{-1/2} P_n^1(x), \tag{6.64}$$

or, upon letting $x = \cos\theta$,

$$\frac{dP_n(\cos\theta)}{d\theta} = -P_n^1(\cos\theta). \tag{6.65}$$

Then Eq. (6.62) becomes

$$-\sum_{l=0}^{\infty} a^l A_l P_l^1(\cos\theta) = -aE_0 P_1^1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)} C_l P_l^1(\cos\theta). \tag{6.66}$$

We will now exploit the orthogonality properties of the associated Legendre polynomials, which obey the relationship [2],

$$\int_{-1}^{+1} P_p^m(x) P_q^m(x) dx = K_{qm}\delta_{pq} \qquad K_{qm} \equiv \frac{2}{2q+1}\frac{(q+m)!}{(q-m)!}, \tag{6.67}$$

or in spherical coordinates

$$\int_0^{\pi} P_p^m(\cos\theta) P_q^m(\cos\theta)\sin\theta d\theta = K_{qm}\,\delta_{pq}. \tag{6.68}$$

There are two solutions for Eq. (6.66),

1. a *special solution* for $l = 1$, and
2. a *general solution* for all $l \neq 1$.

Special solution $(l = 1)$*:* To obtain the special solution let $x = \cos\theta$, multiply both sides by $P_1^1(x)$, and integrate to give

$$-\sum_{l=0}^{\infty} a^l A_l \int P_1^1(x) P_l^1(x) dx = -aE_0 \int P_1^1(x) P_1^1(x) dx + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \int P_l^1(x) P_1^1(x) dx.$$

Utilizing Eq. (6.67), 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 $aA_1 = -aE_0 + a^{-2}C_1$, so that

$$A_1 = -E_0 + \frac{C_1}{a^3} \qquad \text{(Special solution for } l = 1\text{)}. \tag{6.69}$$

General solution $(l \neq 1)$*:* To obtain the general solution, set $x = \cos\theta$, multiply both sides of Eq. (6.66) by $P_k^1(x)$, and integrate to give

$$-\sum_{l=0}^{\infty} a^l A_l \int P_k^1(x) P_l^1(x) dx = -aE_0 \int P_k^1(x) P_1^1(x) dx + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \int P_k^1(x) P_l^1(x) dx.$$

Utilizing the orthogonality relation (6.67), this gives $a^l A_l = a^{-(l+1)} C_l$, which rearranges to

$$A_l = \frac{C_l}{a^{2l+1}} \quad \text{(General solution for } l \neq 1\text{)}. \tag{6.70}$$

Two relations [Eqs. (6.69) and (6.70)] have been obtained among the coefficients using the *tangential matching condition* (6.60) but the contraints imposed by matching at the boundary have not been exhausted: we may use the *normal matching condition* (6.61) to obtain additional constraints on the coefficients. Upon substituting the expansions (6.58) and (6.59) into Eq. (6.61),

$$\varepsilon \sum_{l=0}^{\infty} l a^{l-1} A_l P_l(x) = \varepsilon_0 B_1 P_1(x) + \sum_{l=0}^{\infty} -(l+1) C_l a^{-(l+2)} P_l(x). \tag{6.71}$$

As before, there are two solutions,

1. a *special solution,* when $l = 1$, and
2. a *general solution,* when $l \neq 1$.

Special solution $(l = 1)$*:* The special solution may be found by multiplying Eq. (6.71) by $P_1(x)$ and integrating. Upon exploiting orthogonality properties the result is another relationship among coefficients:

$$\frac{\varepsilon}{\varepsilon_0} A_1 = -E_0 - 2\frac{C_1}{a^3} \quad \text{(Special solution for } l = 1\text{)}. \tag{6.72}$$

General solution $(l \neq 1)$*:* The general solution may be obtained by multiplying both sides of Eq. (6.71) by $P_k(x)$, integrating, and exploiting the orthogonality constraints (4.38) to give

$$\frac{\varepsilon}{\varepsilon_0} l A_l = -(l+1)\frac{C_l}{a^{2l+1}} \quad \text{(General solution for } l \neq 1\text{)}. \tag{6.73}$$

We have obtained four equations—(6.69), (6.70), (6.72), and (6.73)—by exploiting boundary matching conditions on the fields; these must be solved simultaneously for the unknown coefficients. We note that Eqs. (6.70) and (6.73) can be satisfied only if $A_l = C_l = 0$ for all $l \neq 1$, and solving the other two equations (6.69) and (6.72) simultaneously for $l = 1$ gives

$$A_1 = -\left(\frac{3}{\varepsilon/\varepsilon_0 + 2}\right) E_0 \qquad C_1 = \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) a^3 E_0. \tag{6.74}$$

Inserting these results into Eqs. (6.58) and (6.59) gives for the *interior and exterior scalar potentials*

$$\Phi_{\text{in}} = -\left(\frac{3}{2 + \varepsilon/\varepsilon_0}\right) E_0 r\cos\theta = -\left(\frac{3}{2 + \varepsilon/\varepsilon_0}\right) E_0 z \qquad \text{(Interior potential)}, \tag{6.75a}$$

$$\Phi_{\text{out}} = -E_0 z + \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) E_0 a^3 \frac{\cos\theta}{r^2} \qquad \text{(Exterior potential)}, \tag{6.75b}$$

The electric fields then follow from $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$,

$$E_{\text{in}} = \left(\frac{3}{\varepsilon/\varepsilon_0 + 2}\right) E_0 \qquad \text{(Interior electric field)}, \tag{6.76a}$$

$$E_{\text{out}} = E_0 - \frac{P}{4\pi\varepsilon_0 r^3} \qquad \text{(Exterior electric field)}, \tag{6.76b}$$

as shown in Problem 6.5. From these results we conclude that for the interior electric field

1. E_{in} is a constant that is proportional to the constant exterior field E_0,
2. if $\varepsilon = \varepsilon_0$, the interior field is equal to the exterior field, $E_{\text{in}} = E_0$, and
3. if $\varepsilon > \varepsilon_0$, then the strength of the interior field is reduced relative to that of the exterior field, $E_{\text{in}} < E_0$.

In Eq. (6.75b) for the external field, it is clear that the first term is just 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. (3.53)] of the polarized sphere is taken to have magnitude

$$p = 4\pi\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) a^3 E_0, \tag{6.77}$$

then the exterior potential (6.75b) can be written

$$\Phi_{\text{out}} = -E_0 z + \frac{p}{4\pi\varepsilon_0}\frac{\cos\theta}{r^2}, \tag{6.78}$$

making clear that the external potential is the unperturbed external potential (first term) modified by a potential corresponding to an induced electric dipole at the origin that is associated with the polarized dielectric sphere. From Eqs. (6.25) and (6.27), the polarization $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E}, \tag{6.79}$$

and from Eq. (6.11) the polarization may also be defined as the density of dipole moments. Then from Eq. (6.77), the polarization of the dielectric sphere is given by

$$\begin{aligned}\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E} &= \left(\frac{\text{dipole moments}}{\text{unit volume}}\right) = \frac{p}{\frac{4}{3}\pi a^3} \\ &= 3\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) \boldsymbol{E}_0.\end{aligned} \tag{6.80}$$

As you are asked to show in Problem 6.6, the surface charge on the polarized dielectric in Fig. 6.11 is then given by

$$\sigma_{\text{pol}} = 3\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) E_0 \cos\theta. \tag{6.81}$$

The results of this extended example of a spherical dielectric in an external electric field are displayed graphically in Figs. 6.12 and 6.13.

6.12.2 Interpretation of Results

As indicated schematically in Fig. 6.12(a), the external field E_0 aligned in the z direction polarizes the dielectric sphere, with the polarization vector $\boldsymbol{P}$ given in Eq. (6.80) and oriented in the direction of the external field. The polarization is the density of induced dipole moments (6.77). This corresponds to the polarization of charge illustrated in Fig. 6.12(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. 6.12(b) induces an electric field

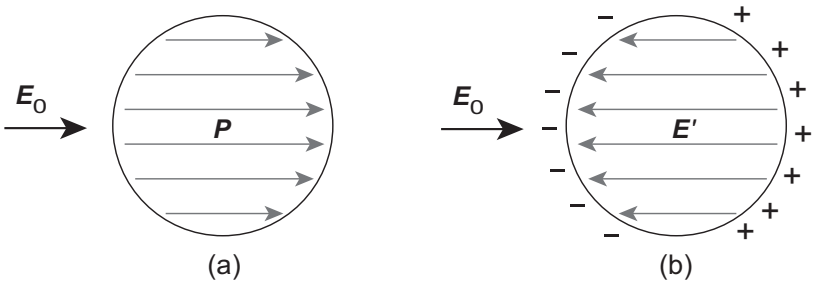


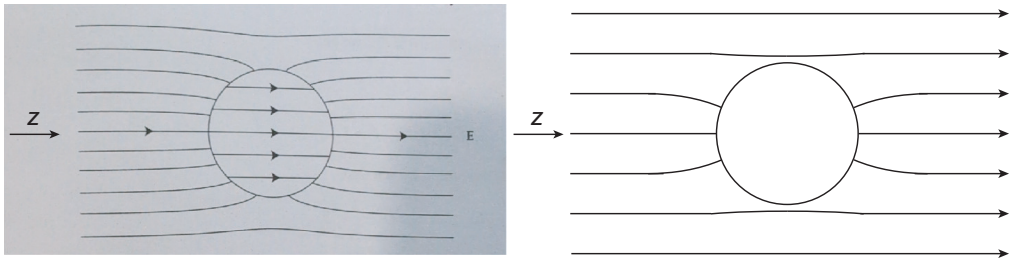
Fig. 6.12

(a) Polarization of a dielectric ball by an external electric field $\boldsymbol{E}_0$. (b) The polarization of charge induces a field $\boldsymbol{E}'$ inside the sphere that opposes and partially cancels the applied field $\boldsymbol{E}_0$.

$\boldsymbol{E}'$ that opposes the applied field $\boldsymbol{E}_0$ but does not completely cancel it for the dielectric-filled sphere as would be the case for a conducting metal-filled sphere. The electric field $\boldsymbol{E}_{\text{in}}$ inside the sphere is given by Eq. (6.76) 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. 6.13(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 $\boldsymbol{E}'$ 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 with density given by Eq. (6.81) that has accumulated there as a consequence of polarization by the external field.

As indicated in Eqs. (6.77) and (6.78), in the presence of the dielectric the external potential has two contributions:



(a) Dielectric filled sphere (a) Metal filled sphere

Fig. 6.13

(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. 4.5.

1. the original potential due to the external field, and
2. a correction term that may be interpreted as the potential generated by an effective electric dipole centered on the dielectric sphere, which has been produced by the charge polarization.

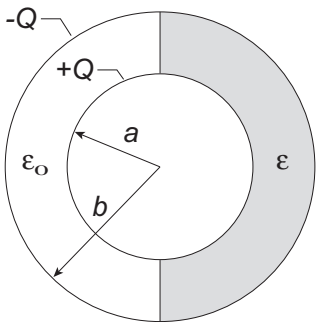
At large distance the external potential is that of the original applied field $\boldsymbol{E}_0$, but near the sphere the field lines in Fig. 6.13(a) are strongly distorted by the increasing importance of the second term in Eq. (6.78), which grows rapidly as the sphere is approached since it scales as r^{-2}.

Background and Further Reading

Good introductions to the material of this chapter may be found in Griffiths [13]. More advanced treatments may be found in Jackson [19], Garg [11], Chaichian et al [5], and Zangwill [42].

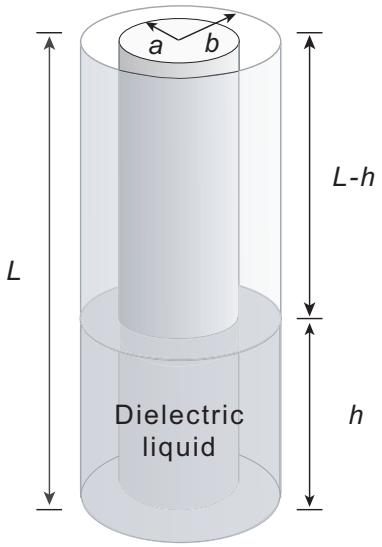
Problems

6.1 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 $r = a$. ***

6.2 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.

6.3 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.

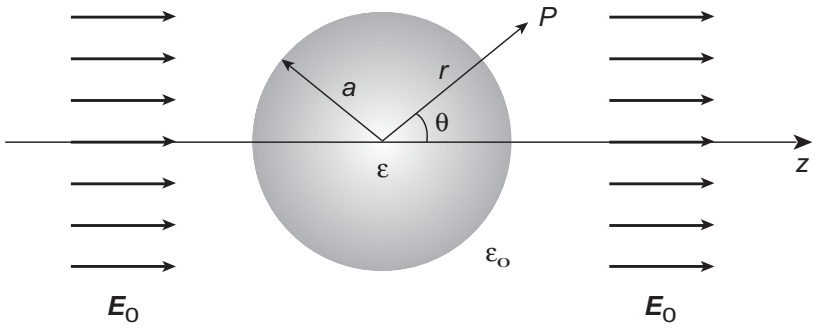
(a) With the charging battery *disconnected* a dielectric sheet that has exactly the same width and length as a plate is inserted between the plates. Find the work done on the dielectric sheet; is it pulled in or must it be pushed in?

(b) Repeat the experiment and analysis of part (a), but with the charging battery *connected* to the plates.

6.4 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.

6.5 Prove that for the dielectric ball in an external electric field considered in Section 6.12, the interior and exterior electric fields are given by Eqs. (6.76a) and (6.76b), respectively. ***

6.6 For the dielectric ball in an external electric field



that is analyzed in Section 6.12, show that the surface charge density induced by the polarization is given by,

$$\sigma_{\rm pol} = 3\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2} \right) E_0 \cos\theta,$$

where ε characterizes the interior and ε_0 characterizes the exterior permittivity.

7 Steady Currents

An important class of problems in electrostatics corresponds to situations where the source of the electric field $\boldsymbol{E}$ is a steady flow of electrical charge. In this chapter we shall study electrostatics in the presence of such steady-current sources.

7.1 Steady-Current Conditions

The current I passing through an arbitrary surface S is given by

$$I = \int_S d\boldsymbol{S} \cdot \boldsymbol{J}. \tag{7.1}$$

where $\boldsymbol{J}$ is the current density. Generally the charge density ρ and the current density $\boldsymbol{J}$ are required to satisfy the continuity equation (1.3),

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

which ensures *local conservation of charge* by demanding that electrical charge variation in some arbitrarily small volume be caused by flow of electrical current through the surface of that volume. If we restrict consideration to charge densities having no explicit time dependence, $\partial \rho / \partial t = 0$ and the continuity equation simplifies to the *steady-current condition*

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = 0 \qquad \text{(Steady-current condition)}. \tag{7.2}$$

Currents satisfying (7.2) can produce static electric and static magnetic fields, which can be studied independently because the Maxwell equations (1.1) do not mix static electric and magnetic fields. The field lines of a current density $\boldsymbol{J}$ consistent with Eq. (7.2) must satisfy the condition that every field line

- begins and ends at infinity, or
- closes on itself, implying that
- any field line entering an infinitesimal volume of space must also exit that volume.

Figure 7.1 shows an example of field lines for a current satisfying these conditions. The resemblance of Fig. 7.1 to streamlines for fluid flow is obvious, and the analogy of a steady current as a flow of fluid is often a useful one as long as the details of microscopic interactions between particles in an ordinary fluid can be ignored. If the motion of a charge density is characterized by a velocity field $\boldsymbol{v}(\boldsymbol{x})$ we may write

$$\boldsymbol{J}(\boldsymbol{x}) = \rho(\boldsymbol{x}) \boldsymbol{v}(\boldsymbol{x}). \tag{7.3}$$

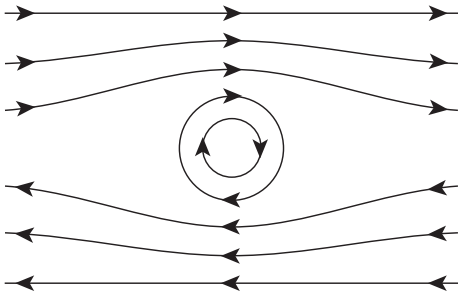


Fig. 7.1 Field lines of a current density that satisfies the condition (7.2) for a steady current. The resemblance to streamlines for flow of an ordinary fluid is apparent.

If the charge density can be viewed as moving rigidly with a uniform velocity $\boldsymbol{v} = v_0$, the current can be expressed as

$$\boldsymbol{J}(\boldsymbol{x}) = \rho(\boldsymbol{x})v_0 \tag{7.4}$$

and Eq. (7.3) is termed the *convection current density*. If N different species carrying charges q_i, with number densities n_i and uniform velocities v_i contribute to the current, the current density is a sum of the different contributions

$$\boldsymbol{J} = \sum_{i=1}^{N} q_i n_i b_i. \tag{7.5}$$

We shall most often exemplify using currents with a single kind of charge carrier, but currents with such mixed charge carriers—perhaps with different charge signs for the constituent carriers—may be found in some realistic applications such as in biological and astrophysical settings.

7.2 Vacuum Currents

We begin with the simplest currents, which can be viewed as a stream of particles flowing in a vacuum, with no influence from any fields or surrounding matter. The properties of vacuum currents may be explored by considering the operation of the *vacuum diode* illustrated in Fig. 7.2(a), which is a parallel plate capacitor with one plate heated so that it expels electrons by *thermionic emission*.[1] These devices may be unfamiliar to the reader as they have been replaced by semiconductors in most modern consumer electronics, but a theoretical analysis of the vacuum diode provides substantial insight into the nature of vacuum currents. In particular, vacuum diodes illustrate clearly a characteristic property of vacuum currents: the current is limited ultimately by interactions between the charge-carrying particles in the current itself.

[1] Thermionic emission is the ejection of particles from the surface of a heated metal because the thermal energy of the metal gives some particles sufficient kinetic energy to escape, with the rate of ejection increasing with temperature. We consider only the case of the emitted particles being electrons.

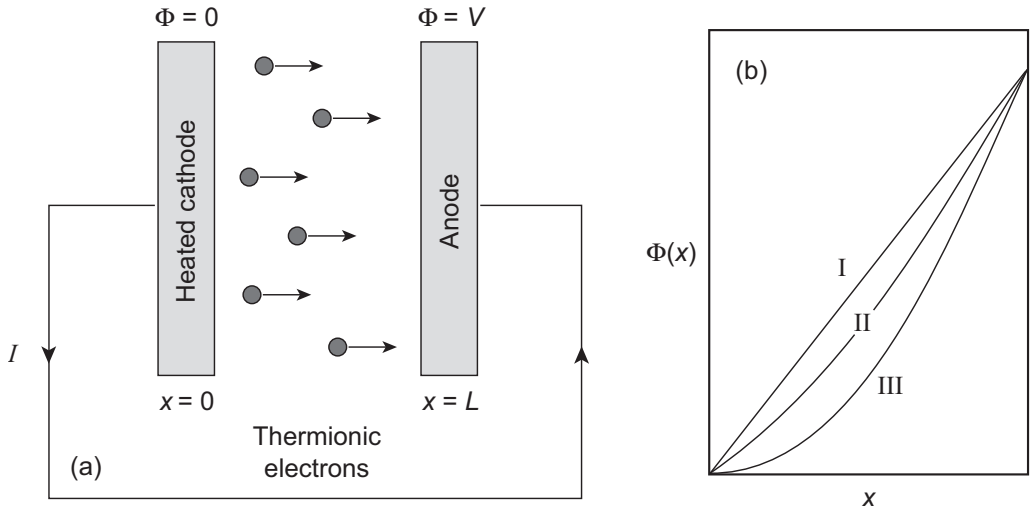


Fig. 7.2 (a) A vacuum diode. The heated cathode emits electrons thermionically that strike the anode, causing a current to flow in the circuit. (b) Curves for the electric potential Φ between the plates for (I) low, (II) medium, and (III) high thermionic electron density.

In our analysis a vacuum will be assumed to exist between the plates in Fig. 7.2(a), and for simplicity the thermionic electrons are taken to have negligible velocity when produced. A vacuum diode behaves differently at different cathode temperatures, which corresponds to different rates of electron emission from the heated plate, as will now be discussed.

7.2.1 Vacuum Diodes at Low Temperature

At low cathode temperatures the rate of electron emission, and thus the density of electrons between the plates, is small. The emitted electrons accelerate in the potential $\Phi(x) = xV/L$, and if the electron density is negligible the behavior of the electrons corresponds to the usual solution of the Laplace equation (2.58) between the plates of a capacitor that is illustrated by Curve I in Fig. 7.2(b). The current per electron in the gap is estimated in Problem 7.2 to be

$$I = \frac{e}{L} v, \tag{7.6}$$

where L is the plate separation and v is the average speed of electrons in the gap.

7.2.2 Vacuum Diodes at Intermediate Temperatures

If the temperature of the cathode is raised the rate of electron emission increases, as does the current in the gap since each additional electron makes a contribution given by Eq. (7.6). Eventually, the charge density in the gap created by the electrons can no longer be ignored; a *space charge*[2] builds up in the gap, and the solution for the potential is no longer

[2] *Space charge* is a general term used to indicate a volume of uncompensated charge that builds up in an otherwise electrically neutral region of space.

given by the Laplace equation $\nabla^2\Phi = 0$, but rather by the Poisson equation (2.54),

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon_0}. \tag{7.7}$$

The transition of the solution from Laplace to Poisson is a signal that the Coulomb repulsion between the emitted electrons in the gap associated with the charge density ρ (which is negative) can no longer be ignored. In one dimension $\nabla^2 = d^2/dx^2$ is a curvature and the Poisson equation (7.7) implies that the potential $\Phi(x)$ begins to develop a positive curvature between the plates as a result of the (negative) space charge built up there (physically because the Coulomb interaction between the charges in the gap begins to retard their acceleration), leading to the behavior exemplified by Curve II in Fig. 7.2(b).

7.2.3 Vacuum Diodes at High (Saturation) Temperature

Despite the retarding effect of the accumulating space charge the charge density and current continue to increase as the temperature of the cathode is raised, although more slowly than at lower temperatures. Eventually however, a temperature is reached where $d\Phi/dx \to 0$ at $x = 0$, which is exemplified by Curve 3 in Fig. 7.2(b). The *current saturates* at this point, since there is no electric field at the cathode ($x = 0$) to accelerate additional emitted electrons.

7.2.4 Space Charge and the Child–Langmuir Law

The saturated saturated steady current illustrated by Curve III in Fig. 7.2(b) still satisfies Eq. (7.2), meaning that

$$\rho(x)v(x) = \text{ constant} \tag{7.8}$$

everywhere. Furthermore, because we assumed $v(0) = 0$, the kinetic energy of each electron is

$$\frac{1}{2}mv^2(x) = e\Phi(x). \tag{7.9}$$

Combining these results with the Poisson equation (7.7) leads to the *Child–Langmuir law* specifying the space-charge limited current density,

$$|\boldsymbol{J}| = \frac{4}{9}\sqrt{\frac{2e}{m}}\frac{\varepsilon_0 V^{3/2}}{L^2}, \tag{7.10}$$

as you are asked to show in Problem 7.3.

7.3 Currents in Matter

Electrical currents flowing in neutral matter are generally more complex than those flowing in a vacuum, requiring the full power of quantum mechanics and statistical mechanics for an adequate theoretical understanding. Nevertheless, it is empirically well established that

the current density for many systems can be described well by *Ohm's law*, which is based on classical considerations and is described in the following subsection.

7.3.1 Ohm's Law

To make a current flow some force must push on the charges (continuously if, as is usually the case, there is any resistance to the charge flow). For most materials the current density $\boldsymbol{J}$ is proportional to the force per unit charge $\boldsymbol{f}$,

$$\boldsymbol{J} = \sigma \boldsymbol{f}, \tag{7.11}$$

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,

$$\boldsymbol{J} = \sigma(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}), \tag{7.12}$$

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 relativistic plasmas, the velocity is small and the $\boldsymbol{v} \times \boldsymbol{B}$ term can be neglected, leaving

$$\boldsymbol{J} = \sigma \boldsymbol{E} \qquad \text{(Ohm's law)}. \tag{7.13}$$

This is called *Ohm's law*.[3] 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 given

$$I \equiv JA\sigma EA = \frac{\sigma A}{V}. \tag{7.14}$$

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,

$$V = IR \qquad \text{(Ohm's law)}, \tag{7.15}$$

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. In more complex situations (for example, crossed electric and magnetic fields) the simple scalar form of Ohm's law described here is replaced by a tensor formulation, as described in Box 7.1.

7.3.2 The Drude Model of Metallic Conduction

An unavoidable difficulty for any classical treatment of electromagnetism is that important phenomena involving interaction with matter, such as the dielectric or magnetic properties of various materials, depend essentially on the microscopic structure of the matter, which

[3] Ohm's law is not a true "law" in the same sense as say Coulomb's law or the law of gravity. It is a "rule of thumb" that is often but not always obeyed. The subset of conductors that do obey Ohm's law are called *ohmic conductors*.

Box 7.1 Resistivity and Conductivity Tensors

In the simplest cases the current and the electric field may be assumed collinear and the elementary form of *Ohm's law* described in Section 7.3.1 holds: $V = IR$, where V is voltage, I is current, and R is resistance. This can also be expressed in terms of the electric field E and current density j as $j = \sigma E$ where σ is the scalar *conductivity,* or the inverted expression $E = \rho j$, where the scalar *resistivity* ρ is given by $\rho = \sigma^{-1}$. In more complex situations such as in the Hall effect described in Box 8.1 where there are both electric and magnetic fields, the resistivities and related quantities become tensors. Restricting to two dimensions for the Hall problem, the rank-2 resistivity tensor ρ and rank-2 conductivity tensor σ may be expressed as the matrices

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix} = \begin{pmatrix} \rho_{xx} & \rho_{xy} \\ \rho_{yx} & \rho_{yy} \end{pmatrix} \qquad \sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix},$$

where the components are $\rho_{ij} = E_i/j_j$ and $\sigma_{ij} = j_i/E_j$, with ρ and σ related by matrix inversion, $\rho = \sigma^{-1}$. Then $E_i = \rho_{ij} j_j$ and $j_i = \sigma_{ij} E_j$, or

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} \rho_{xx} & \rho_{xy} \\ \rho_{yx} & \rho_{yy} \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix} \qquad \begin{pmatrix} j_x \\ j_y \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix},$$

written out explicitly. Notice that if $\rho_{xy} = \rho_{yx}$ vanishes the resistivity matrix becomes diagonal and we revert to the simple Ohm's law description of Section 7.3.1, with a scalar resistivity.

requires arguments from quantum theory and quantum statistical mechanics that are outside the scope of classical electrodynamics and classical mechanics. In a presentation of classical electrodynamics, one is forced then into one of two alternatives:

1. graft sufficient quantum theory and results into the presentation, assuming that the reader has sufficient understanding of quantum theory, or
2. approximate quantum phenomena by classical or semiclassical models.

An example of the second alternative is the *Drude model of electrical conduction* [8, 9], which predates modern quantum mechanics by 25 years, but is still widely used in condensed matter physics to give a qualitative overview of metallic conduction [3]. We begin this section with a brief introduction to the Drude model.

Although at the time there was little precise theoretical justification, in the year 1900 Drude proposed a model of metallic conduction that *assumed* the heavy, positively charged ions (atomic nucleus plus core electrons in modern language) to be fixed spatially in a crystal lattice, and that the light valence electrons became detached from the ions and wandered freely, scattering randomly from the fixed ions as in a *miniature pinball machine;*

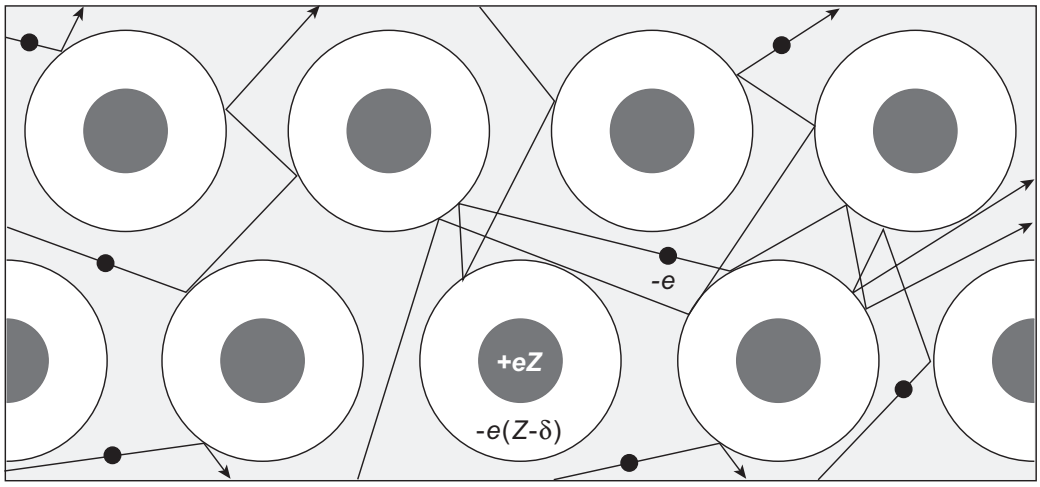


Fig. 7.3

Drude model of electrical conduction in metal with atoms of atomic number Z, which is basically a classical analysis of a pinball machine in which the ions are stationary and act as the posts of the machine and the valence electrons are detached from the ions, free to roam through the crystal lattice and scatter from the ions as in a pinball machine. In this diagram (which is *not to scale*) each large circle is an ion, which consists of the dark gray nucleus with charge $+eZ$, surrounded by a white region containing the core electrons carrying charge $-e(Z-\delta)$, where δ is the number of electrons ionized from each ion. The small black balls indicate electrons that have been freed from the ions and scatter randomly from the fixed ions. In this example an electric field has been assumed that causes a net drift of the randomly scattering electrons to the right.

Fig. 7.3 illustrates.[4] These free (valence) electrons are called the *conduction electrons*, and in the Drude model they are viewed as forming a gas. There are four basic assumptions that enter into this *Drude model*.

1. Between collisions with ions the conduction electrons move freely in straight lines, if there are no applied fields. If external fields are applied, the electrons move according to Newton's laws of motion in those fields (this is called the *independent electron approximation*), neglecting fields produced by other electrons or ions (this is called the *free electron approximation*). The independent electron approximation is found to be fairly good in many contexts, but even a qualitative understanding of the behavior of metals often requires improving on the free electron approximation.

[4] As indicated in Fig. 7.3, the number of ionized valence electrons per atom is denoted by δ, which will depend on the valence of the atoms comprising the metal. For example, metallic copper in a $+1$ valence state has $\delta(\mathrm{Cu}) = 1$ and metallic iron in a $+2$ valence state has $\delta(\mathrm{Fe}) = 2$. One should bear in mind that when the Drude model was formulated in 1900 the electron had only recently been discovered as a distinguishable particle (by J. J. Thomson in 1897), but it was not yet certain that atoms even existed, or what their structure was if they did. In particular, it is clear that our discussion above distinguishing core electrons from valence electrons in atoms is an interpolation into the Drude model of knowledge that would only come later: even the rudimentary Bohr model of the atom was still 15 years in the future, and it would be another 25 years before quantum mechanics would be invented when Drude published his model.

2. As in kinetic theory, collisions are assumed to alter electron velocity instantaneously.[5] Although the simple picture of Fig. 7.3 with instantaneous interactions misses the mark in details, many qualitative features of metallic behavior survive.
3. A parameter τ called the *relaxation time* or *mean collision time* is of fundamental importance in the Drude model. The probability per unit time for an electron collision to occur is τ^{-1} so that the probability that an electron experiences a collision in a time interval dt is dt/τ. With these assumptions, an electron chosen at random at a given time will travel an average time τ before the next collision, and will have traveled an average time τ since its last collision. In the simplest approximation, the relation time τ is assumed *independent of electron position and velocity*. Perhaps surprisingly, this assumption often works reasonably well.
4. Electrons reach thermal equilibrium only through collisions. After each collision an electron emerges with a velocity having a random direction and a magnitude proportional to the local temperature at the site of the collision.

The following example illustrates an application of the Drude model to the relationship of current, voltage, and resistance for a current flowing in a wire.

Example 7.1 According to the empirical Ohm's law described in Section 7.3.1, for a current I flowing in a wire, $V = IR$, where V is the voltage drop and R is the resistance, which depends on the geometry of the wire but is independent of V and I. Let us show that the Drude model provides a reasonable understanding of Ohm's law, and allows an estimate of the resistance R. The dependence of the resistance on the geometry of the wire can be eliminated by introducing the *resistivity* ρ, which depends only on the nature of the metal in the wire, through the requirement that it be the constant of proportionality between the electric field $\boldsymbol{E}$ and the current density $\boldsymbol{J}$ that it induces,

$$\boldsymbol{E} = \rho \boldsymbol{J}, \tag{7.16}$$

where we assume the simplest case that $\boldsymbol{E}$ and $\boldsymbol{J}$ are parallel. [If they were not parallel, the scalar resistivity in Eq. (7.16) would become a *resistivity tensor*, as described in Box 7.1.] The magnitude of $\boldsymbol{J}$ is the charge per unit time crossing a unit area perpendicular to the current. For uniform flow of a current I through a wire of length L and cross sectional area A, the current density will be $J = I/A$ and since the voltage drop along the wire is EL, from Eq. (7.16) $V = I\rho L/A$ and the resistance is related to the resistivity of the metal and the geometry of the wire by

$$R = \frac{\rho L}{A}. \tag{7.17}$$

If n electrons per unit volume, each carrying charge $-e$ move with velocity $\boldsymbol{v}$, in a time dt the charge passing through a cross section A will be $-nevAdt$, so the current density is

$$\boldsymbol{J} = -ne\boldsymbol{v}. \tag{7.18}$$

[5] As illustrated in Problem 7.1, the density of the "gas of conduction electrons" is typically of order 1000 times larger than that of classical gases at standard temperature and pressure; nevertheless, the Drude model assumes the normal kinetic theory of gases to apply in relatively unaltered form.

In the absence of an applied field electrons may be expected to move randomly in all directions and the velocity $\boldsymbol{v}$ in Eq. (7.18) averages to zero, but if an electric field $\boldsymbol{E}$ is applied there will be a non-zero net velocity of the electrons opposite the direction of $\boldsymbol{E}$ (since the electron charge is negative).

This velocity can be computed as follows. For a typical electron at time zero, let Δt be the time elapsed since its last collision. Its velocity will be a sum of the velocity $\boldsymbol{v}_0$ after the last collision plus a velocity $-e\boldsymbol{E}\delta t/m$ (where m is the electron mass) gained from the electric field. Since the Drude model assumes the direction of the velocity after the previous collision to be random, $\boldsymbol{v}_0$ averages to zero and the average electronic velocity is given entirely by the average of $-e\boldsymbol{E}\Delta t/m$. But the average of Δt is the Drude relaxation time τ and therefore the average velocity is

$$\overline{\boldsymbol{v}} = -\frac{e\boldsymbol{E}\tau}{m}, \tag{7.19}$$

and the current density is

$$\boldsymbol{J} = \sigma\boldsymbol{E} \qquad \sigma \equiv \frac{ne^2\tau}{m}, \tag{7.20}$$

where $\sigma = \rho^{-1}$ is called the *conductivity*. [As for the resistivity, if $\boldsymbol{E}$ and $\boldsymbol{J}$ are not parallel the scalar conductivity of Eq. (7.20) becomes a *conductivity tensor*, as described in Box 7.1.] The Drude relaxation time is then

$$\tau = \frac{m}{\rho ne^2}, \tag{7.21}$$

from which measured resistivities ρ can be used to estimate relaxation times. These are found to be $\tau \sim 10^{-14}$ seconds for many metals at room temperature.

The Drude model describes some conduction phenomena qualitatively, but fails dramatically for others. The successes point to aspects of conduction theory that can be understood in simple classical terms, such in Example 7.1; the failures can be understood and ameliorated only with the aid of modern quantum theories of electrical conduction.

Background and Further Reading

The properties of vacuum diodes are described in Ref. [42]. An introduction to the Drude model of classical conduction may be found in Ref. [3].

Problems

7.1 In the Drude model, estimate the number density of the gas of conduction electrons for a typical metal like copper. Compare with the number density for a typical gas by calculating the number density for air molecules under standard conditions.

7.2 For the vacuum diode of Fig. 7.2(a), assume that the cathode temperature is low enough that electron density between the plates is small. Find an expression for the average current per electron in the gap between the plates in terms of plate separation L and average electron speed v.

7.3 Using Eqs. (7.8) and (7.9), and the Poisson equation (7.7), prove the space-charge limiting result given in Eq. (7.10). ***

8 Magnetostatics in Vacuum

Chapters 2-6 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 produced by those charges at arbitrary locations in space. This can be accomplished 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.

8.1 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]

$$\boldsymbol{\nabla}\cdot\boldsymbol{B}(\boldsymbol{x}) = 0, \tag{8.1a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B}(\boldsymbol{x}) = \mu_0\boldsymbol{J}(\boldsymbol{x}). \tag{8.1b}$$

Applying the divergence operation to both sides of Eq. (8.1b) using the vector identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0$ of Eq. (A.6) gives

$$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{B}(\boldsymbol{x})) = 0 = \mu_0\boldsymbol{\nabla}\cdot\boldsymbol{J}(\boldsymbol{x}). \tag{8.2}$$

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 (8.1) specify the divergence and the curl of the vector field $\boldsymbol{B}(\boldsymbol{x})$. This should be sufficient to specify $\boldsymbol{B}(\boldsymbol{x})$ uniquely by the Helmholtz theorem of Box 3.1, if the fields and sources are sufficiently well-behaved.

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. (8.4) 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. 6), 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, occurs much more commonly than than the dielectric counterpart of *ferroelectricity* where a material spontaneously polarizes in the absence of an external electric field [42].

From the static (all time derivatives set to zero) vacuum Maxwell equations (1.1), the basic equations governing electrostatics are

$$\boldsymbol{\nabla} \times \boldsymbol{E}(\boldsymbol{x}) = 0 \tag{8.3a}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{E}(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x})}{\varepsilon_0}. \tag{8.3b}$$

Comparing with the basic equations (8.1) 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 in one inertial frame looks like a magnetic field in a different inertial frame. Thus the distinction between electric and magnetic phenomena isn't consistent with special relativity and Lorentz invariance.

[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. Intrinsic spin is inherently quantum in nature and any attempt to treat it classically raises issues at a fundamental level.

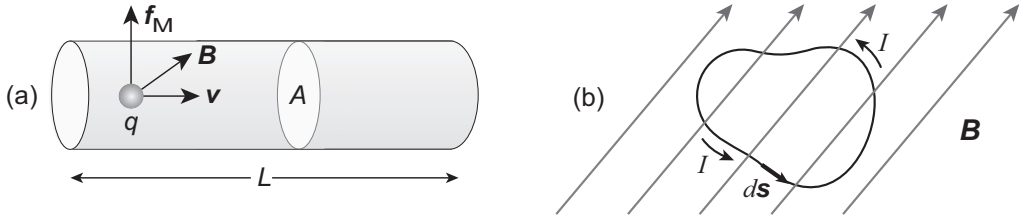


Fig. 8.1 (a) The magnetic force $\boldsymbol{f}_{\mathrm{M}}$ acting on a charged particle moving with velocity $\boldsymbol{v}$ in a segment of conductor of length L and cross-sectional area A, subject to a magnetic field $\boldsymbol{B}$. (b) Current loop carrying a steady current I in a uniform magnetic field $\boldsymbol{B}$. The vector line segment $d\boldsymbol{s}$ has a magnitude equal to the segment length and direction equal to the local current direction.

We shall take that up in later chapters addressing 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 of most everyday and laboratory experience,[3] there is utility in a formalism that distinguishes between electric and magnetic phenomena through the use of equations that are (secretly) Lorentz covariant, but are not manifestly so.

8.2 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*,

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \qquad \text{(Lorentz force law)}. \tag{8.4}$$

As you are asked to show in Problem 9.2, the magnetic force acting on a single charged particle in the short segment of conductor illustrated in Fig. 8.1(a) is

$$\boldsymbol{f}_{\mathrm{M}} = q\boldsymbol{v} \times \boldsymbol{B}, \tag{8.5}$$

and is directed upward by the right-hand rule for the cross product. The total magnetic force acting on the segment in Fig. 8.1(a) is given by

$$\boldsymbol{F}_{\mathrm{M}} = I(\boldsymbol{L} \times \boldsymbol{B}), \tag{8.6}$$

[3] Although it is not completely apparent from the appearance of the equations in SI units without examining the constants carefully [recall from Eq. (2.4) that the SI constants μ_0 and ε_0 are related to each other through the speed of light c], *magnetic fields are intrinsically much weaker than electric fields under typical laboratory conditions*. But this is largely because the characteristic speeds of charged particles are much less than the speed of light c in those circumstances. If the speed of charged particles approaches c, magnetic forces and electric forces become comparable in intrinsic strength. This is a consequence of special relativity, which requires not only space and time, but also electric and magnetic phenomena, to enter the theory on an equal footing, and whether a field is seen as electric or magnetic by an observer depends on the inertial frame of that observer.

where $\boldsymbol{L}$ has magnitude L and direction equivalent to that of the current, and the magnitude of the current is $I = nqvA$. A finite length of wire having uniform cross section can be partitioned into segments of length ds as in Fig. 8.1(b), and the total force acting on the wire is found to be

$$\boldsymbol{F}_{\mathrm{M}} = I \int_a^b (d\boldsymbol{s} \times \boldsymbol{B}), \tag{8.7}$$

where a and b are the endpoints. For the special case of a steady current, uniform magnetic field $\boldsymbol{B}$, and a closed loop ($a = b$), the force acting on the loop is

$$\boldsymbol{F}_{\mathrm{M}} = I \oint (d\boldsymbol{s} \times \boldsymbol{B}) = I \left(\oint d\boldsymbol{s} \right) \times \boldsymbol{B} = 0, \tag{8.8}$$

which vanishes because for a closed loop the set of length-element vectors $d\boldsymbol{s}$ forms a closed polygon so $\oint d\boldsymbol{s} = 0$.

For any closed current loop in a uniform magnetic field the total magnetic force acting on the loop is zero.

Although in a uniform field the net force on current loops vanishes, the *net torque* does not, as you are asked to show in Problem 9.3.

The quite different phenomenology of magnetism relative to electrostatics in everyday experience is partially due to characteristic charged-particle velocities being much less than the speed of light in typical observations (see Section 8.1), and partially due to the nature of the Lorentz 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 8.1 and Fig. 8.2. Two important physical consequences of the force law (8.4) are illustrated in Box 8.1, *cyclotron orbits* and the classical *Hall effect*.

How much work can be done by the Lorentz force? We calculated previously in Eq. (2.37) 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 (8.4) with $\boldsymbol{B} = 0$ the same result is obtained (which is reassuring!). On the other hand, let's set $\boldsymbol{E} = 0$ in Eq. (8.4) 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

$$dW = \boldsymbol{F} \cdot d\boldsymbol{L} = Q(v \times \boldsymbol{B}) \cdot \boldsymbol{v}\, dt = 0. \tag{8.9}$$

The magnetic field does no work, since the cross product $\boldsymbol{v} \times \boldsymbol{B}$ is perpendicular to the velocity $\boldsymbol{v}$, so that the scalar product $(v \times \boldsymbol{B}) \cdot \boldsymbol{v}$ *vanishes identically*.

As exemplified in the cyclotron motion discussed in Box 8.1, magnetic forces can *alter the direction* of charged-particle velocity but *not its magnitude*.

If a magnetic field appears to us to be doing work, Eq. (8.9) indicates that we have

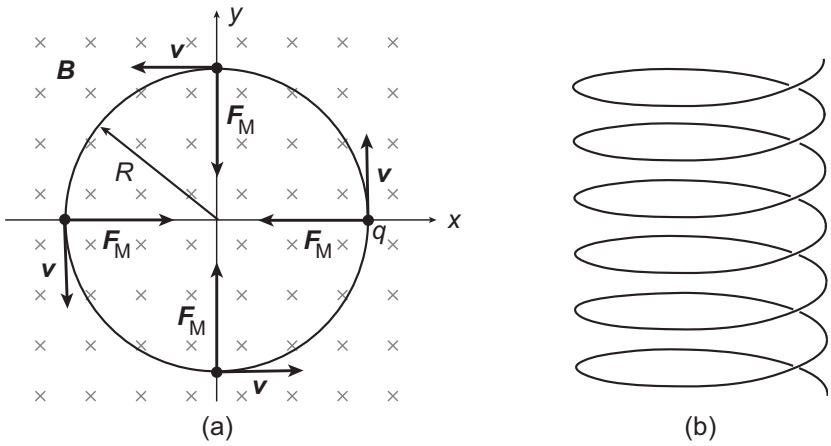


Fig. 8.2 (a) Cyclotron motion caused by magnetic field $\boldsymbol{B}$ pointing into the page along the z axis acting on a charge q with a velocity $\boldsymbol{v}$, as explained in Box 8.1. We have used the standard notation of $\otimes$ or $\times$ to indicate a magnetic field pointing into the page (and $\odot$ to indicate one pointing out of the page). (b) An electric field $\boldsymbol{E}$ perpendicular to the magnetic field converts circular cyclotron motion into spiral motion because of the Lorentz force law (8.4).

misunderstood what is happening. Often, when at first glance a magnetic field appears to be doing work, what is actually happening is that a magnetic field acts to alter the direction of motion, which then permits an electric field to do the observed work. In such a case the magnetic field enables the work to be done by changing the direction of motion, but the actual work is done by an electric field, never by the magnetic field.

8.3 The Law of Biot and Savart

The magnetic field $\boldsymbol{B}(\boldsymbol{x})$ of a steady line current described by the density $\boldsymbol{J}(\boldsymbol{x})$ is given by the empirical *Biot–Savart law*,[4]

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x' = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times\hat{\boldsymbol{R}}}{R^2}\,d^3x', \tag{8.10}$$

where the *permeability of free space* μ_0 is defined in Eq. (1.2) and the second form employs the compact notation of Eqs. (2.10) and (2.11). The integral in Eq. (8.10) can be performed analytically only for simple geometrical arrangements. 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 points to a line element $d\boldsymbol{L}$ of the wire as in Fig. 8.3, substitution of $Id\boldsymbol{L}$ for $\boldsymbol{J}d^3x$ in Eq. (8.10) yields

[4] The law embodied in Eq. (8.10) was formulated by Jean-Baptiste Biot and Félix Savart in 1820. The SI unit for $\boldsymbol{B}$ is the *tesla* (T), where $1\,\mathrm{T} \equiv 1\,\mathrm{N\,A^{-1}\,m^{-1}}$. Although discouraged by champions of international standards, it is common also to use the gaussian (CGI) unit of *gauss* (G) for $\boldsymbol{B}$, even when using SI units overall, where $1\,\text{tesla} = 10^4$ gauss. For reference, the magnetic field at the surface of the Earth is of order one gauss, and the strongest known magnetic fields are $\sim 10^{16}$ gauss, which are observed in highly magnetized, rapidly rotating neutron stars called *magnetars*.

Box 8.1 **Motion of Charged Particles in Magnetic and Electric Fields**

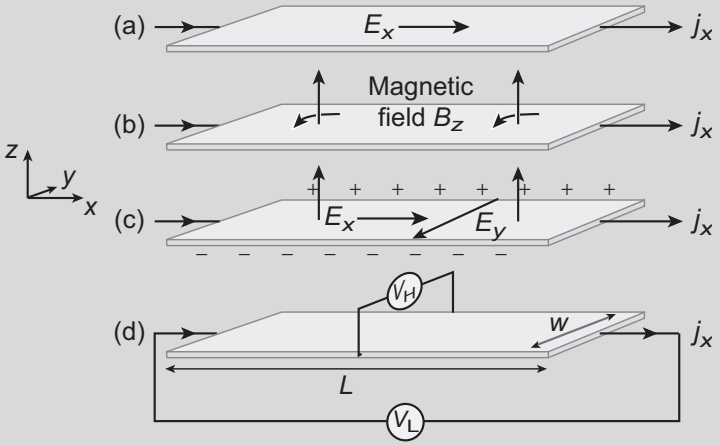
Characteristic *cyclotron motion* of a charged particle in a magnetic field is circular with the centripetal acceleration that curves the path provided by the magnetic force.

Cyclotron Motion

As illustrated in Fig. 8.2(a) and Problem 8.11, the magnetic field $\boldsymbol{B}$, oriented into the page, produces through the $q\boldsymbol{v}\times\boldsymbol{B}$ component of the Lorentz force (8.4) 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 both an electric field and magnetic field, the circular cyclotron motion due to the magnetic field can be converted into a the spiral motion depicted in Fig. 8.2(b) by the $q\boldsymbol{E}$ component of the Lorentz force.

Classical and Quantum Hall Effects

An important consequence of the Lorentz force is exemplified by the *classical Hall effect*, depicted in the following diagram for a 2D conductor.



(a) A current density j_x is produced by an applied electric field E_x. (b) A uniform magnetic field $\boldsymbol{B}$ applied in the $+z$ direction deflects electrons in the current in the $-y$ direction (Lorentz force). (c) Negative charge accumulate on one edge and positive charge on the opposite edge, producing a transverse electric field E_y (*Hall field*) and associated force that just cancels the Lorentz force; 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 Hall effect is a diagnostic for charge carriers in materials. At very high magnetic fields, the classical Hall effect is modified dramatically by quantum mechanics. Corresponding *quantum Hall experiments* revealed topological effects that initiated the topological matter revolution in modern condensed matter and materials science.

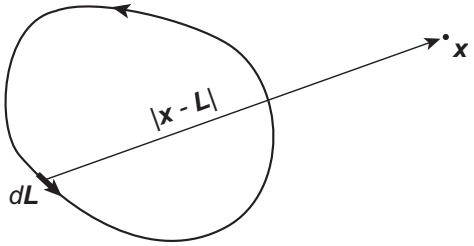


Fig. 8.3 Current loop carrying a steady current I for Eq. (8.11).

the Biot–Savart law in the form

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0 I}{4\pi}\int \frac{d\boldsymbol{L}\times(\boldsymbol{x}-\boldsymbol{L})}{|\boldsymbol{x}-\boldsymbol{L}|^3} = \frac{\mu_0 I}{4\pi}\int \frac{d\boldsymbol{L}\times\hat{\boldsymbol{R}}}{R^2}. \tag{8.11}$$

for a steady current.

Similar to the case for electric flux defined in Eq. (2.49), it is convenient to introduce the *magnetic flux* $\mathscr{F}_{\rm M}$ through a surface S, which is defined as the normal component of $\boldsymbol{B}$ integrated over S,

$$\mathscr{F}_{\rm M} = \int_S \boldsymbol{B}\cdot d\boldsymbol{a}. \tag{8.12}$$

In electrostatics one can use lines of force to describe an electric field. In a similar way, for magnetic fields, one can draw lines of the magnetic field, which are everywhere tangent to the direction of $\boldsymbol{B}$ at that point.

Since steady currents normally flow in closed loops, the integral in the Biot–Savart law is typically evaluated over a closed curve. However, the Biot–Savart law can also be applied to infinitely long wires. Let us use the Biot-Savart law to calculate the magnetic field due to current in a very long straight wire carrying a steady current I, as illustrated in Fig. 8.4. An element $Id\boldsymbol{L}$ of the current will produce a magnetic field $d\boldsymbol{B}$ with magnitude

$$dB = \frac{\mu_0 I}{4\pi}\frac{dL\sin\phi}{r^2},$$

as illustrated in Fig. 8.4. Expressing the variables dL, $\sin\phi$, and r^2 in terms of the angle θ,

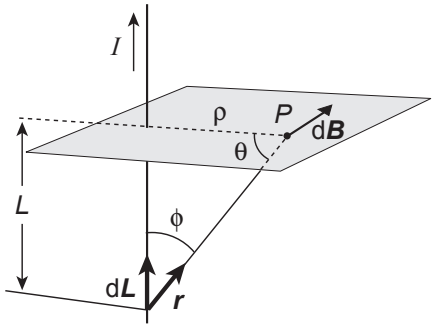


Fig. 8.4 Magnetic field $d\boldsymbol{B}$ produced by an element $Id\boldsymbol{L}$ of an infinitely long straight wire. The vector $d\boldsymbol{B}$ lies in a plane perpendicular to the wire at the point P.

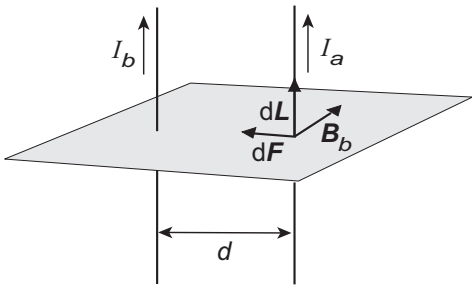


Fig. 8.5 Geometry for calculating the force per unit length acting between two long, straight parallel wires.

the magnitude of B is given by

$$B = \frac{\mu_0 I}{4\pi\rho}\int_{-\pi/2}^{+\pi/2}\cos\theta d\theta = \frac{\mu_0 I}{2\pi\rho}, \tag{8.13}$$

where we've assumed the length of the wire to be infinite.

This result can be generalized easily to calculate the force acting between the two long parallel wires illustrated in Fig. 8.5, as you are asked to show in Problem 8.7. The resulting force per unit length is

$$\frac{dF}{dI_a} = \frac{\mu_0 I_a I_b}{2\pi d}, \tag{8.14}$$

with the force attractive if the currents I_a and I_b flow in the same direction and repulsive if they flow in opposite directions.[5]

Example 8.1 Let us use the Biot–Savart law (8.11) to calculate the magnetic field produced along the z axis by the circular current loop in Fig. 8.6. 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

$$\boldsymbol{B}(z) = \hat{\boldsymbol{z}}\frac{\mu_0 I}{4\pi}\frac{\cos\theta}{R^2+z^2}\oint dL = \hat{\boldsymbol{z}}\frac{\mu_0 I}{2}\frac{R^2}{(R^2+z^2)^{3/2}}, \tag{8.15}$$

where $\oint dL = 2\pi R$ and $\cos\theta = R/\sqrt{R^2+z^2}$ were used. 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 the corresponding electrostatics problem of 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 constructively

[5] Until the year 2019, this equation provided the official definition of the ampere (amp) unit of current in the SI system: one ampere is the current passing through two long parallel wires 1 meter apart with thickness negligible relative to their separation that produces a magnetic force of 2×10^{-7} newtons per meter. In a 2019 revision of the SI units system, the elementary charge e was fixed to be exactly 1.602176634×10^{19} C, and an ampere (amp) was defined to be an electric current equivalent to one Coulomb (C) of charge moving past a reference point per second.

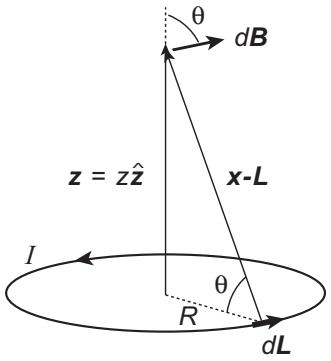


Fig. 8.6 Geometry for Biot–Savart calculation of the magnetic field on the symmetry axis of a circular current loop, as illustrated in Example 8.1.

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 electric fields produced by stationary charges and magnetic fields produced by charges in motion.

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 empirical starting point for magnetostatics, just as Coulomb's law may be viewed as the empirical starting point for electrostatics.

Both laws exhibit an inverse-square distance dependence, but otherwise they differ substantially because of the vector character of the magnetic law.

8.4 Differential Form of the Biot–Savart Law

The Biot–Savart law (8.10) for a line current 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 for line currents, 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 [19]. From Eq. (A.11),

$$\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3} = -\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = \boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right), \tag{8.16}$$

which allows converting the Biot–Savart equation (8.10) into

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{x}) &= \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x' \\ &= \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \end{aligned} \tag{8.17}$$

where $\boldsymbol{\nabla}$ has been pulled out of the integral because it operates on $\boldsymbol{x}$ but not $\boldsymbol{x}'$.[6] Now take the divergence of Eq. (8.17)

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\times\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'.$$

But from Eq. (A.6) we have the identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0$ so

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0, \tag{8.18}$$

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. (8.17),

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{8.19}$$

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 becomes

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{B} &= \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int \boldsymbol{J}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x' - \frac{\mu_0}{4\pi}\int \boldsymbol{J}(\boldsymbol{x})\,\nabla^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x' \\ &= -\frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int \boldsymbol{J}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x' + \mu_0\boldsymbol{J}(\boldsymbol{x}), \end{aligned} \tag{8.20}$$

where we have used the identities of Eq. (A.11),

$$\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) \qquad \nabla^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}') \tag{8.21}$$

(with $\boldsymbol{\nabla}$ operating on $\boldsymbol{x}$ and $\boldsymbol{\nabla}'$ operating on $\boldsymbol{x}'$) in the last step. Integrating the remaining integral in Eq. (8.20) by parts then gives

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J} + \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{8.22}$$

But for steady-state magnetism $\boldsymbol{\nabla}\cdot\boldsymbol{J} = 0$ and we obtain finally

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J}, \tag{8.23}$$

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)].

[6] Remember that in these manipulations, using our standard notation, 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. However, we will usually use the abbreviated notation $\boldsymbol{\nabla}\equiv\boldsymbol{\nabla}_x$ and $\boldsymbol{\nabla}'\equiv\boldsymbol{\nabla}_{x'}$.

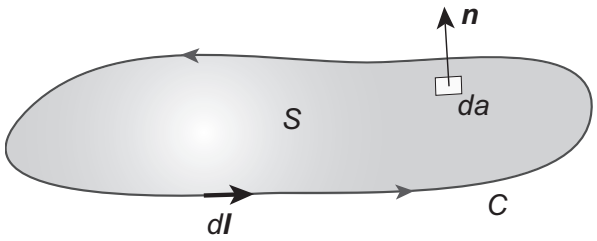


Fig. 8.7 The 2D surface S and the contour C bounding the surface for Stokes' theorem. An infinitesimal line element on C is indicated by dl and an infinitesimal surface element on S is indicated by da, with the normal to da denoted by $\boldsymbol{n}$. The path in the line integration is traversed in a right-hand rule sense relative to $\boldsymbol{n}$, as indicated by arrows.

The integral equivalent of Ampère's law (8.23) may be obtained from Stokes' theorem (2.25),

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

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 8.7 illustrates. Applying Stokes' theorem to Eq. (8.23),

$$\int_S (\boldsymbol{\nabla}\times\boldsymbol{B})\cdot\boldsymbol{n}\,da = \oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \mu_0\int_S \boldsymbol{J}\cdot\boldsymbol{n}\,da \tag{8.24}$$

and therefore Eq. (8.23) becomes

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \mu_0\int_S \boldsymbol{J}\cdot\boldsymbol{n}\,da. \tag{8.25}$$

Using that the total current I passing through the closed curve C is given by the surface integral on the right side of Eq. (8.25),

$$I = \int_S \boldsymbol{J}\cdot\boldsymbol{n}\,da, \tag{8.26}$$

gives us Ampère's law in integral form,

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \mu_0 I, \tag{8.27}$$

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 8.2 Let's determine the magnetic field for the long solenoid illustrated in Fig. 8.8, 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, each coil may be assumed to be perpendicular to the symmetry axis of the cylinder. One expects then on symmetry and general grounds that the magnetic field is oriented along the cylinder axis, and that it

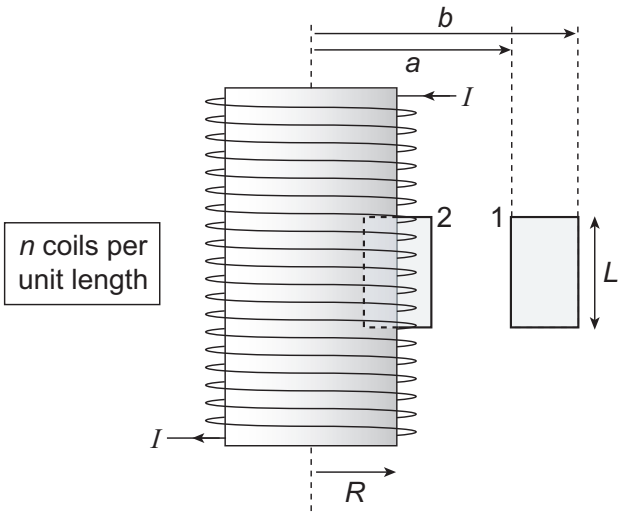


Fig. 8.8 Analysis of a long, tightly wound solenoid using Ampère's law. Two rectangular loops are shown, Loop 1 is outside the solenoid and Loop 2 overlaps half inside and half outside of the solenoid.

must be zero at large distances from the solenoid. Let's apply Ampère's law (8.27) to the two rectangular *Ampèrian loops* shown in the diagram.[7] 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 $B = |\boldsymbol{B}|$. Thus,

1. $B(a) = B(b)$ and the magnetic field outside is independent of the distance from the solenoid, but
2. the boundary conditions require that $B = 0$ at infinity,

which implies that $B = 0$ *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, since there is no contribution from the half rectangle outside the cylinder because $B = 0$ 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

[7] We shall refer to such constructions employed in the use of Ampère's law as *Ampèrian loops*, by analogy with the *Gaussian surfaces* introduced in Section 2.4 in the use of Gauss's law.

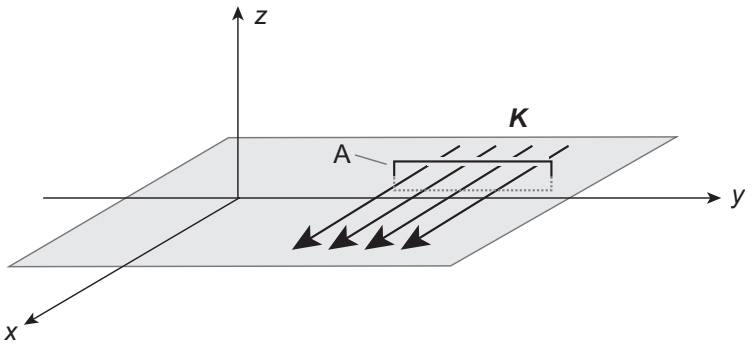


Fig. 8.9 Using the Biot–Savart law in Example 8.3 to find the magnetic field produced by an infinite sheet of surface current $\boldsymbol{K} = K\hat{\boldsymbol{x}}$ in the $x-y$ plane.

if a problem has sufficient symmetry to allow B to be pulled out of the integral $\oint \boldsymbol{B}\cdot d\boldsymbol{l}$, permitting Eq. (8.27) to be solved easily for the magnetic field.

8.5 Biot–Savart Law for Surface and Volume Currents

The integral form of the Biot–Savart Law for line currents given in Eq. (8.10) may be generalized easily to apply to surface currents and volume currents. For a 2D surface current $\boldsymbol{K}$, the Biot–Savart law takes the form

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{K}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,da' = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{K}(\boldsymbol{x}')\times\hat{\boldsymbol{R}}}{R^2}\,da' \tag{8.28}$$

while for a 3D volume current $\boldsymbol{J}$

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x' = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times\hat{\boldsymbol{R}}}{R^2}\,d^3x', \tag{8.29}$$

where the second form of each expression employs the compact notation of Eqs. (2.10) and (2.11).

Example 8.3 Let us use Eq. (8.28) to calculate the magnetic field due to a surface current $\boldsymbol{K} = K\hat{\boldsymbol{x}}$ in the $x-y$ plane, as illustrated in Fig. 8.9. From the Biot–Savart law the magnetic field must be perpendicular to $\boldsymbol{K}$ so it cannot have an x component and cannot have a z component by symmetry in y since any vertical contribution at $+y$ would be cancelled by one at $-y$. Thus the magnetic field has only y components and by the right-hand rule it points left above the plane of Fig. 8.9 and right below it. The rectangular Ampèrian loop A is parallel to the yz plane and is half above and half below the xy surface. Applying Ampère's law (8.27) to the loop A gives

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = 2BL = \mu_0 I_{\rm enc} = \mu_0 KL,$$

where L is the width of the rectangle and I_{enc} is the current enclosed by the rectangle. Thus $B = \frac{1}{2}\mu_0 K$ and the magnetic field produced by the sheet of current $\boldsymbol{K} = K\hat{\boldsymbol{x}}$ in Fig. 8.9 is

$$\boldsymbol{B} = \begin{cases} +\dfrac{1}{2}\,\mu_0 K\hat{\boldsymbol{y}} & (z<0), \\ -\dfrac{1}{2}\,\mu_0 K\hat{\boldsymbol{y}} & (z>0), \end{cases} \tag{8.30}$$

where $\hat{\boldsymbol{y}}$ is a unit vector in the y direction.

In Example 8.3 the magnetic field produced by a uniform sheet of current is independent of the distance from the surface containing the current in Example 8.3, just as would be the case for the electric field produced by a uniform surface charge in electrostatics.

8.6 Vector Potentials and Gauge Invariance

We have seen in the preceding section that the basic laws of magnetostatics are defined in differential form through Eqs. (8.23) and (8.18),

$$\boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J}, \qquad \boldsymbol{\nabla}\cdot\boldsymbol{B} = 0, \tag{8.31}$$

which must be solved for the magnetic field $\boldsymbol{B}$. In electrostatics the electric potential Φ proved to be 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 $\boldsymbol{B} = -\boldsymbol{\nabla}\Phi_{\text{M}}$. Then in Eq. (8.31) $\boldsymbol{\nabla}\times\boldsymbol{B} = 0$, and $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ reduces to the Laplace equation for Φ_{M}, implying that the methods for solving Laplace's equation developed for electrostatics in Chs. 2-6 become applicable. However, this approach has limited utility because it is valid only for regions where the current density vanishes. As we now discuss, an approach with potentially broader application 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 $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ will hold everywhere if $\boldsymbol{B}$ is the curl of some vector field $\boldsymbol{A}$ (the vector potential),

$$\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{\nabla}\times\boldsymbol{A}(\boldsymbol{x}), \tag{8.32}$$

since then the identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0$ ensures that the divergence of $\boldsymbol{B}$ vanishes under all conditions. In fact, $\boldsymbol{B}$ was already written in this form in Eq. (8.17),

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\,\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \tag{8.33}$$

and upon comparing Eqs. (8.32) and (8.33), a vector potential $\boldsymbol{A}$ consistent with the phenomenology of magnetism takes the general form

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x' + \boldsymbol{\nabla}\chi(\boldsymbol{x}), \tag{8.34}$$

where the addition of the *arbitrary scalar function* $\chi(\boldsymbol{x})$ has no effect on $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ because of the identity $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}f) = 0$ [see Eq. (A.7)] for a scalar function f. Since $\chi(\boldsymbol{x})$ is an arbitrary scalar function of $\boldsymbol{x}$, the vector potential $\boldsymbol{A}$ can be transformed freely according to

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi, \tag{8.35}$$

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 $\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi$ 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. 18.

From the Helmholtz theorem (Box 3.1), a vector field with suitable boundary conditions is specified uniquely by its curl and its divergence. From Eq. (8.32), specifying the magnetic field requires only the curl of $\boldsymbol{A}$. Thus we are free to make gauge transformations on 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 classical electromagnetism.

Substituting $\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A}$ into Eq. (8.23) gives

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A}) = \mu_0\boldsymbol{J},$$

which becomes

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \nabla^2\boldsymbol{A} = \mu_0\boldsymbol{J}, \tag{8.36}$$

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 let us exploit the freedom of gauge transformations (8.35) by choosing a gauge where

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0 \qquad \text{(Coulomb gauge condition)}, \tag{8.37}$$

is satisfied, 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. (8.36) is transformed into the Poisson equation,

$$\nabla^2\boldsymbol{A} = -\mu_0\boldsymbol{J} \qquad \text{(Coulomb gauge)}. \tag{8.38}$$

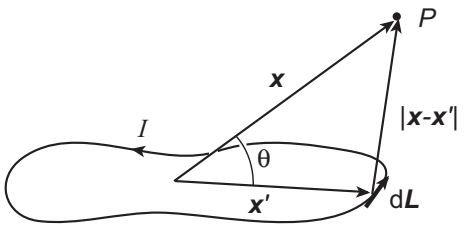


Fig. 8.10 A small current loop that produces a vector potential $\boldsymbol{A}(\boldsymbol{x})$ and corresponding magnetic field $\boldsymbol{B}(\boldsymbol{x})$ at the point P, where it is assumed that $\boldsymbol{x} \gg \boldsymbol{x}'$.

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. (8.34) with $\chi =$ constant,

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x', \tag{8.39}$$

provided that $\boldsymbol{J}(\boldsymbol{x}') \to 0$ sufficiently rapidly at infinity. Classical gauge transformations will be discussed in more depth in Ch. 11.

8.7 Magnetic Fields of Localized Currents

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; Fig. 8.10 illustrates, where a loop that is small compared with the distance to an observer at P carries a steady current I. A full treatment of this problem by analogy with electric multipole expansions is possible using *vector spherical harmonics*, which are described briefly in Box 8.2. We shall avoid using them and confine ourselves instead to lowest-order multipole expansions.

Let us assume that $\boldsymbol{x} \gg \boldsymbol{x}'$ and expand the denominator of Eq. (8.39) in a multipole series of the form (3.48) for the current distribution in Fig. 8.10,

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \frac{1}{r} \sum_{n=0}^{\infty} \left(\frac{r'}{r}\right)^n P_n(\cos\theta) d\boldsymbol{l}, \tag{8.40}$$

where $r \equiv |\boldsymbol{x}|$, $r' \equiv |\boldsymbol{x}'|$, and θ is indicated in Fig. 8.10. Thus the vector potential can be expanded as

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0 I}{4\pi} \left[\frac{1}{r} \oint d\boldsymbol{l}' + \frac{1}{r^2} \oint r' \cos\theta d\boldsymbol{l}' + \frac{1}{r^3} \oint (r')^2 \left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right) d\boldsymbol{l}' + \cdots \right]. \tag{8.41}$$

The first term in Eq. (8.41) vanishes because the integral of the current vector over all

Box 8.2 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. [4], 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 $\boldsymbol{r}$).

orientations will average to zero.[8] 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 ($\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$). Defining the *magnetic moment* $\boldsymbol{m}$ by the integral

$$\boldsymbol{m} = \frac{1}{2}\int \boldsymbol{x}'\times\boldsymbol{J}(\boldsymbol{x}')\,d^3x', \tag{8.42}$$

and the *magnetic moment density* or *magnetization* $\boldsymbol{M}(\boldsymbol{x})$ by

$$\boldsymbol{M}(\boldsymbol{x}) = \frac{1}{2}\left[\boldsymbol{x}\times\boldsymbol{J}(\boldsymbol{x})\right], \tag{8.43}$$

for $I = |\boldsymbol{J}|$ the multipole expansion (8.41) of the vector potential can be written as

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\hat{\boldsymbol{r}}}{r^2} + \text{ higher-order multipole terms},$$

where the first term has the form of a *magnetic dipole*. The higher-order multipoles in this expression have increasingly larger powers of $|\boldsymbol{x}|$ in their denominators and at large distances from the current source the vector potential for a steady-state current distribution can be approximated well by retaining only the magnetic dipole term (see Problem 8.8,

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\hat{\boldsymbol{r}}}{r^2}. \tag{8.44}$$

[8] This argument is intuitive; a more formal proof that the first term vanishes is given in Section 5.6 of Ref. [19].

Then the magnetic field may be found by taking the curl of $\boldsymbol{A}$, giving the components (see Problem 8.9)

$$\boldsymbol{B}_r = \frac{\mu_0 m}{2\pi r^3}\cos\theta \qquad \boldsymbol{B}_\theta = \frac{\mu_0 m}{4\pi r^3}\sin\theta \qquad \boldsymbol{B}_\phi = 0. \tag{8.45}$$

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,

$$|\boldsymbol{m}| = IA, \tag{8.46}$$

where I is the current and 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 ith 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

$$\boldsymbol{m} = \sum_i \frac{q_i}{2M_i}\boldsymbol{L}_i. \tag{8.47}$$

If all the particles have the same charge-to-mass ratio $q_i/M_i = e/M$,

$$\boldsymbol{m} = \frac{e}{2M}\sum_i \boldsymbol{L}_i = \frac{e}{2M}\boldsymbol{L}, \tag{8.48}$$

where $\boldsymbol{L}$ is the total orbital angular momentum. Equation (8.48) is the well-known classical connection between angular momentum and magnetic moment. It holds approximately for orbital angular momentum even on the atomic scale, but fails for moments arising from intrinsic particle spin, which is a uniquely 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 [13]. More advanced treatments may be found in Jackson [19], Garg [11], Chaichian et al [5], and Zangwill [42].

Problems

8.1 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

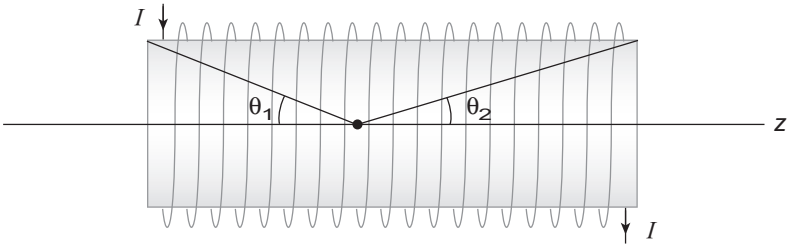
$$A_\phi^{\text{in}}(r,\theta) = -\frac{\mu_0}{4\pi}\sum_l m_l r^l P_l^1(\cos\theta),$$

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.

8.2 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.

8.3 It was asserted in Eq. (8.39) that the vector potential in Coulomb gauge is given by

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',$$

provided that the current distribution $\boldsymbol{J}(\boldsymbol{x}')$ vanishes sufficiently rapidly at infinity. Prove that this solution does indeed satisfy the Coulomb gauge condition.

8.4 Show that the vector potential $\boldsymbol{A} = \frac{1}{2}\boldsymbol{B}_0 \times \boldsymbol{x}$ corresponds to a magnetic field $\boldsymbol{B}_0$. ***

8.5 (a) Starting from the classical Lorentz-force equation, show that $E_y = v_x B_z$ in the Hall effect experimental diagram shown in Box 8.1.

(b) The electron velocity $\boldsymbol{v}$ in the Hall experiment depends on interactions with the lattice in the 2D Hall device. In the classical *Drude model* of electron transport (see

Section 7.3.2) the force is given by $\boldsymbol{F} = m(d\boldsymbol{v}/dt + \boldsymbol{v}/\tau)$, where m is the effective electron mass and τ is the Drude-model relaxation time. Assuming steady-state electron flow and defining the *cyclotron frequency* $\omega_c \equiv eB/m$, show that the classical equations of motion for Hall-device electrons in the Drude approximation are

$$v_x = -\frac{e\tau}{m}E_x - \omega_c \tau v_y \qquad v_y = -\frac{e\tau}{m}E_y + \omega_c \tau v_x \qquad v_z = -\frac{e\tau}{m}E_z,$$

where e is the electronic charge, m is the electron mass, $\boldsymbol{B}$ is the magnetic field, and $\boldsymbol{E}$ is the electric field.

(c) Use the results from parts (a) and (b) to show that in the figure of Box 8.1,

$$E_y = -\frac{eB\tau}{m}E_x.$$

Hint: At equilibrium, $v_y = 0$.

(d) The current density $\boldsymbol{J}$ in the Drude model is

$$\boldsymbol{J} = \frac{e^2 n_e}{m}\tau\boldsymbol{E} = \sigma\boldsymbol{E} \qquad \sigma \equiv \frac{e^2 n_e \tau}{m},$$

where n_e is the electron number density, σ is the conductance, and τ is the Drude-model relaxation time. Using the figure in Box 8.1, prove that

$$E_y = \frac{V_H}{w} \qquad I = w j_x \qquad R_H \equiv \frac{V_H}{I} = -\frac{B}{en_e} \qquad R_L \equiv \frac{V_L}{I} = \frac{L}{w\sigma},$$

where V_H is the *Hall voltage* (associated with the induced electric field E_y), I is the total current, J_x is the current density in the x direction, $q = -e$ was assumed for electrons, R_H is the *Hall resistance,* and R_L is the *longitudinal resistance* (along the x axis) of the Hall device shown in Box 8.1.

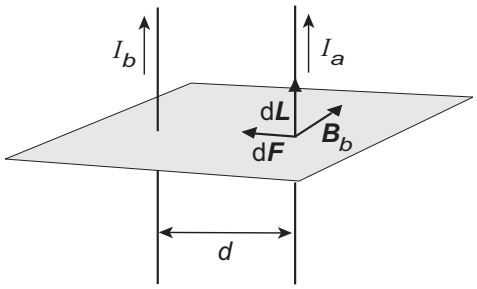
8.6 (a) Show that for a 2D Hall device like the one shown in Box 8.1, with length L and width w, that $J = \sigma E$ (where J is the current density, E is the electric field, and σ is the conductivity) is equivalent to the simple form of Ohm's law, $V = IR$. *Hint*: See Section 7.3.1 and Box 7.1.

(b) Show that for the Hall device shown in Box 8.1, the resistance components $R_{ij} = V_i/I_j$ and resistivity components $\rho_{ij} = E_i/j_j$ are related by

$$R_{yx} = \rho_{yx} \qquad R_{xx} = (L/w)\rho_{xx}.$$

Verify the final forms of the ρ and σ tensors in Box 7.1. ***

8.7 Show that the magnitude of the force per unit length acting between two infinitely long parallel wires carrying steady currents I_a and I_b, respectively, and separated by a distance d,



is given by

$$\frac{dF}{dI_a} = \frac{\mu_0 I_a I_b}{2\pi d},$$

and is attractive if the currents are in the same direction and repulsive if they are in the opposite directions.

8.8 Prove that for the current loop in Fig. 8.10 the vector potential is at large distance from the source is given by

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\hat{\boldsymbol{r}}}{r^2},$$

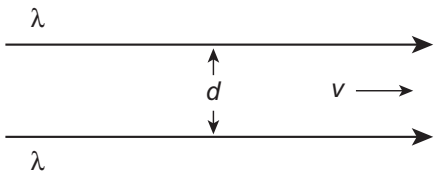
where we define $r \equiv |\boldsymbol{x}|$ and $r' \equiv |\boldsymbol{x}'|$. *Hint*: Use a multipole expansion assuming $r \gg r'$, and that

$$\oint (\hat{\boldsymbol{r}}\cdot\boldsymbol{r}')d\boldsymbol{L}' = -\hat{\boldsymbol{r}}\times\int d\boldsymbol{a}' = -\hat{\boldsymbol{r}}\times\boldsymbol{a}$$

in Fig. 8.10, where $\boldsymbol{a}$ is the "vector area" of the loop. ***

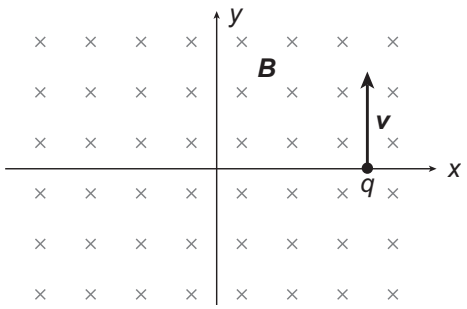
8.9 Find the formula for the magnetic field corresponding to the vector potential derived in Problem 8.8

8.10 Consider two parallel, infinite line charges λ separated by a distance d and and moving in the same direction at a speed v, as viewed in the laboratory frame. The following figure illustrates.



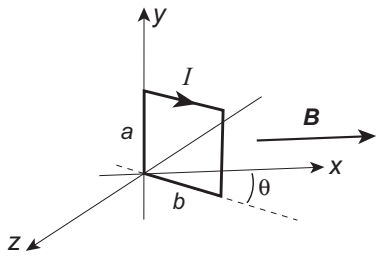
Show that for the electrostatic repulsion between the charges to be equal in magnitude to the magnetic attraction, v must be equal to the speed of light c. *Hint*: Remember that the SI constants ε_0 and μ_0 are related to each other through c. ***

8.11 Consider a particle with positive charge q and mass m moving in a uniform magnetic field $\boldsymbol{B}$ directed into the page with initial velocity $\boldsymbol{v}$ in the $x-y$ plane perpendicular to the field and to the x axis, as illustrated in the following diagram.



Show that the resulting force acting on the particle has constant magnitude $F_M = qvB$, with orientation perpendicular to both $\boldsymbol{B}$ and $\boldsymbol{v}$. Thus show that the positively charged particle executes clockwise circular motion in the $x-y$ plane with the force always directed toward the center of a circular orbit of radius $R = mv^2/qvB$, with orbital period $T = 2\pi m/qB$. ***

8.12 A rectangular loop consists of $N = 100$ closely wrapped turns and has dimensions $a = 0.4$ m and $b = 0.3$ m. The loop is hinged along the y-axis and its plane makes an angle of $\theta = 30^\circ$ with the x-axis.



What is the magnitude of the torque exerted on the loop by a uniform magnetic field $B = 0.8$ T directed along the x-axis when the steady current is $I = 1.2$ A in the direction shown? What is the expected direction of rotation of the loop? ***

9 Magnetic Fields in Matter

Just as it was shown in Ch. 6 that electric fields can polarize matter electrostatically, magnetic fields can polarize matter magnetically. 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. 6, 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. *diamagnetic materials* acquire a magnetization 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.

9.1 Alignment of Current Loops in Magnetic Fields

In classical electromagnetism, magnetic interactions in matter are assumed to be a consequence of small current loops arising at the atomic and molecular level.[1] As shown in Probs. 9.2 and 9.3, these current loops experience *no net force* in a uniform magnetic field

[1] Some early investigators of magnetism were enamored of the idea that magnetic effects were attributable to Coulomb-like interactions with magnetic dipoles formed from north and south magnetic poles, analogous to electric dipoles formed from a positive and negative charge (this is often called the *Gilbert model*). Today we know that there are no distinct magnetic charges and a more realistic picture of magnetism is based on small loops of electrical current (this is often called the *Ampèrian model*). These fictitious magnetic charges can be used as a basis for theoretical approaches that give approximately "correct" results for long-distance behavior, but that tend to fail at short distance. We shall always use the Ampèrian picture of magnetism arising from tiny current loops.

if they carry steady currents, but they do experience *a net torque* that tends to align them with the magnetic field. It is this alignment of current loops by the magnetic field that leads to magnetization of matter. If the magnetic interaction is expanded in a multipole series (see Section 8.7), only the lowest orders contribute when observing from a large distance. For a magnetic interaction, the $l = 0$ monopole term is the total "charge", but there are no magnetic charges (no magnetic monopoles) so the $l = 0$ term vanishes and the lowest-order magnetic multipole is the dipole with $l = 1$. Thus, to good approximation we may treat the current loops as a set of magnetic dipoles.[2] Just as found for electric multipole moments in Section 6.2.2, the low-order moments dominate for a distant observer. The difference here is that there is no monopole moment for magnetism, so it is dominated typically by the magnetic dipole moments.

Thus the current loops may be replaced by magnetic dipoles in the analysis of magnetized matter. The magnetic dipole moment vector $\boldsymbol{m}$ for a loop carrying a current I is defined by

$$\boldsymbol{m} = I \int d\boldsymbol{a} = I\boldsymbol{a}, \tag{9.1}$$

where $\boldsymbol{a}$ is the *vector area of a surface S* bounded by the loop:

$$\boldsymbol{a} = \int_S d\boldsymbol{a} = \frac{1}{2} \oint \boldsymbol{r} \times d\boldsymbol{l}, \tag{9.2}$$

where the second integral is around the boundary curve of the area S. If the loop is flat, $\boldsymbol{a}$ has a magnitude equal to the usual area enclosed and the direction of $\boldsymbol{a}$ (and of $\boldsymbol{m}$) is given by the right-hand rule applied to the direction of the current in the loop. Then the magnitude of the magnetic dipole moment vector $\boldsymbol{m}$ is

$$|\boldsymbol{m}| = Ia, \tag{9.3}$$

where I is the magnitude of the current and a is the area bounded by the loop. The torque $\boldsymbol{\tau}$ exerted on a current loop of area $\boldsymbol{a}$ carrying current I, or on the corresponding magnetic dipole moment $\boldsymbol{m}$, is

$$\boldsymbol{\tau} = I\boldsymbol{a} \times \boldsymbol{B} = \boldsymbol{m} \times \boldsymbol{B}, \tag{9.4}$$

with the direction given by the right-hand rule applied to the direction of the current within the loop.

9.2 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 since the magnetic polarization of the matter will modify current densities. Now we must develop methods to solve for the fields in the presence of such polarization.

[2] Retaining only the dipoles and not higher-order multipoles is justified if the distance to the observer greatly exceeds the size of the current loop. This is almost always true, since the current loops are atomic-scale.

9.2.1 Macroscopic Averaging

The atomic currents in matter produce rapidly fluctuating current densities on a microscopic scale. Just as discussed in Section 6.7 for electric fields in matter, only averages over small macroscopic volumes are known and only these enter into the classical equations of electromagnetism. The first step in introducing averaging in matter for electric fields in Section 6.7 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 $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$. In a similar manner, for magnetic fields the macroscopic averaging procedure leads to the same equation $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ that we introduced in Eq. (8.18). 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, $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$.

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

$$\boldsymbol{M}(\boldsymbol{x}) = \sum_i N_i \langle \boldsymbol{m}_i \rangle, \tag{9.5}$$

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

$$\Delta \boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \left[\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} + \frac{\boldsymbol{M}(\boldsymbol{x}') \times (\boldsymbol{x} - \boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|^3} \right] \Delta V, \tag{9.6}$$

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. (8.44). Letting ΔV tend to d^3x', the total vector potential at $\boldsymbol{x}$ is given by an integral over all space,

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \int \left[\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} + \frac{\boldsymbol{M}(\boldsymbol{x}') \times (\boldsymbol{x} - \boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|^3} \right] d^3x'. \tag{9.7}$$

The second (magnetization) term can be cast in another form by utilizing Eq. (8.16) to write

$$\int \frac{\boldsymbol{M}(\boldsymbol{x}') \times (\boldsymbol{x} - \boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|^3} d^3x' = \int \boldsymbol{M}(\boldsymbol{x}') \times \boldsymbol{\nabla}' \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) d^3x'.$$

Integration by parts then allows Eq. (9.7) to be written

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \int \frac{\left[\boldsymbol{J}(\boldsymbol{x}') + \boldsymbol{\nabla}' \times \boldsymbol{M}(\boldsymbol{x}') \right]}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x', \tag{9.8}$$

where a surface term has been discarded by assuming $\boldsymbol{M}(\boldsymbol{x}')$ to be localized and well behaved.

9.2.2 The Auxiliary Field H and Constituitive Relations

From the results in the preceding section we see that the magnetization contributes an effective current density

$$\boldsymbol{J}_{\mathrm{M}} = \boldsymbol{\nabla} \times \boldsymbol{M}. \tag{9.9}$$

Then $\boldsymbol{J} + \boldsymbol{J}_{\mathrm{M}}$ plays the role of the effective current such that

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0(\boldsymbol{J} + \boldsymbol{\nabla} \times \boldsymbol{M}). \tag{9.10}$$

It is then convenient to define a new macroscopic field $\boldsymbol{H}$ by[3]

$$\boldsymbol{H} \equiv \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M}. \tag{9.11}$$

The introduction of $\boldsymbol{H}$ in the presence of magnetically polarized materials is a matter of convenience, analogous to the introduction in Section 6.7 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 $\boldsymbol{B}$.[4] 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. (8.31) are

$$\begin{aligned} \boldsymbol{\nabla} \times \boldsymbol{H} &= \boldsymbol{J}, \\ \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0, \end{aligned} \tag{9.12}$$

which are analogous to the macroscopic fields for electrostatics in medium

$$\begin{aligned} \boldsymbol{\nabla} \times \boldsymbol{E} &= 0, \\ \boldsymbol{\nabla} \cdot \boldsymbol{D} &= \rho, \end{aligned} \tag{9.13}$$

that were derived in Sections 6.7 and 6.8.

Just as for Eqs. (9.13), the description of macroscopic magnetostatics in Eqs. (9.12) 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,

$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{9.14}$$

[3] 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 [19] 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).

[4] This is a classical statement. As we shall discuss in Ch. 18, 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*; also called *minimal substitution*).

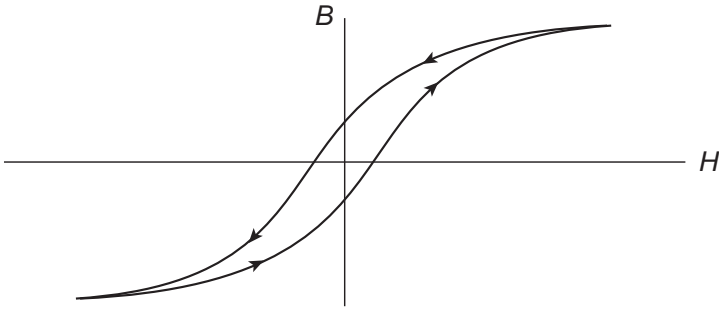


Fig. 9.1 Schematic illustration of magnetic hysteresis.

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

$$\chi = \left(\frac{\mu}{\mu_0} - 1 \right). \tag{9.15}$$

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 9.1.

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.[5] 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. 9.1 illustrates. These complex behaviors lead to a constituitive relationship

$$\boldsymbol{B} = \boldsymbol{F}[\boldsymbol{H}], \tag{9.16}$$

where the notation $\boldsymbol{F}[\boldsymbol{H}]$ indicates that $\boldsymbol{F}$ is a *non-linear function of* $\boldsymbol{H}$. The complicated nonlinear 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.

[5] 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 $T > T_c$, a ferromagnet then behaves as a paramagnet. An introduction to spontaneously broken symmetry may be found in Chs. 17 and 18 of Ref. [17].

Box 9.1 **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

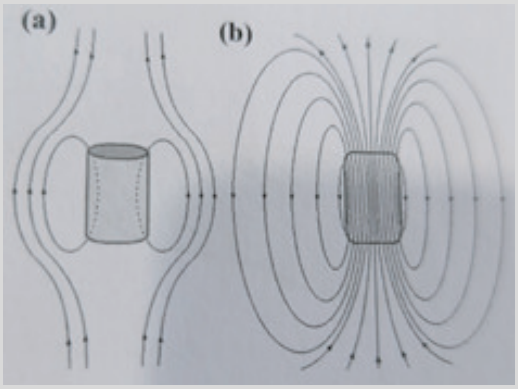
In diamagnetic media an external field $\boldsymbol{H}$ creates a magnetization $\boldsymbol{M}$ opposite to $\boldsymbol{H}$, so $\chi < 0$. Physically, external fields induce currents associated with orbital motion in the diamagnetic material. (The external field also interacts with electron spins, 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 [5].



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. 18, this *Higgs mode of spontaneous symmetry breaking* is fundamental for the Standard Model of elementary particle physics.

9.2.3 Magnetic Boundary-Value Conditions

Boundary conditions for $\boldsymbol{B}$ and $\boldsymbol{H}$ at media interfaces were considered in Section 6.11; Eqs. (6.56b) and (6.57b) 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

$$(\boldsymbol{B}_2 - \boldsymbol{B}_1) \cdot \boldsymbol{n} = 0, \tag{9.17a}$$

$$\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{K}, \tag{9.17b}$$

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 (9.14) and have finite conductivities so that $\boldsymbol{J} = \sigma \boldsymbol{E}$ and $\boldsymbol{K} = 0$, boundary conditions can be expressed as [19],

$$\boldsymbol{B}_2 \cdot \boldsymbol{n} = \boldsymbol{B}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{B}_2 \times \boldsymbol{n} = \frac{\mu_2}{\mu_1} \boldsymbol{B}_1 \times \boldsymbol{n}, \tag{9.18}$$

or as

$$\boldsymbol{H}_2 \cdot \boldsymbol{n} = \frac{\mu_1}{\mu_2} \boldsymbol{H}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{H}_2 \times \boldsymbol{n} = \boldsymbol{H}_1 \times \boldsymbol{n}. \tag{9.19}$$

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. In the next section we shall address methods of solving magnetostatic boundary value problems.

9.3 Solving Magnetostatic Boundary-Value Problems

Magnetostatic boundary value problems require solving Eqs. (9.12) subject to constituitive relations as in Eqs. (9.14) and (9.16). 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 [19].

9.3.1 Solutions Using a 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 (9.16) 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

$$\boldsymbol{\nabla} \times \left(\frac{1}{\mu} \boldsymbol{\nabla} \times \boldsymbol{A} \right) = \boldsymbol{J}. \tag{9.20}$$

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

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \nabla^2\boldsymbol{A} = \mu\boldsymbol{J}. \tag{9.21}$$

Invoking the Coulomb gauge condition (8.37) by setting $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$ we obtain the Poisson equation

$$\nabla^2\boldsymbol{A} = -\mu\boldsymbol{J}. \tag{9.22}$$

Comparing with Eq. (8.38) 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. (9.22) across interfaces between different (linear) media can be implemented using the boundary conditions (9.18) or (9.19).

9.3.2 Solutions Using a Magnetic Scalar Potential

As mentioned in Section 8.6, 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

$$\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}}. \tag{9.23}$$

If the medium is linear and μ is constant the magnetic scalar potential satisfies the Laplace equation

$$\nabla^2\Phi_{\mathrm{M}} = 0, \tag{9.24}$$

for which the boundary conditions (9.18) 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 9.3.3.

9.3.3 Solutions for Hard Ferromagnets

A *hard ferromagnet* is one having a magnetization sufficiently stable 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.

Using the Magnetic Scalar Potential: Since $\boldsymbol{J} = 0$ for a hard ferromagnetic, the magnetic scalar potential method of Section 9.3.2 is applicable. Then $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ becomes

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = \mu_0\boldsymbol{\nabla}\cdot(\boldsymbol{H}+\boldsymbol{M}) = 0, \tag{9.25}$$

where Eq. (9.11) was used, and since $\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}}$ from Eq. (9.23), this becomes a *magnetic Poisson equation*,

$$\nabla^2\Phi_{\mathrm{M}} = \boldsymbol{\nabla}\cdot\boldsymbol{M} = -\rho_{\mathrm{M}}, \tag{9.26}$$

where an *effective magnetic-charge density*

$$\rho_{\rm M} \equiv -\boldsymbol{\nabla}\cdot\boldsymbol{M} \tag{9.27}$$

has been introduced. By analogy with the second term of Eq. (6.21) for electrostatics in electrically polarized matter, if there are no boundary surfaces the solution is expected to be [19]

$$\Phi_{\rm M}(\boldsymbol{x}) = -\frac{1}{4\pi}\int \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x = -\frac{1}{4\pi}\boldsymbol{\nabla}\cdot\int \frac{\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x, \tag{9.28}$$

where the second form results from an integration by parts and Eq. (8.16), 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 [19]

$$\sigma_{\rm M} = \boldsymbol{n}\cdot\boldsymbol{M}, \tag{9.29}$$

where $\boldsymbol{n}$ is the outward normal at the surface. Then the first form of the potential (9.28) is modified to

$$\Phi_{\rm M}(\boldsymbol{x}) = -\frac{1}{4\pi}\int_V \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x + \frac{1}{4\pi}\oint_S \frac{\boldsymbol{n}'\cdot\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da'. \tag{9.30}$$

An important special case is for uniform magnetization of the volume. Then $\boldsymbol{J}' = 0$ inside the volume and the first term of Eq. (9.30) vanishes, leaving only the surface term,

$$\Phi_{\rm M}(\boldsymbol{x}) = \frac{1}{4\pi}\oint_S \frac{\boldsymbol{n}'\cdot\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da' = \frac{1}{4\pi}\oint_S \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da', \tag{9.31}$$

where Eq. (9.29) 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. (9.29).

Using the Vector Potential: If we choose to satisfy $\boldsymbol{\nabla}\cdot\boldsymbol{B}$ automatically by introducing a vector potential $\boldsymbol{A}$ and defining the magnetic field by $\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A}$, then the first of Eqs. (9.12) is

$$\boldsymbol{\nabla}\times\boldsymbol{H} = \boldsymbol{\nabla}\times\left(\frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M}\right) = 0, \tag{9.32}$$

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. (9.9) on the right side,

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \nabla^2\boldsymbol{A} = \mu_0\boldsymbol{J}_{\rm M}, \tag{9.33}$$

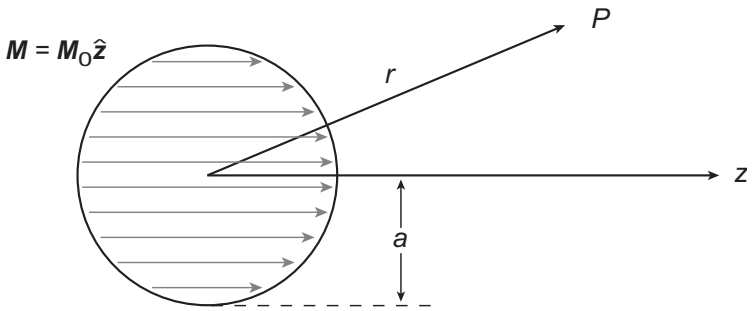


Fig. 9.2 A uniformly magnetized sphere of radius a, with a sharp surface and a surface magnetic charge density $\sigma_{\rm M}(\theta)$; $\boldsymbol{M} = \boldsymbol{M}_0\hat{\boldsymbol{z}}$ is parallel to the z axis, so that $\sigma_{\rm M} = \boldsymbol{n}\cdot\boldsymbol{M} = M_0\cos\theta$.

where $\boldsymbol{J}_{\rm M}$ is the effective magnetic current density defined in Eq. (9.9). Transforming to Coulomb gauge (8.37) by setting $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$ then leads to a Poisson equation for the vector potential,

$$\nabla^2\boldsymbol{A} = -\mu_0\boldsymbol{J}_{\rm M}. \tag{9.34}$$

If there are no bounding surfaces, the solution is

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x \tag{9.35}$$

[compare Eq. (9.8) with $\boldsymbol{J} = 0$]. If the magnetization is discontinuous (as in the uniformly magnetized sphere with sharp surface in Section 9.3.4), it is necessary to add a surface integral to Eq. (9.28). For the case of $\boldsymbol{M}$ falling to zero at the surface S bounding the volume V, starting from Eq. (9.7) the proper generalization of Eq. (9.35) may be shown to be [19]

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x + \frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}')\times\boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da'. \tag{9.36}$$

For the special case that the magnetization is constant over the volume V, only the second (surface integral) term can contribute in Eq. (9.36).

9.3.4 Example: Uniformly Magnetized Sphere

As an example of solving boundary problems in magnetostatics using the methods just discussed, let us consider a spherical ball with uniform permanent magnetization that has a sharp transition at the surface from magnetized to non-magnetized matter; Fig. 9.2 illustrates.

Solution Using the Magnetic Scalar Potential: The sphere is uniformly magnetized, so there are no volume currents and the scalar magnetic potential method of Section 9.3.2 is applicable. As discussed in Section 9.3.3, 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 $\sigma_{\rm M} = \boldsymbol{n}\cdot\boldsymbol{M} = \boldsymbol{M}_0\cos\theta$ in spherical coordinates, from Eq. (9.28) the scalar magnetic potential is

$$\Phi_{\rm M}(r,\theta) = \frac{1}{4\pi}\oint_S \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da' = \frac{M_0a^2}{4\pi}\int \frac{\cos\theta}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega'.$$

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. (3.48) and using that $P_1(\cos\theta) = \cos\theta$ gives

$$\begin{aligned}\Phi_{\rm M}(r,\theta) &= \frac{M_0a^2}{4\pi}\int\frac{P_1(\cos\theta)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega' \\ &= \frac{M_0a^2}{4\pi}\sum_{l=0}^{\infty}\frac{r_<^l}{r_>^{l+1}}\int P_l(\cos\theta)P_1(\cos\theta)\,d\Omega' \\ &= \frac{1}{3}M_0a^2\frac{r_<}{r_>^2}P_1(\cos\theta) = \frac{1}{3}M_0a^2\frac{r_<}{r_>^2}\cos\theta, \end{aligned} \tag{9.37}$$

where the orthogonality condition of Eq. (4.39) has eliminated all terms in the sum except for $l = 1$, and $r_<$ and $r_>$ are the smaller and larger of r and a, respectively.

Interior solution: Inside the sphere, $r_< = r$ and $r_> = a$, so the magnetic scalar potential is given by

$$\Phi_{\rm M}^{\rm in} = \frac{1}{3}M_0r\cos\theta = \frac{1}{3}M_0z.$$

Then from Eq. (9.23)

$$\boldsymbol{H}_{\rm in} = -\boldsymbol{\nabla}\Phi_{\rm M}^{\rm in} = -\frac{\partial}{\partial z}\left(\frac{1}{3}M_0z\right)\hat{\boldsymbol{z}} = -\frac{1}{3}M_0\hat{\boldsymbol{z}} = -\frac{1}{3}\boldsymbol{M},$$

and from Eq. (9.11),

$$\boldsymbol{B}_{\rm in} = \mu_0(\boldsymbol{H}_{\rm in}+\boldsymbol{M}) = \mu_0\left(-\frac{1}{3}\boldsymbol{M}+\boldsymbol{M}\right) = \frac{2\mu_0}{3}\boldsymbol{M}.$$

Therefore, the interior solution is

$$\Phi_{\rm M}^{\rm in} = \frac{1}{3}M_0z \qquad \boldsymbol{H}_{\rm in} = -\frac{1}{3}\boldsymbol{M} \qquad \boldsymbol{B}_{\rm in} = \frac{2\mu_0}{3}\boldsymbol{M}, \tag{9.38}$$

where we see that the fields are constant inside the sphere, with $\boldsymbol{B}$ parallel and $\boldsymbol{H}$ antiparallel to $\boldsymbol{M}$.

Exterior solution: For the exterior solution, $r_< = a$ and $r_> = r$, giving from Eq. (9.37) for the exterior potential,

$$\Phi_{\rm M}^{\rm out} = \frac{1}{3}M_0a^3\frac{\cos\theta}{r^2}, \tag{9.39}$$

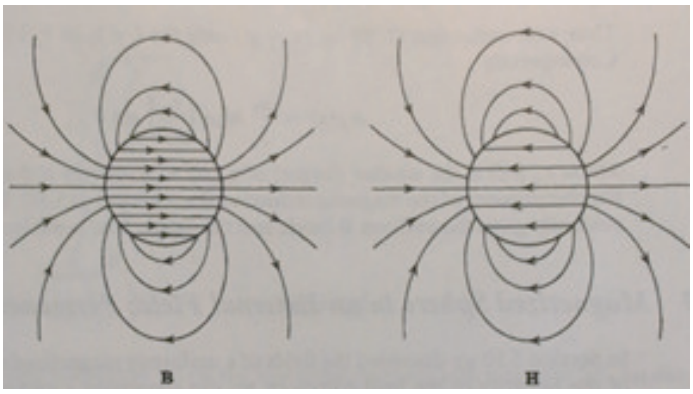


Fig. 9.3 Lines of $\boldsymbol{B}$ and $\boldsymbol{H}$ for a uniformly magnetized sphere having a sharp boundary [19].

which may be recognized as a dipole potential with dipole moment

$$\boldsymbol{m} = \frac{4\pi a^3}{3}\boldsymbol{M}. \tag{9.40}$$

Then, proceeding as above the exterior solution is

$$\boldsymbol{H}_{\text{out}} = -\boldsymbol{\nabla}\Phi_{\text{M}}^{\text{out}} = -\frac{1}{3}\frac{a^3}{r^3}\boldsymbol{M} \qquad \boldsymbol{B}_{\text{out}} = \mu_0(\boldsymbol{H}+\boldsymbol{M}) = \left(\frac{3r^3-a^3}{3r^3}\right)\mu_0\boldsymbol{M}. \tag{9.41}$$

The $\boldsymbol{B}$ and $\boldsymbol{H}$ fields for the interior and exterior solutions are plotted in Fig. 9.3

Solution Using the Vector Potential: Alternatively, the vector potential and Eq. (9.36) can be used to obtain the solution for the problem posed in Fig. 9.2. The magnetization is assumed uniform inside the sphere so the volume current $\boldsymbol{J}_{\text{M}} = 0$ and the first (volume) term in Eq. (9.36) make no contribution. However, there is a surface charge because of the sharp boundary on magnetization so the second term in Eq. (9.36) will be non-zero and

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}')\times\boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da'. \tag{9.42}$$

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

$$A_\phi = \frac{\mu_0}{4\pi}M_0a^2\int\frac{\sin\theta'\cos\phi'}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega', \tag{9.43}$$

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

$$\sin\theta'\cos\phi' = -\sqrt{\frac{8\pi}{3}}\,\mathrm{Re}\,[Y_{11}(\theta',\phi')],$$

where $\mathrm{Re}\,(x)$ denotes the real part of x. Thus,

$$A_\phi = -\sqrt{\frac{3}{8\pi}}\frac{\mu_0}{4\pi}M_0 a^2 \int \frac{\mathrm{Re}\,[Y_{11}(\theta',\phi')]}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega'.$$

Then if the denominator is expanded using Eq. (3.51),

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

the spherical harmonic orthogonality relation (3.56),

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

ensures that only the $l=1, m=1$ term survives the summation and the vector potential is

$$A_\phi(\boldsymbol{x}) = \frac{\mu_0}{3}M_0 a^2\left(\frac{r_<}{r_>^2}\right)\sin\theta. \tag{9.44}$$

For the inside solution, $r_< = r$ and $r_> = a$, so

$$A_\phi^{\mathrm{in}}(\boldsymbol{x}) = \frac{\mu_0}{3}M_0 r\sin\theta. \tag{9.45}$$

You are asked to calculate the corresponding $\boldsymbol{B}$ and $\boldsymbol{H}$ fields for the interior of the sphere in Problem 9.1.

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 9.1 illustrates.

Example 9.1 Consider Fig. 9.4, 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 9.3.2 is applicable and from Eq. (9.23)

$$\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}},$$

and since $\boldsymbol{B} = \mu\boldsymbol{H}$, the equation $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$ is replaced by $\boldsymbol{\nabla}\cdot\boldsymbol{H} = 0$ in all regions. Therefore, the magnetic potential Φ_{M} satisfies the Laplace equation everywhere and the problem reduces to solving the Laplace equation subject to the boundary conditions of Eq. (9.19) at $r = a$ and $r = b$.

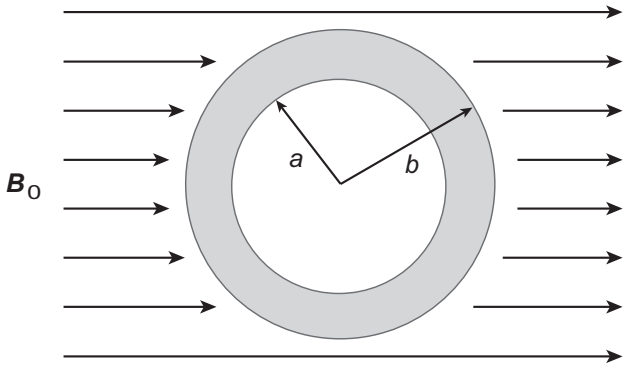


Fig. 9.4 A shell of material with permeability μ is placed in a previously uniform magnetic field $\boldsymbol{B}_0$. Example 9.1 demonstrates that if the shell contains highly permeable material the cavity inside the shell is shielded strongly from the magnetic field.

Exterior solution: If $r > b$, the potential is of the form

$$\Phi_{\rm M} = -H_0 r\cos\theta + \sum_{l=0}^{\infty} \frac{\alpha_l}{r^{l+1}} P_l(\cos\theta) \qquad (r > b), \tag{9.46}$$

which gives a uniform field $\Phi_{\rm M} = \boldsymbol{H}_0$ at large distance.

Interior solution: Likewise, the potential must take the form

$$\Phi_{\rm M} = \sum_{l=0}^{\infty} \left(\beta_l r^l + \gamma_l \frac{1}{r^{l+1}}\right) P_l(\cos\theta). \qquad (a < r < b), \tag{9.47a}$$

$$\Phi_{\rm M} = \sum_{l=0}^{\infty} \delta_l r^l P_l(\cos\theta) \qquad (r < a). \tag{9.47b}$$

in interior regions.

Matching boundary conditions: The boundary conditions at $r = a$ and $r = b$ require the components H_θ and B_r be continuous, which implies that

$$\frac{\partial\Phi_{\rm M}}{\partial\theta}(b_+) = \frac{\partial\Phi_{\rm M}}{\partial\theta}(b_-) \qquad \frac{\partial\Phi_{\rm M}}{\partial\theta}(a_+) = \frac{\partial\Phi_{\rm M}}{\partial\theta}(a_-), \tag{9.48a}$$

$$\mu_0\frac{\partial\Phi_{\rm M}}{\partial r}(b_+) = \mu\frac{\partial\Phi_{\rm M}}{\partial r}(b_-) \qquad \mu\frac{\partial\Phi_{\rm M}}{\partial r}(a_+) = \mu_0\frac{\partial\Phi_{\rm M}}{\partial r}(a_-), \tag{9.48b}$$

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 (9.48) determine the unknown constants in Eqs. (9.46) and (9.47). One finds that all coefficients with $l \neq 1$ vanish and the $l = 1$ coefficients satisfy the simultane-

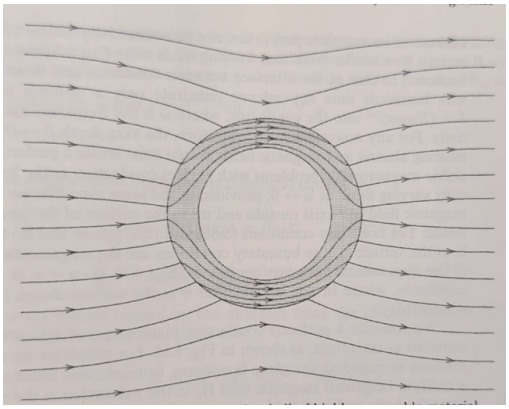


Fig. 9.5 The magnetic shielding effect of a shell of highly permeable material. Lines for the magnetic field $\boldsymbol{B}$ are shown [19]. Note the absence of field lines for the cavity inside the shell. Calculation described in Example 9.1.

ous equations

$$\begin{aligned} \alpha_1 - b^3\beta_1 - \gamma_1 &= b^3 H_0, \\ 2\alpha_1 + \mu' b^3\beta_1 - 2\mu'\gamma_1 &= -b^3 H_0, \\ a^3\beta_1 + \gamma_1 - a^3\delta_1 &= 0, \\ \mu' a^3\beta_1 - 2\mu'\gamma_1 - a^3\delta_1 &= 0, \end{aligned} \tag{9.49}$$

where $\mu' \equiv \mu/\mu_0$, which may be solved simultaneously for the unknown coefficients. For example, the solution of Eqs. (9.49) for α_1 and δ_1 are [19],

$$\begin{aligned} \alpha_1 &= \left(\frac{(2\mu'+1)(\mu'-1)}{(2\mu'+1)(\mu'+2) - 2\dfrac{a^3}{b^3}(\mu'-1)^2} \right) (b^3 - a^3) H_0, \\ \delta_1 &= -\left(\frac{9\mu'}{(2\mu'+1)(\mu'+2) - 2\dfrac{a^3}{b^3}(\mu'-1)^2} \right) H_0. \end{aligned} \tag{9.50}$$

The corresponding magnetic field lines are shown in Fig. 9.5. 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

$$\alpha_1 \longrightarrow b^3 H_0 \qquad -\delta_1 \longrightarrow \frac{9\mu_0}{2\mu(1 - a^3/b^3)} H_0. \tag{9.51}$$

Thus the field in the inner cavity of Fig. 9.5 is proportional to μ^{-1} and a shield made of highly permeable material causes a large reduction of the magnetic field inside the cavity.

Example 9.1 illustrates that even thin shells of material with $\mu/\mu_0 \sim 10^3 - 10^6$ can greatly reduce the interior magnetic field, as is clear from Fig. 9.5.

Background and Further Reading

The presentation in this chapter has followed Jackson [19] rather closely. See also Garg [11], Chaichian et al [5], and Zangwill [42].

Problems

9.1 Find the fields $\boldsymbol{B}_{\text{in}}$ and $\boldsymbol{H}_{\text{in}}$ corresponding to the vector potential of Eq. (9.45),

$$A_{\phi}^{\text{in}}(\boldsymbol{x}) = \frac{\mu_0}{3} M_0 r \sin\theta,$$

inside the sphere of Fig. 9.2.

9.2 Classical magnetism can be viewed as resulting from current loops in matter. Analyze the magnetic forces acting on a rectangular current loop carrying a steady current I in a uniform magnetic field $\boldsymbol{B}$. Show that the total magnetic force acting on the loop is

$$\boldsymbol{F}_{\text{M}} = I \oint (d\boldsymbol{s} \times \boldsymbol{B}),$$

and that the total force vanishes for a loop carrying a steady current in a uniform magnetic field. ***

9.3 In Problem 9.2 and Section 8.2 we showed that there is no net force acting on a loop carrying a steady current that is immersed in a uniform magnetic field. Show that there *is however a net torque* acting on current loops in uniform magnetic fields, and that the net effect of this torque is to align magnetic dipole moments with the applied magnetic field by aligning the normal vector of the plane of current loops with $\boldsymbol{B}$.

10 Dynamical Charges and Currents

In the discussion to this point we have considered primarily static charges (no relative motion) as sources of electric fields and steady-state charge currents (no change in time) as the source of magnetic fields. However, many important electromagnetic phenomena involve the motion of charges and/or non-steady electrical currents. Hans Christian Ørsted (1777-1827) discovered the magnetic effect of the electric current, establishing the first connection between electric and magnetic phenomena, and André–Marie Ampère (1775-1827) extended that work, finding in 1823 the circuit law between electric current passing through a loop and magnetic field around the loop and establishing a formula describing the interaction of two currents. However, it was Michael Faraday who performed the most quantitative experiments on time-dependent electric and magnetic field, and their relationship, beginning in 1831. As we shall see, these experiments implied that a time-varying electric field produces a magnetic field and a time-varying magnetic field produces an electric field, and paved the way for the understanding of classical electromagnetism that was elucidated fully by Maxwell.

10.1 Faraday and the Law of Induction

Faraday extended the earlier insights of Ørsted and Ampère by concentrating on time-varying electric and magnetic phemomena. In a nutshell, 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, Faraday established experimentally that a current flows in a test circuit if a current in a nearby circuit changes over time, or if the two circuits move with respect to each other, or if a magnetic field varies near the test circuit. Faraday interpreted these results as originating in a *changing magnetic flux* that

1. is produced an electric field around the test circuit,
2. which leads to an *electromotive force* $\mathscr{E}$ (EMF) that drives a current in the test circuit governed by Ohm's law (see Section 7.3.1).

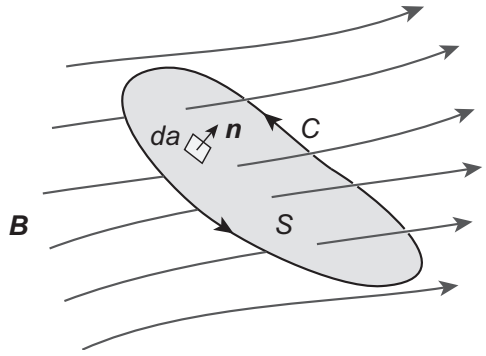


Fig. 10.1

Magnetic flux $\boldsymbol{B}$ through a circuit C bounding a surface S, with a unit normal to the surface $\boldsymbol{n}$ and surface element da.

Thus, Faraday's results and his interpretation of those results represent the first steps in the unification of electricity and magnetism into a single theory of *electromagnetism*.

10.2 Mathematical Description of Faraday's Results

Faraday was a talented experimentalist with incisive intuition, but without much formal training in the sciences or mathematics. Maxwell and others are largely responsible for transforming Faraday's brilliant insights and measurements into mathematical language. Let us express the results of Faraday's observations mathematically. Consider Fig. 10.1, where the magnetic flux passing through the area enclosed by the circuit indicated by the contour C is

$$\mathscr{F}_{\mathrm{M}} = \int_S \boldsymbol{B} \cdot \boldsymbol{n} da \tag{10.1}$$

and the electromotive force $\mathscr{E}$ around the circuit that causes a current flow according to Ohm's law is

$$\mathscr{E} = \oint_C \boldsymbol{E}' \cdot d\boldsymbol{l}, \tag{10.2}$$

where $\boldsymbol{E}'$ is the electric field at element $d\boldsymbol{l}$ of the circuit C. All of Faraday's observations listed above may be explained by the relationship

$$\mathscr{E} = -k \frac{d\mathscr{F}_{\mathrm{M}}}{dt}.$$

between the rate of change of the magnetic flux $d\mathscr{F}_{\mathrm{M}}/dt$ and the EMF $\mathscr{E}$, where k is a constant of proportionality that can be determined by requiring Galilean invariance at low velocities.[1] This indicates that $k = 1$ in SI units (and $k = c^{-1}$ in Gaussian units). Therefore,

[1] As we shall see in Ch. 15, 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.

in SI units

$$\mathscr{E} = -\frac{d\mathscr{F}_{\mathrm{M}}}{dt}, \tag{10.3}$$

where the sign is specified by *Lenz's law*.[2]

> ***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. (10.3) with Eqs. (10.1) and (10.2) gives Faraday's law in integral form,

$$\oint_C \boldsymbol{E}' \cdot d\boldsymbol{l} = -\frac{d}{dt}\int_S \boldsymbol{B}\cdot\boldsymbol{n}\, da \qquad \text{(Faraday's law)}, \tag{10.4}$$

indicating that the induced EMF is proportional to the *total time derivative* of the magnetic flux.[3] Notice that this reduces to the result $\oint_c \boldsymbol{E}\cdot d\boldsymbol{l} = 0$ of Eq. (2.26) for the static case with $d\boldsymbol{B}/dt = 0$. Equation (10.4) 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. (10.4) 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 varies with time at a given point, or
2. translation of the circuit changes the location of the circuit in an external field.

This can be taken into account by computing the *convective derivative*,

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \boldsymbol{v}\cdot\boldsymbol{\nabla}, \tag{10.5}$$

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 [19],

$$\frac{d}{dt}\int_S \boldsymbol{B}\cdot\boldsymbol{n}\, da = \int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}\, da + \oint_C (\boldsymbol{B}\times\boldsymbol{v})\cdot d\boldsymbol{l}, \tag{10.6}$$

which can be used to write Eq. (10.4) in the form

$$\oint_C \left[\boldsymbol{E}' - (\boldsymbol{v}\times\boldsymbol{B})\right]\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}\, da. \tag{10.7}$$

This is Faraday's law applied to the moving circuit C, but if we think of the circuit C and

[2] Griffiths [13] states the Lenz law in somewhat more picturesque form "Nature abhors a change in flux". For a dramatic example of just how much Nature detests magnetic flux changes, see Problem 10.1.

[3] The flux can be changed by changing the magnetic field, or the shape, position, or orientation of the circuit. The total time derivative d/dt appearing in Eq. (10.4) 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. (10.4) to that circuit at fixed location gives,

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}\, da, \tag{10.8}$$

where $\boldsymbol{E}$ is now the electric field in the laboratory frame. Galilean invariance (valid at low velocities) then requires the left sides of Eqs. (10.7) and (10.8) to be equivalent.

If the circuit is held fixed in a reference frame so that the electric and magnetic fields are defined in the same frame, Faraday's law (10.4) 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. (10.4) 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 expression in parentheses must vanish identically, giving

$$\boldsymbol{\nabla} \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{10.9}$$

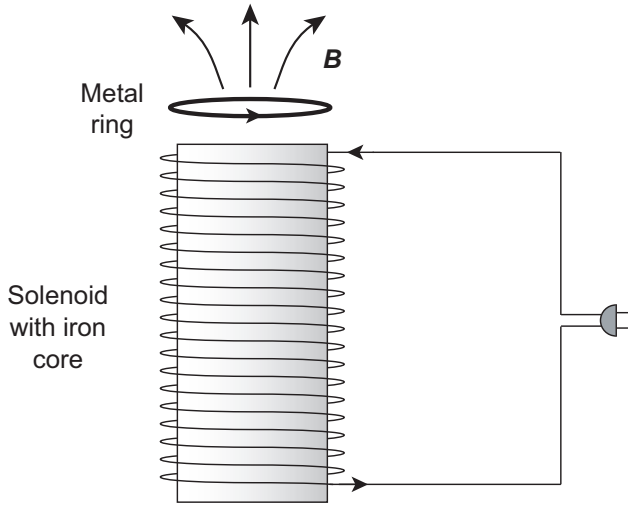
which is Faraday's law (1.1b) in differential form. Note that this is the time-dependent generalization of $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$ from Eq. (2.27) in electrostatics. In a static system the curl of $\boldsymbol{E}$ vanishes but that is no longer true in the presence of a time-dependent magnetic field.

Background and Further Reading

See Griffith's [13] for an introduction and Garg [11] and Zangwill [42] for more advanced discussions.

Problems

10.1 Consider a metal ring placed on top of a strong solenoid (enhanced by an iron core), as illustrated in the following figure.



If the solenoid is plugged in the metal ring will jump into the air, by as much as several feet in typical demonstrations. Explain why. *Hint*: See Refs. [36, 39] for a discussion of this demonstration. ***

11 Maxwell's Equations

The preceding chapters have provided a systematic explanation and validation of the various pieces of electromagnetic theory that James Clerk Maxwell synthesized into what we now call the *Maxwell equations*. To this point we have treated electricity and magnetism largely as separate subjects. As Faraday's discoveries that were discussed in Ch. 10 make clear, this distinction begins to fail when one moves from static to time-dependent phenomena, and we will now begin to address a unified picture of electromagnetism, valid for electric and magnetic phenomena in both static and dynamical contexts. 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.

11.1 The Almost-but-Not-Quite Maxwell's Equations

The basic equations of electricity and magnetism that we have studied in Chs. 2-9 can be summarized (in medium, in differential form, in SI units) as

$$\boldsymbol{\nabla}\cdot\boldsymbol{D}=\rho \qquad \text{(Gauss's law)}, \tag{11.1a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E}+\frac{\partial\boldsymbol{B}}{\partial t}=0 \qquad \text{(Faraday's law)}, \tag{11.1b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B}=0 \qquad \text{(No magnetic charges)}, \tag{11.1c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{H}=\boldsymbol{J} \qquad \text{(Ampère's law)}, \tag{11.1d}$$

Expressed here 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 that were introduced in Eqs. (6.50). 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 (6.50d).

11.1.1 Ampère's Law and the Displacement Current

It is important to recognize that all but Faraday's law in Eqs. (11.1) were derived from data taken under steady-state conditions, so we should expect that modifications might be required for the description of time-dependent fields. Maxwell realized that the fundamental equations of electromagnetism (11.1) known in his time were inconsistent as they then stood. Since the divergence of a curl vanishes by a basic vector-calculus identity,

$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{H})=0$, it follows from Ampère's law (11.1d) for steady currents that

$$\boldsymbol{\nabla}\cdot\boldsymbol{J}=\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{H})=0. \tag{11.2}$$

But from the continuity equation (1.3) ensuring conservation of charge,

$$\frac{\partial\rho}{\partial t}=-\boldsymbol{\nabla}\cdot\boldsymbol{J},$$

so $\boldsymbol{\nabla}\cdot\boldsymbol{J}=0$ can hold only if the charge density doesn't change with time. Therefore, Maxwell modified Ampère's law to accomodate a time dependence by introducing a *displacement current* term $\partial\boldsymbol{D}/\partial t$,[1]

$$\boldsymbol{\nabla}\times\boldsymbol{H}=\boldsymbol{J}\quad\longrightarrow\boldsymbol{\nabla}\times\boldsymbol{H}=\boldsymbol{J}+\frac{\partial\boldsymbol{D}}{\partial t}, \tag{11.3}$$

thus converting Eqs. (11.1) into what we now call the Maxwell equations (6.50) (in medium, in differential form),[2]

$$\boldsymbol{\nabla}\cdot\boldsymbol{D}=\rho \qquad \text{(Gauss's law)}, \tag{11.4a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E}+\frac{\partial\boldsymbol{B}}{\partial t}=0 \qquad \text{(Faraday's law)}, \tag{11.4b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B}=0 \qquad \text{(No magnetic charges)}, \tag{11.4c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{H}-\frac{\partial\boldsymbol{D}}{\partial t}=\boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{11.4d}$$

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. As you were asked to show way back in Problem 1.1, these (Maxwell) equations are now consistent with the continuity equation. Taking the divergence of both sides of the Ampère–Maxwell law,

$$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{H})=\boldsymbol{\nabla}\cdot\boldsymbol{J}+\frac{\partial(\boldsymbol{\nabla}\cdot\boldsymbol{D})}{\partial t},$$

and using $\boldsymbol{\nabla}\cdot\boldsymbol{D}=\rho$ from Gauss's law and the identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{H})=0$ from Eq. (A.6), this becomes

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

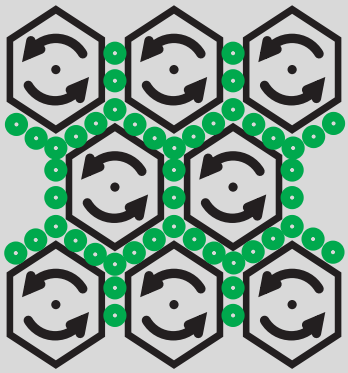
which is the continuity equation (1.3).

[1] As elaborated in Box 11.1, Maxwell's original detailed physical reason for the displacement term and corresponding name is partially bogus, being based on a now-discredited mechanical model of electric and magnetic fields in the fictitious aether. However, his insight that the displace term is required was seminal, and the name itself survives in the modern form of the Maxwell equations.

[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 Maxwell equations in medium in integral form are given by Eqs. (6.55).

Box 11.1 Maxwell, Steampunk, and the Displacement Current

Maxwell's reasons for introducing the displacement current were based partially on a now-discredited mechanical model of the (fictitious) aether, which was thought in the 19th century to be the special medium through which electromagnetic waves traveled. These ideas of a mechanical aether have strong overlap with the modern literary and animé genre called *steampunk*, which is a fictional alternative history in which modern electrical and internal combustion technologies fail to develop and all mechanical devices (including flying machines) are driven by steam, as described in Ref. [31]. A schematic illustration of Maxwell's mechanical model (the "elastic vortex aether") is illustrated in the following figure.



Maxwell calculated that his elastic vortex aether propagated transverse waves at a speed of $3.107 \times 10^8\,\mathrm{m\,s^{-1}}$, which was close to the $3.149 \times 10^8\,\mathrm{m\,s^{-1}}$ for the speed of light that Hippolyte Fizeau had measured in 1847 using a rotating toothed wheel with mirror apparatus. Maxwell remarked [28, 29],

> "we can scarcely avoid the inference that light consists of the transverse undulations of the same medium which is the cause of electric and magnetic phenomena" ,

which was the first direct claim that light had an electromagnetic origin. The displacement term in the Maxwell equations and the electromagnetic origin of light have stood the test of time, despite their origin in a fictitious steampunk-like mechanical model of vortices and idle wheels inspired by the cogs and gears in the machinery of the Victorian age.

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.

1. This requires conservation of charge within a volume of space that can be arbitrarily small, thus implying that *charge is conserved locally*.
2. 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: de-

Maxwell's equations (11.4), 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 experimentally observable particle motion are thought to describe all of classical electromagnetism.

11.1.2 Implications of the Ampère–Maxwell Law

As was introduced in Section 1.4, 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 meaningfully 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 Maxwell's theory (see Section 13.2.1):

   > The Maxwell equations have wave solutions, and the corresponding electromagnetic waves may be interpreted as light,

   which *unifies* the field of electricity, not only with the field of magnetism, but also with the field of optics.
4. That Maxwell's equations obey the continuity equation and thus conserve charge locally leads to the idea of *electromagnetic gauge invariance*, which will provide powerful tools for dealing with classical electromagnetism, and 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 it is important to appreciate that they 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. [14].

11.2 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 8.6). The Maxwell equations (11.4) 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. 11.4. As we have noted in Section 8.6, 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]

$$\boldsymbol{B}=\boldsymbol{\nabla}\times\boldsymbol{A}. \tag{11.5}$$

Then Faraday's law (11.4b) may be written as,

$$\boldsymbol{\nabla}\times\left(\boldsymbol{E}+\frac{\partial\boldsymbol{A}}{\partial t}\right)=0. \tag{11.6}$$

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

$$\boldsymbol{E}=-\boldsymbol{\nabla}\Phi-\frac{\partial\boldsymbol{A}}{\partial t}. \tag{11.7}$$

For simplicity, let us restrict consideration to the vacuum form of the Maxwell equations that is given in Eq. (1.1), which are formulated in terms of the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$,

$$\boldsymbol{\nabla}\cdot\boldsymbol{E}=\frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)}, \tag{11.8a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E}+\frac{\partial\boldsymbol{B}}{\partial t}=0 \qquad \text{(Faraday's law)}, \tag{11.8b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B}=0 \qquad \text{(No magnetic charges)}, \tag{11.8c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B}-\frac{1}{c^2}\frac{\partial\boldsymbol{E}}{\partial t}=\mu_0\boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}. \tag{11.8d}$$

Substituting $\boldsymbol{B}=\boldsymbol{\nabla}\times\boldsymbol{A}$ and $\boldsymbol{E}=-\boldsymbol{\nabla}\Phi-\partial\boldsymbol{A}/\partial t$ into the Maxwell equations (11.8), 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 (11.8b) and (11.8c) are *satisfied identically*, while the inhomogeneous equations (11.8a)

[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.

and (11.8d) are transformed into coupled second-order differential equations,

$$\nabla^2\Phi + \frac{\partial}{\partial t}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) = -\frac{\rho}{\varepsilon_0}, \tag{11.9a}$$

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t}\right) = -\mu_0\boldsymbol{J}, \tag{11.9b}$$

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 equations (11.9). These equations can be decoupled by a suitable gauge transformation, as will now be shown.

11.3 Exploiting Gauge Symmetry

In Section 8.6 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 gause 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

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \to \Phi - \frac{\partial\chi}{\partial t}, \tag{11.10}$$

on the vector potential $\boldsymbol{A}$ and the scalar potential Φ, where χ is an arbitrary scalar function. A gauge transformation leaves the electric and magnetic fields, and thus Maxwell's equations, unchanged.

Notice that in the gauge transformation (11.10) the same scalar function χ must be used in both equations, but the choice of χ is arbitrary.

Mathematically, the invariance of electromagnetism under the gauge transformation (11.10) means that an electromagnetic field may be viewed as an *equivalence class* of potentials $\{\Phi,\boldsymbol{A}\}$, where equivalence classes are defined in Box 11.2. 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 the gauge transformation of Eq. (11.10)" and the equivalence class is the set of 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 classical electromagnetic physics when inserted in the Maxwell equations.

Box 11.2 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 $a \in \mathrm{S}$, one has $a \sim a$.
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.

11.3.1 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*

$$\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t} = 0 \qquad \text{(Lorenz gauge).} \tag{11.11}$$

A constraint like Eq. (11.11) is termed a *gauge-fixing condition* and imposing such a constraint is termed *fixing the gauge*. The gauge choice implied by Eq. (11.11) is called the *Lorenz gauge*.[4] If the Lorenz gauge condition is inserted into the coupled equations (11.9), the equations decouple and solving the Maxwell equations has now been reduced to solv-

[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 an early book [14]. Ludvig Lorenz did significant work on electromagnetism contemporary with, but independent of, Maxwell. For example, he proposed independently that electromagnetic waves might be lightwaves. An overview of Lorenz's scientific work is given in Ref. [24], and the history of his work and the incorrect attribution of his famous gauge to Lorentz is discussed in Ref. [20].

ing two second-order differential equations,

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = -\frac{\rho}{\varepsilon_0}, \tag{11.12a}$$

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J}, \tag{11.12b}$$

which are *uncoupled:* solution of Eq. (11.12a) gives the scalar potential Φ independent of the vector potential $\boldsymbol{A}$, while solution of Eq. (11.12b) gives the vector potential, independent of the scalar potential. Equations (11.12) along with the gauge condition (11.11) are completely equivalent in physical content to the original Maxwell equations (11.4).

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. (11.9), but does not satisfy Eqs. (11.11). 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

$$\nabla^2\chi - \frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2} = -\left(\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t}\right), \tag{11.13}$$

the new potentials will satisfy the Lorenz gauge conditions (11.11) and the simultaneous equations (11.12).

The Lorenz condition Eq. (11.11) does not exhaust the gauge degrees of freedom in Lorenz gauge. The *restricted gauge transformation*

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \to \Phi - \frac{\partial\chi}{\partial t}, \tag{11.14}$$

where the scalar function χ (which is arbitrary for general gauge transformations) is *restricted to those that satisfy the constraint*

$$\nabla^2\chi - \frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2} = 0, \tag{11.15}$$

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. (11.11). The Lorenz gauge is often used for two reasons:

1. It leads to the decoupled equations (11.12) that treat Φ and $\boldsymbol{A}$ on an equal footing.
2. As will be elaborated in Ch. 15, the Lorenz gauge condition is manifestly invariant under Lorentz transformations, which fits naturally into special relativity.[5]

[5] Note that the equations are in (Ludvig) *Lorenz gauge*, but we shall see in Ch. 15 that they are invariant under (Hendrik) *Lorentz transformations*. This perhaps helps to explain the historical confusion in the name of the gauge.

There are infinitely many valid gauge transformations that can be made using Eq. (11.10) 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 8.6 is the transformation to Coulomb gauge.

11.3.2 Coulomb Gauge

We have already introduced the *Coulomb gauge condition* in Eq. (8.37),

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}=0 \qquad \text{(Coulomb gauge condition).}$$

From Eq. (11.9a), in Coulomb gauge the scalar potential obeys a Poisson equation

$$\nabla^2\Phi=-\frac{\rho}{\varepsilon_0}, \tag{11.16}$$

which has a solution

$$\Phi(\boldsymbol{x},t)=\frac{1}{4\pi\varepsilon_0}\int\frac{\rho(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{11.17}$$

This is the instantaneous Coulomb potential caused by the charge density $\rho(\boldsymbol{x})$, which is the motivation for the appelation *Coulomb gauge*.

> As Box 11.3 discusses, the Coulomb potential (11.17) suggests instantaneous transmission of information in a universe where lightspeed c is the speed limit. In Chs. 15-18 we will resolve this issue and show that electromagnetism is in fact causal because *no information is being transmitted at a speed $v>c$*.

From Eq. (11.9b), in Coulomb gauge the vector potential obeys

$$\nabla^2\boldsymbol{A}-\frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2}=-\mu_0\boldsymbol{J}+\frac{1}{c^2}\frac{\partial\Phi}{\partial t}. \tag{11.18}$$

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

$$\boldsymbol{J}=\boldsymbol{J}_{\mathrm{L}}+\boldsymbol{J}_{\mathrm{T}}, \tag{11.19}$$

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*.

The longitudinal and transverse currents are given explicitly by

$$\boldsymbol{J}_{\mathrm{L}}=-\frac{1}{4\pi}\boldsymbol{\nabla}\int\frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \tag{11.20a}$$

$$\boldsymbol{J}_{\mathrm{T}}=\frac{1}{4\pi}\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{11.20b}$$

Box 11.3 Causality and the Coulomb and Gravitational Potentials

Causality (that actions must precede results) is a fundamental principle underlying all of modern science. The Coulomb potential of Eq. (11.17) appears to violate that principle.

Causality and the Coulomb Potential

The Coulomb potential in Eq. (11.17) is said to be "instantaneous" because the associated force acts without delay at any $\boldsymbol{x}$ corresponding to 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 Gravitational 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 a distance of 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, we conclude that *the speed of gravity is* c (as predicted by the general theory of relativity). More extensive discussion of physical implications for GW170817 may be found in Ref. [15] and in Section 24.7 of Ref. [16].

[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.

From Eq. (11.17) and the continuity equation (1.3),

$$\frac{1}{c^2}\boldsymbol{\nabla}\frac{\partial\Phi}{\partial t} = \mu_0 \boldsymbol{J}_{\mathrm{T}}, \tag{11.21}$$

which gives when inserted into Eq. (11.9b), upon using the identity $\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) = 0$ and Eq. (11.19),

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J}_{\mathrm{T}}. \tag{11.22}$$

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 that in describing electromagnetic radiation by quantum electrodynamics, only the vector potential (which is determined by the transverse components) need be quantized.

Notice that Eq. (11.22) 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 given above implies that for the scalar potential the propagation speed is infinite. The full resolution of this seeming paradox is beyond our scope because it requires relativistic quantum field theory, but in essence the propagating classical field has only transverse components and one finds that in quantum electrodynamics only the vector potential (and thus only the transverse components) need be quantized.

As we will see in Ch. 15, Lorentz covariance is most transparent if the 3-vector potential $\boldsymbol{A}$ and the scalar potential Φ are combined into a spacetime 4-vector A^μ with the components of the 4-vector given by

$$A^\mu = (A^0, A^1, A^2, A^3) = (\Phi, \boldsymbol{A}) = (\Phi, A^1, A^2, A^3), \tag{11.23}$$

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]

It was 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

[6] A small terminology inconsistency must be dealt with. We have been calling the 3-vector $\boldsymbol{A}$ 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); see Section 16.8.3. We adopt a policy that where it is clear that we are working non-relativistically, $\boldsymbol{A}$ will often be called the vector potential, while if it is clear that we are working in a relativistic context, A^μ will often be called the vector potential. If there is a chance for confusion the explicit names "3-vector potential" for $\boldsymbol{A}$ and "4-vector potential" for A^μ will be used. Similar terminology considerations apply to other quantities such as the 3-vector current $\boldsymbol{J}$ and the 4-vector current J^μ, each of which is called the "current vector" in various contexts.

degrees of freedom? A systematic discussion of relativistic quantum field theory applied to electrons and photons (QED) is beyond our present scope. However, the short answer is 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.

11.4 Retarded Green Function

From Eqs. (11.12), 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,

$$\Box\psi = -f, \tag{11.24}$$

where we introduce for convenience the d'Alembertian operator $\Box$ with[7]

$$\Box \equiv -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2, \tag{11.25}$$

because if one knows how to solve (11.24), one can solve Eqs. (11.12). Lorenz gauge has been assumed, so Eq. (11.11) must also be satified, but the retarded solution of Eqs. (11.12) with sources that will be found shortly will in fact satisfy Eq. (11.11) automatically, provided that the solution decreases rapidly enough at infinity. Generalizing the electrostatics case, a Green function $G(t,\boldsymbol{x};t'\boldsymbol{x}')$ can be defined by

$$\Box G(t,\boldsymbol{x};t'\boldsymbol{x}') = -\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t-t'), \tag{11.26}$$

where the derivative operators in $\Box$ are understood to operate on the unprimed variables, and in contrast to Eq. (3.21) 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. (11.24) is given by the principle of superposition,

$$\psi(t,\boldsymbol{x}) = \int G(t,\boldsymbol{x};t'\boldsymbol{x}')f(t',\boldsymbol{x}')\,d^3x'\,dt', \tag{11.27}$$

assuming that the integral converges. Let us seek a solution of Eq. (11.26) using *Fourier transforms*, guided by the discussion in Wald [40].

[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. 15.

11.4.1 Fourier Transforms

For an integrable function $F : \mathbb{R} \to \mathbb{R}$, define its *Fourier transform* $\hat{F}$ as[8]

$$\hat{F}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(x) e^{-ikx} dx. \tag{11.28}$$

Fourier transforms can also be extended to distributions such as the Dirac delta function; for example, the Fourier transform of the delta function is,

$$\hat{\delta}_{x_0}(k) = \frac{1}{\sqrt{2\pi}} e^{ilx_0}. \tag{11.29}$$

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*,

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{F}(k) e^{+ikx} dk. \tag{11.30}$$

This formula applies also to distributions and the inverse of the Fourier transform for a delta function is

$$\delta_{x_0}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{\delta}_{x_0}(k) e^{+ikx} dk = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ikx_0} e^{+ikx} dk. \tag{11.31}$$

One of the most useful properties of Fourier transforms is that differentiation in real space corresponds to multiplication by ik in Fourier transform space. For example,

$$\widehat{\frac{dF}{dx}}(k) = ik\hat{F}(k), \tag{11.32}$$

for the Fourier transform of $dF(x)/dt$, as you are asked to show in Problem 11.1.

> Thus a partial differential equation with constant coefficients expressed in real space can be converted to a corresponding algebraic equation in Fourier transform space.

Now let's solve Eq. (11.26) for G using Fourier transform methods.

11.4.2 Green Functions Using Fourier Transforms

To simplify notation let us temporarily set $x' = t' = 0$ and $c = 1$, and define the 4D Fourier transform of G as[9]

$$\hat{G}(\omega, \boldsymbol{k}) = \frac{1}{(2\pi)^2} \int_{-\infty}^{+\infty} G(t, \boldsymbol{x}) e^{+i\omega t} e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} dt\, d^3x \tag{11.33}$$

[8] There are different conventions for choosing the normalization factor in defining Fourier transforms. The definitions in Eqs. (11.28) and(11.30) employ the "symmetric convention", which uses a normalization factor of $1/\sqrt{2\pi}$ for both the transform (11.28) and the inverse transform (11.30). Note that a simple example of a solution using Green functions for a driven mechanical oscillator was described in Box 3.2, where a different normalization for Fourier transforms was used.

[9] 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. 15.

Taking the Fourier transform of Eq. (11.26) with respect to t and $\boldsymbol{x}$, using Eq. (11.29) with $x_0 = 0$, and using Eq. (11.32) yields

$$(\omega^2 - k^2)\,\hat{G}(\omega,\boldsymbol{k}) = -\frac{1}{4\pi^2}, \tag{11.34}$$

where $k \equiv |k|$. Naively, this suggest the solution

$$\hat{G}(\omega,\boldsymbol{k}) = -\frac{1}{4\pi^2}\frac{1}{(\omega^2-k^2)} = -\frac{1}{4\pi^2}\frac{1}{(\omega+k)(\omega-k)}, \tag{11.35}$$

but division by $(\omega^2 - k^2)$ is illegal since it is possible that $\omega = k$. To see the difficulty 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 $\tilde{G}$,

$$\begin{aligned}\tilde{G}(t,\boldsymbol{k}) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\hat{G}(\omega,\boldsymbol{k})e^{-i\omega t}\,d\omega. \\ &= -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k)(\omega-k)}\,d\omega,\end{aligned} \tag{11.36}$$

where Eq. (11.35) was used. This has *logarithmic divergences* of the integral[10] if $\omega \to \pm k$, so it is ill-defined.

The fundamental reason that the integral in (11.36) is not well defined as it stands is that many different Green functions satisfy Eq. (11.26), so a unique solution for G in (11.36) cannot be expected without providing further information to identify the particular Green function that we seek. To proceed it is necessary to regularize the integration in Eq. (11.36) in such a way that

1. equation (11.34) remains valid,
2. the right side of Eq. (11.36) becomes well defined, and
3. the resulting solution has the desired boundary and causality properties.

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 this solution the poles should be displaced into the lower half of the complex ω-plane; thus we define the retarded Green function by

$$\tilde{G}(t,\boldsymbol{k})_{\rm ret} = -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k+i\varepsilon)(\omega-k+i\varepsilon)}\,d\omega, \tag{11.37}$$

where ε is positive and the limit $\varepsilon \to 0$ is to be taken after the integral is evaluated. Viewing the integral in Eq. (11.37) as a contour integral in the complex ω-plane, if $t < 0$ the integration contour can be closed by a large semicircle in the upper half plane since the exponential factor $\exp(-i\omega t)$ is damped there, as illustrated in Fig. 11.1(a). The poles are

[10] All diverges are trouble, but some are worse than others. The prototypical logarithmic divergence occurs in an integral of the form $f(x) = \int_{x_0}^{x} d\Lambda/\Lambda$. This integral diverges as $x \to \infty$, which is relatively mild, growing only as $f(x) \sim \log(x)$ at large x.

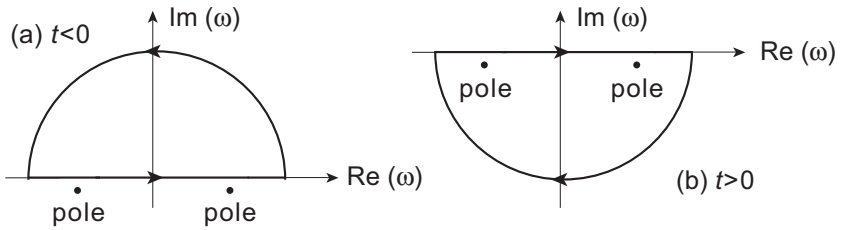


Fig. 11.1 Contour integration paths for the the retarded Green function. (a) Path for $t < 0$, which encloses no poles so it gives zero. (b) Path for $t > 0$, which encloses both poles and the value of the integral comes from the sum of residues at the two poles.

in the negative half-plane so the contour does not enclose them and by the *Cauchy residue theorem*,

$$\tilde{G}(t,\boldsymbol{k})_{\rm ret} = 0 \qquad (t<0). \tag{11.38}$$

That the Green function vanishes for $t = 0$ is characteristic of the retarded solution, corresponding to 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 integration contour can be closed in the lower half-plane, as illustrated in Fig. 11.1(b). Now the contour encloses poles at $\omega = \pm k - i\varepsilon$ and by the Cauchy theorem the integral evaluates to $2\pi i$ times a sum of residues at the two poles. The residue $\operatorname{res} f(a)$ of a function $f(z)$ at a pole $z = a$ is given by

$$\operatorname{res} f(a) = \frac{1}{(m-1)!}\lim_{z\to 0}\left(\frac{d^{m-1}}{dz^{m-1}}(z-a)^m f(z)\right), \tag{11.39}$$

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. (11.37) the poles are of order $m = 1$ and we obtain

$$\tilde{G}(t,\boldsymbol{k})_{\rm ret} = \frac{2\pi i}{4\pi^2\sqrt{2\pi}}\left(\frac{e^{-ikt}}{2k} - \frac{e^{+ikt}}{2k}\right) = \frac{1}{2\pi\sqrt{2\pi}}\frac{\sin kt}{k} \qquad (t>0). \tag{11.40}$$

The retarded Green function in real (position) space then results from taking the inverse transform of Eq. (11.40) with respect to k (homework),

$$\begin{aligned} G(t,\boldsymbol{x})_{\rm ret} &= \frac{1}{(2\pi)^{3/2}}\frac{1}{2\pi\sqrt{2\pi}}\int \frac{\sin kt}{k}\, e^{i\boldsymbol{k}\cdot\boldsymbol{x}} d^3k \\ &= \frac{\delta(t-|\boldsymbol{x}|)}{4\pi|\boldsymbol{x}|} \qquad (t>0). \end{aligned} \tag{11.41}$$

Restoring t', $\boldsymbol{x}'$, and c, this result becomes

$$G(t,\boldsymbol{x};t',\boldsymbol{x}')_{\rm ret} = \begin{cases} 0 & (t<t'), \\ \dfrac{\delta(t-t'-|\boldsymbol{x}-\boldsymbol{x}'|/c)}{4\pi|\boldsymbol{x}-\boldsymbol{x}'|} & (t>t'). \end{cases} \tag{11.42}$$

This propagator now has the desired causal behavior.[11]

Using the language of special relativity that will be introduced in Ch. 15, $G(t,\boldsymbol{x};t',\boldsymbol{x}')_{\text{ret}}$ is non-zero only *on the future lightcone* of the source point $(t',\boldsymbol{x}')$, meaning that it vanishes unless $|\boldsymbol{x}-\boldsymbol{x}'| = c(t-t')$. That is, if a field satisfying the wave equation $\Box\psi = -f$ of Eq. (11.24) is zero at early times and

1. a δ-function source is placed at $(t',\boldsymbol{x}')$,
2. the resulting disturbance of the field (electromagnetic waves) will propagate away from the source at the speed of light.

Because the retarded Green function respects the principle of causality (cause precedes effect) it is said to be a *causal Green function*.

Thus, light corresponds to a wave solution of the Maxwell equations that is fixed by the theory to have a constant speed of propagation $v = c$, independent of inertial frame. Electricity and magnetism are now unified in the science of electromagnetism, and optics has now become the science of electromagnetic waves.

11.4.3 Causal Solutions for Vector and Scalar Potentials

The retarded solution of Eq. (11.24) is the solution obtained using the retarded Green function in Eq. (11.27),

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t-|\boldsymbol{x}-\boldsymbol{x}'|/c,\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{11.43}$$

This can also be written more compactly as

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t',\boldsymbol{x}')_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \tag{11.44}$$

where the notation means that $f(t',\boldsymbol{x}')$ is to be evaluated at the retarded time

$$t' = t - \frac{|\boldsymbol{x}-\boldsymbol{x}'|}{c}, \tag{11.45}$$

and where t' isn't independent but rather is a function of t, $\boldsymbol{x}$, and $\boldsymbol{x}'$.[12] Written in this form the retarded solution looks like a solution of the Poisson equation, but evaluated at the retarded time (11.45).

[11] The retarded time has a simple meaning. When there is a local change in an electromagnetic field, the field at some distant point doesn't reflect that change instantaneously. Instead, the signal of a change in the field propagates at the speed of light and there is a time delay in its action at a distant point; the time at the distant point minus the time for the signal to propagate is termed the *retarded time*.

[12] In Eq. (11.44), the integration is not over all of space at an instant of time (as the notation might suggest). Rather, it is restricted to the *surface defined by 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 and causality in Ch. 15).

The Maxwell equations (11.12a) and (11.12b) for the scalar and vector potentials are of the form (11.24), so we can write immediately the corresponding retarded solutions,

$$\Phi(t,\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{[\rho(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x', \tag{11.46a}$$

$$\boldsymbol{A}(t,\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{J}(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'. \tag{11.46b}$$

Finally, substitution of Eqs. (11.46a) and (11.46b) in Eq. (11.11) and some manipulation indicates that these solutions are indeed consistent with the Lorenz gauge condition (see Problem 11.3). Therefore, we may conclude that the solution of the Maxwell equations for a charge density ρ and a current density $\boldsymbol{J}$, with no initial incoming radiation, is given by Eqs. (11.46). This has been shown here in for the special case of 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 [19], Garg [11], and Wald [40]. Further useful discussion may be found in Refs. [5, 13, 42].

Problems

11.1 One of the most useful aspects of Fourier transforms is that differentiation in real space corresponds to multiplication by a factor ik in Fourier transform space, which converts differential equations into algebraic equations. As an example, prove that

$$\widehat{\frac{dF}{dx}}(k) = ik\hat{F}(k),$$

for the Fourier transform of dF/dx. ***

11.2 Supply and explain the steps in verifying the result of Eq. (11.41).

11.3 Prove that Eqs. (11.46) are consistent with the Lorenz gauge condition (11.11). ***

11.4 For problems with a magnetic field oriented along the z axis, two gauge choices that can be made for the vector potential $\boldsymbol{A} = (A_x, A_y, A_z)$ are

$$\boldsymbol{A} = (0, Bx, 0) \qquad \text{(Landau gauge)},$$

$$\boldsymbol{A} = \frac{1}{2}B(-y, x, 0) \qquad \text{(Symmetric gauge)}.$$

Demonstrate that both of these gauges lead to the same magnetic field $\boldsymbol{B} = (B_x, B_y, B_z) = (0, 0, B)$, where $B = |\boldsymbol{B}|$.

12 Conservation Laws

In Box 1.1 we summarized concisely the symmetries that the equations of classical electromagnetism exhibit. In this chapter we elaborate on the corresponding conservation laws that are implied by the symmetries of the Maxwell equations. The prototype conservation law in electromagnetism is charge conservation, exmplified in the continuity equation (1.3)

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{J} = 0 \qquad \text{(Continuity equation)},$$

discussed in Section 1.3, which relates the change of charge density in a local region to flow of charge through the boundaries of that region. Various other electromagnetic conservation laws can be formulated in a similar manner.

12.1 Conservation of Energy

Problems

12.1 *No problems yet for this chapter.*

13 Electromagnetic Waves

Perhaps the consequence of classical electromagnetism having the greatest everyday import is that the Maxwell equations have traveling wave solutions that can propagate through space carrying energy, momentum, and angular momentum. The continuous arrival of life-sustaining energy from the Sun, your smartphone, and the satellite navigation system in your phone or car are but a few implications. These propagating electromagnetic fields are termed *free fields*, because the electric field lines do not terminate in charges and the magnetic field lines do not encircle current for these solutions.

In the 19th century it was commonly held that waves required a medium for transmission, based on experience with waves like water waves that *did* propagate through a medium. Thus was hypothesized the existence of a medium specifically for transmission of the electromagnetic waves predicted by Maxwell's theory that was called the *luminiferous aether* or just *aether* for short,[1] and electromagnetic waves were hypothesized to travel in a preferred inertial frame that was at rest with respect the aether. The aether hypothesis held sway for decades, but its death knell sounded from both observational and theoretical quarters as the 19th century turned into the 20th century:

1. The *Michelson–Morley light interferometry experiment* in 1887 found no evidence for motion of the Earth through the aether (this hypothesized effect was called *aether drift*).
2. The publication of Einstein's *special theory of relativity* in 1905 rendered the aether superfluous, by asserting that there are no preferred inertial frames and that electromagnetic waves traveled at a constant speed c in every inertial frame.

The idea that electromagnetic waves travel at the same speed in all inertial frames was radical, but correct, and special relativity implied that both electromagnetism and particle mechanics were governed by Lorentz invariance. With the aether now irrelevant, let us speak of it no more and consider electromagnetic waves, first in vacuum, and then in matter.

13.1 The Classical Wave Equation

Wave phenomena in classical physics are described by a *wave equation*, which is a second-order partial differential equation that in one spatial dimension and cartesian coordinates

[1] To accomplish its task, while remaining faithful to the rest of observed reality, the aether was required to possess rather miraculous properties. It was assumed that the aether filled all of space but that it was invisible (since no one could see it), that it was rigid (because electromagnetic waves are transverse and normal transverse waves require a rigid body for transmission), and that it influenced only electromagnetic waves (so that it didn't disturb our understanding of non-electromagnetic phenomena).

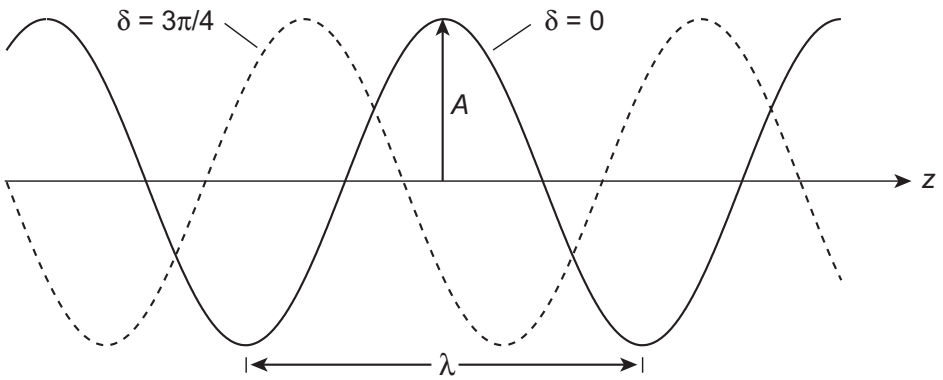


Fig. 13.1 Sinusoidal solutions (13.4) of a 1D wave equation with amplitude A and wavelength $\lambda = 2\pi$. The solid curve has phaseshift $\delta = 0$ and the dashed curve has a phaseshift $\delta = 3\pi/4$. The two waves have the same amplitude and wavelength, but they are shifted in phase by $3\pi/4$.

takes the form

$$\frac{\partial^2 \psi(z,t)}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2 \psi(z,t)}{\partial t^2} = \mathscr{F}(z,t), \tag{13.1}$$

where $\psi(z,t)$ is a scalar field, v is the speed of the wave,[2] and we have allowed the possibility of a source term $\mathscr{F}(z,t)$. The 1D classical wave equation admits solutions of the form

$$\psi(z,t) = \psi(z \pm vt), \tag{13.2}$$

with the $\pm$ sign indicating whether the wave is traveling in the positive or negative z direction. The wave equation is linear, meaning that a sum of solutions is also a solution, and the most general 1D solution is of the form

$$\psi(z,t) = f(z+vt) + g(z-vt), \tag{13.3}$$

which represents a superposition of left-going (negative-z direction) and right-going (positive-z direction) waves.

Example 13.1 The sinusoidal waveform

$$\psi(z,t) = A\cos[k(z-vt) + \delta], \tag{13.4}$$

that is illustrated in Fig. 13.1 is a representative and instructive solution of the classical wave equation (13.1), both because many simple wave solutions have a similar form, and because complex waveforms can be built up by a linear combination of such waves (*Fourier series*). As illustrated in Fig. 13.1, the *amplitude* A is the maximum displacement of the

[2] Specifically v is the *phase velocity* of the wave described by the function ψ, which will depend on the nature of the wave. For example, for transverse waves on a string $v = (T/\mu)^{1/2}$, where T is the string tension and μ is the mass per unit length of the string, while we shall see that electromagnetic waves in vacuum propagate at the speed of light c.

wave. The argument of the cosine function is called the *phase* of the wave, and k is the *wave number*, which is related to the wavelength λ by

$$k = \frac{2\pi}{\lambda} \qquad \text{(Wavenumber)}, \tag{13.5}$$

with the wavelength displayed in Fig. 13.1. As illustrated in Fig. 13.1, if the wave is assumed to be propagating to the right in the positive x direction, the *phaseshift* δ represents a shift to the left (a delay) of the waveform. As time passes the the entire wave proceeds to the right in Fig. 13.1 with a constant speed v, such that a distance $\Delta x = 2\pi/k$ corresponds to one complete cycle of the cosine function, and one wavelength. At a fixed point x, the wave amplitude vibrates up and down, executing one complete cycle in a *period* T given by

$$T = \frac{2\pi}{kv}. \tag{13.6}$$

The corresponding *frequency* ν represents the number of oscillations per unit time and is given by

$$\nu = \frac{1}{T} = \frac{k\nu}{2\pi} = \frac{\nu}{\lambda}, \tag{13.7}$$

which may also be expressed as the *angular frequency* ω, with

$$\omega \equiv 2\pi\nu = k\nu, \tag{13.8}$$

which allows Eq. (13.4) to be written more concisely as

$$\psi(x,t) = A\cos(kz - \omega t + \delta). \tag{13.9}$$

It is sometimes useful to write a waveform in complex notation using the *Euler formula*,

$$e^{i\theta} = \cos\theta + i\sin\theta, \tag{13.10}$$

which allows writing Eq. (13.9) in the form

$$\psi(z,t) = \mathrm{Re}\left(Ae^{i(kz-\omega t+\delta)}\right), \tag{13.11}$$

where $\mathrm{Re}\,(\cdots)$ takes the real part of a complex argument.

Extending to three dimensions, Eq. (13.1) generalizes to the 3D wave equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\psi(\boldsymbol{x}) = \mathscr{F}(\boldsymbol{x},t),, \tag{13.12}$$

which can be written in the compact form

$$\Box\,\psi(\boldsymbol{x}) = f, \tag{13.13}$$

by utilizing the d'Alembertian operator $\Box$ given in Eq. (11.25) and the Laplacian operator ∇^2 defined in Eq. (2.55).

$$\Box \equiv -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2 \qquad \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Let us now consider wave solutions of the Maxwell equations, first in vacuum, and then in medium.

13.2 Electromagnetic Waves in Vacuum

Electromagnetic waves are solutions of the Maxwell equations in the absence of sources. Removing the source terms ρ/ε_0 and $\mu_0 \boldsymbol{J}$ from the vacuum Maxwell equations (1.1) gives the *source-free Maxwell equations* in vacuum,

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = 0, \tag{13.14a}$$

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{13.14b}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0, \tag{13.14c}$$

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \frac{1}{c^2} \frac{\partial \boldsymbol{E}}{\partial t}. \tag{13.14d}$$

We may interpret these four equations in the following way:

1. The two equations (13.14b) and (13.14d) involving curls are coupled partial differential equations describing the time evolution of the electric field $\boldsymbol{E}(\boldsymbol{x},t)$ and magnetic field $\boldsymbol{B}(\boldsymbol{x},t)$, respectively.
2. The two equations (13.14a) and (13.14c) involving divergences provide initial conditions for the integration of (13.14b) and (13.14d), and we may deduce from them that

$$\frac{\partial}{\partial t}(\boldsymbol{\nabla} \cdot \boldsymbol{E}) = 0 \qquad \frac{\partial}{\partial t}(\boldsymbol{\nabla} \cdot \boldsymbol{B}) = 0. \tag{13.15}$$

   Thus if the fields are free of divergence initially because of conditions set by Eqs. (13.14a) and (13.14c), they will remain divergence-free as time evolves in the integration because of Eq. (13.15).

We shall now show that there are solutions of the source-free Maxwell equations in vacuum (13.14) that are vector versions of the classical wave solution for a scalar field. However, these wave solutions for the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields

1. are 3D vector fields, and
2. the vector $\boldsymbol{E}$ and $\boldsymbol{B}$ fields are not independent because they are coupled through one of the first-order Maxwell equations, for example the Faraday equation (13.14b).

Thus electromagnetic wave solutions may be expected to be much richer than those for a scalar wave equation.

13.2.1 Electromagnetic Wave Equations for the Fields

First, applying the curl operation to Eq. (13.14b) and using the vector identity of Eq. (A.8),

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}, \tag{13.16}$$

we obtain

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{E}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{E}) - \nabla^2\boldsymbol{E} = -\boldsymbol{\nabla}\times\left(\frac{\partial\boldsymbol{B}}{\partial t}\right)$$
$$= -\frac{\partial}{\partial t}(\boldsymbol{\nabla}\times\boldsymbol{B}) = -\frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2},$$

where Eq. (13.14d) was used in the last step. Thus,

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{E}) - \nabla^2\boldsymbol{E} = -\frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2},$$

But from the 1st vacuum Maxwell equation $\boldsymbol{\nabla}\cdot\boldsymbol{E} = 0$, and we obtain finally,

$$\nabla^2\boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2}, \tag{13.17}$$

Next apply the curl operation to Eq. (13.14d) and use the identity (13.16) to obtain in a similar manner

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{B}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{B}) - \nabla^2\boldsymbol{B} = \frac{1}{c^2}\boldsymbol{\nabla}\times\left(\frac{\partial\boldsymbol{E}}{\partial t}\right)$$
$$= \frac{1}{c^2}\frac{\partial}{\partial t}(\boldsymbol{\nabla}\times\boldsymbol{E}) = -\frac{1}{c^2}\frac{\partial^2\boldsymbol{B}}{\partial t^2},$$

where Eq. (13.14b) was used in the last step. But from the 3rd vacuum Maxwell equation, $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$, and we obtain

$$\nabla^2\boldsymbol{B} = \frac{1}{c^2}\frac{\partial^2\boldsymbol{B}}{\partial t^2}, \tag{13.18}$$

Finally, collecting the results in Eqs. (13.17) and (13.18), the wave solutions of the Maxwell equationa are expressed by

$$\nabla^2\boldsymbol{E} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2} = 0, \tag{13.19a}$$

$$\nabla^2\boldsymbol{B} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{B}}{\partial t^2} = 0, \tag{13.19b}$$

from which one sees that each cartesian component of $\boldsymbol{E}$ and $\boldsymbol{B}$ satisfies a scalar wave equation of the form (13.1) with no source term ($\mathscr{F} = 0$). Thus the coupled first-order Maxwell equations (13.14b) and (13.14d) have been decoupled in the vector wave equations (13.19), but at a price of the decoupled equations becoming second-order PDEs.[3]

Equations (13.19) indicate that each cartesian component of $\boldsymbol{E}(\boldsymbol{x},t)$ and $\boldsymbol{B}(\boldsymbol{x},t)$ satisfies a scalar wave equation of the form (13.1), but this is no longer true in spherical or cylindrical coordinates. Functions that happen to satisfy Eqs. (13.19) are not automatically guaranteed to be solutions of the Maxwell equations (13.19); they only describe valid electric and magnetic fields if they satisfy also the divergence conditions of Eqs. (13.14a) and (13.14c), and

[3] Note that the wave equations (13.19) require second derivatives of the fields, which appear only because of the displacement term $(1/c^2)\partial\boldsymbol{E}/\partial t$ that Maxwell added to Ampère's law to convert it into the Ampère–Maxwell law given in Eq. (13.14d). Thus one of the several fundamental consequences following from Maxwell's modification of Ampère's law that were discussed in Section 11.1.2 is the existence of electromagnetic wave solutions of the Maxwell equations.

satisfy Eqs. (13.14b) and (13.14d) when coupled together. These Maxwell-equation constraints are easy to demonstrate for simple plane-wave solutions of Eqs. (13.19), but more difficult for more complicated situations such as spherical waves or beam-like solutions. In such cases it may be simpler to use a set of wave equations based on the electromagnetic potentials rather than the fields, as illustrated in the next section.

13.2.2 Electromagnetic Wave Equations for the Potentials

An uncoupled description of electromagnetic waves can be obtained using the potentials rather than the fields.

Coulomb Gauge: For example, if we work in Coulomb gauge the gauge constraint $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$ of Eq. (8.37) for the vector potential $\boldsymbol{A}$ implies that $\nabla^2\Phi = 0$, which has a solution $\Phi = 0$, and the electric field in the presence of a vector potential is the electrostatic result given in Eq. (2.35) plus a time derivative of the vector potential:

$$\boldsymbol{E} = -\boldsymbol{\nabla}\Phi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{13.20}$$

Thus, substituting

$$\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A} \qquad \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{A}}{\partial t}.$$

into the Faraday law (13.14b) and applying $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$ once again gives

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = 0, \tag{13.21}$$

which can be solved for $\boldsymbol{A}$ and the electric and magnetic fields obtained by applying the $\partial/\partial t$ and curl operations to $\boldsymbol{A}$.

Lorenz Gauge: In a similar way, the Lorenz gauge condition of Eq. (11.11)

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t} = 0, \tag{13.22}$$

can be used to deduce the wave equations

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 0, \tag{13.23}$$

$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = 0, \tag{13.24}$$

for the scalar potential Φ and vector potential $\boldsymbol{A}$ in Lorenz gauge.

13.2.3 Monochromatic Plane-Wave Solutions

Comparing Eqs. (13.1) and (13.19), we may interpret the solutions of Eqs. (13.19) as waves in the electric and magnetic fields propagating at the speed of light c in vacuum, with the

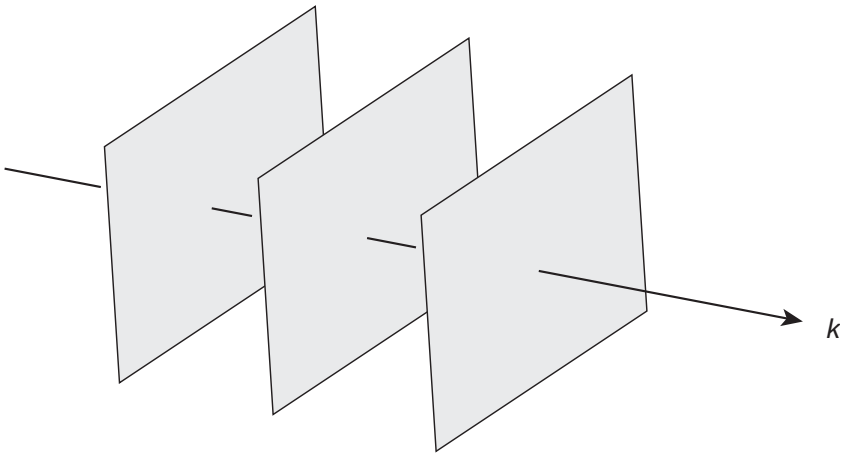


Fig. 13.2 For a plane wave the fields are constant on any plane transverse to the direction of propagation given by $\boldsymbol{k}$.

wavespeed c arising because of the relationship in Eq. (2.4) between the SI constants ε_0 and μ_0, and the speed of light,

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = 3.0 \times 10^8 \,\mathrm{m\,s^{-1}}. \tag{13.25}$$

The essential nature of these wave solutions may be illustrated by considering sinusoidal waves of a single angular frequency ω (monochromatic waves), since more complex waveforms correspond to a Fourier superposition of such waves. We assume these monochromatic and sinusoidal waves to propagate in the positive-z, direction, and assume them to be *plane-waves* (no x and y dependence); so-called, because the fields are uniform over a plane perpendicular to the direction of propagation,[4] as illustrated schematically in Fig. 13.2. For plane waves the wave equations for $\boldsymbol{E}$ and $\boldsymbol{B}$ can be solved directly relatively easily as in Section 13.2.1, so it isn't necessary to solve for the potentials first, as in Section 13.2.2.

Using the exponential notation introduced in Eq. (13.11), the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ corresponding to wave solutions of Eqs. (13.19) may be expressed in the general form,

$$\boldsymbol{E}(\boldsymbol{x},t) = \boldsymbol{\mathcal{E}}_0 e^{i(\boldsymbol{k}\cdot\boldsymbol{x}-\omega t)}, \tag{13.26a}$$

$$\boldsymbol{B}(\boldsymbol{x},t) = \boldsymbol{\mathcal{B}}_0 e^{i(\boldsymbol{k}\cdot\boldsymbol{x}-\omega t)}, \tag{13.26b}$$

where $\boldsymbol{\mathcal{E}}_0$ and $\boldsymbol{\mathcal{B}}_0$ are complex vectors [that have absorbed the phaseshift δ of Eq. (13.4)], and $\boldsymbol{k}$ is the *wavevector*, with

$$\boldsymbol{k} = \frac{\omega}{c}\hat{\boldsymbol{n}} \qquad \boldsymbol{\mathcal{E}}_0 \perp \hat{\boldsymbol{n}} \qquad \boldsymbol{\mathcal{B}}_0 = \hat{\boldsymbol{k}} \times \boldsymbol{\mathcal{E}}_0, \tag{13.27}$$

where $\hat{\boldsymbol{n}}$ is the direction of wave propagation. Generally every solution of the Maxwell equations satisfies Eqs. (13.26), but not every solution of Eqs. (13.26) is consistent with the Maxwell equations, which impose additional constraints. Taking the wave propagation

[4] Monochromatic plane waves are a fiction, since no physical wave can have infinite extent. However, they are a useful fiction because (1) they are simple mathematically, (2) realistic waves often resemble plane waves in a local-enough region, and (3) linear combinations of plane waves can approximate physical waves.

direction to be along the z-axis, from Eq. (13.19a) the wave equation for the z component of $\boldsymbol{E}$ is

$$\frac{\partial^2 E_z}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 E_z}{\partial t^2} = 0, \tag{13.28}$$

and the constraint $\boldsymbol{\nabla}\cdot\boldsymbol{E} = 0$ of Eq. (13.14a) imposes $\partial E_z/\partial t = 0$, so E_z is independent of z and the wave equation (13.28) reduces to

$$\frac{\partial^2 E_z}{\partial t^2} = 0. \tag{13.29}$$

Integrating gives $E_z = at + b$ where a and b are constants. The first term at can be discarded as unphysical because it grows in time without bound, giving $E_z =$ constant, which we choose to be zero. By a similar procedure, we can choose $B_z = 0$. Therefore,

$$E_z = B_z = 0. \tag{13.30}$$

In addition, Faraday's law requires a relationship between electric and magnetic fields,

$$\boldsymbol{B}(\boldsymbol{x},t) = \frac{1}{c}\,\hat{\boldsymbol{n}}\times\boldsymbol{E}(\boldsymbol{x},t). \tag{13.31}$$

where $\hat{\boldsymbol{n}}$ is the direction of wave propagation. In summary, for a plane wave traveling in the direction $\hat{\boldsymbol{n}}$,

$$\boldsymbol{E}\perp\hat{\boldsymbol{n}} \qquad \boldsymbol{B}\perp\hat{\boldsymbol{n}} \qquad \boldsymbol{B}(\boldsymbol{x},t) = \frac{1}{c}\,\hat{\boldsymbol{n}}\times\boldsymbol{E}(\boldsymbol{x},t). \tag{13.32}$$

These results indicate that

1. electromagetic waves are *transverse:* oscillations in the electric and magnetic fields are
   - perpendicular to the direction of wave propagation,
   - perpendicular to each other, and
   - in phase with each other.
2. The fields $\boldsymbol{B}$ scaled by a factor of c and $\boldsymbol{E}$ have the same magnitude at all points of spacetime, $B_0 = E_0/c$.

A transverse electromagnetic field propagating in the vacuum along the z-axis with the electric field oscillating in the $x-z$ plane is illustrated schematically in Fig. 13.3.

13.2.4 Energy and Momentum of Electromagnetic Waves

From Ch. 12, the general expression for the energy density u associated with an electromagnetic wave is

$$u = \frac{1}{2}\left(\varepsilon_0 E^2 + \frac{1}{\mu_0}B^2\right). \tag{13.33}$$

From Eq. (13.32), for monochromatic plane waves

$$B^2 = \frac{1}{c^2}E^2 = \mu_0\varepsilon_0 E^2, \tag{13.34}$$

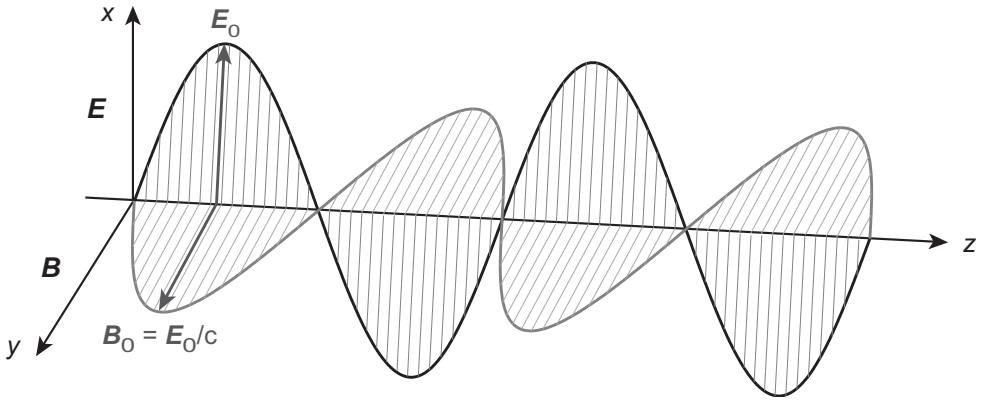


Fig. 13.3 An electromagnetic wave in vacuum propagating along the z axis. The electric field $\boldsymbol{E}$ oscillates sinusoidally in the $x-z$ plane, implying that the magnetic field $\boldsymbol{B}$ oscillates sinusoidally in the $y-z$ plane, in phase with the electric field, with the strength of the waves related by $cB_0 = E_0$ at each spacetime point.

and the contributions of the electric and magnetic fields to the energy density are equal,

$$\begin{aligned} u = \varepsilon_0 E^2 &= \varepsilon_0 \left(\mathrm{Re}\, E_0 e^{i(kz-\omega t+\delta)} \right)^2 \\ &= \varepsilon_0 E_0^2 \cos^2(kz-\omega t+\delta). \end{aligned} \tag{13.35}$$

The energy flux density associated with an electromagnetic wave (energy per unit area per unit time) is given by the Poynting vector

$$\boldsymbol{S} = \frac{1}{\mu_0}(\boldsymbol{E}\times\boldsymbol{B}), \tag{13.36}$$

and for a monochromatic plane wave propagating in the z-direction,

$$\boldsymbol{S} = c\varepsilon_0 E_0^2 \cos^2(kz-\omega t+\delta)\hat{\boldsymbol{z}} = cu\hat{\boldsymbol{z}}. \tag{13.37}$$

The corresponding momentum density is

$$\boldsymbol{g} = \frac{1}{c^2}\boldsymbol{S}, \tag{13.38}$$

and for monochromatic plane waves propagating in the $\hat{\boldsymbol{z}}$ direction,

$$\boldsymbol{g} = \frac{1}{c}\varepsilon_0 E_0^2 \cos^2(kz-\omega t+\delta)\hat{\boldsymbol{z}} = \frac{1}{c}u\hat{\boldsymbol{z}}. \tag{13.39}$$

Typically for electromagnetic waves the period is so short[5] that any macroscopic measurement will encompass many cycles and for energy and momentum densities we are interested in their average value over a cycle. The average of the rapidly oscillating cosine

[5] For example, in the visible spectrum the characteristic wavelength of yellow light is $\sim 6\times10^{-7}$ m, corresponding to a frequency of $\sim 5\times10^{-14}$ Hz.

squared function over one period is

$$\langle \cos^2(kz - 2\pi t/T + \delta)\rangle = \frac{1}{T}\int_0^T \cos^2(kz - 2\pi t/T + \delta)dt = \frac{1}{2}, \tag{13.40}$$

where $\langle\ \rangle$ denotes the average (over one cycle). Thus, the energy density, Poynting vector, and momentum density averaged over one cycle become

$$\langle u\rangle = \frac{1}{2}\varepsilon_0 E_0^2, \tag{13.41}$$

$$\langle \boldsymbol{S}\rangle = \frac{1}{2}c\varepsilon_0 E_0^2\,\hat{\boldsymbol{z}} \tag{13.42}$$

$$\langle \boldsymbol{g}\rangle = \frac{1}{2c}\varepsilon_0 E_0^2\,\hat{\boldsymbol{z}} \tag{13.43}$$

for a wave propagating in the $\hat{\boldsymbol{z}}$ direction.

13.3 Polarization of Electromagnetic Waves

The polarization of an electromagnetic wave characterizes the orientation of its field vectors over time. This polarization can be measured (see Section 13.3.5) and is often of considerable physical interest, since the polarization carries information about how the wave was formed, and about the environments that it has encountered while propagating from the source.[6] By convention, the polarization of electromagnetic waves is defined by the plane in which the electric field $\boldsymbol{E}$ is oscillating, as illustrated in Fig. 13.4.

13.3.1 The Polarization Ellipse

Confining ourselves to monochromatic plane electromagnetic waves propagating in vacuum, we shall now show that as the phase of the wave advances by 2π, the tip of the electric field vector traces out an ellipse in the plane perpendicular to the propagation direction $\boldsymbol{k}$ that is called the *polarization ellipse*. Let us define real unit vectors $\hat{\boldsymbol{\varepsilon}}_1$ and $\hat{\boldsymbol{\varepsilon}}_2$ such that $(\hat{\boldsymbol{\varepsilon}}_1, \hat{\boldsymbol{\varepsilon}}_2, \hat{\boldsymbol{k}})$ constitutes a right-handed orthogonal triad,

$$\hat{\boldsymbol{\varepsilon}}_1 \cdot \hat{\boldsymbol{\varepsilon}}_2 = 0 \qquad \hat{\boldsymbol{\varepsilon}}_1 \times \hat{\boldsymbol{\varepsilon}}_2 = \hat{\boldsymbol{k}}. \tag{13.44}$$

The vectors $(\hat{\boldsymbol{\varepsilon}}_1, \hat{\boldsymbol{\varepsilon}}_2)$ constitute a polarization basis for $\boldsymbol{\mathcal{E}}_0$ in Eq. (13.26a) and the electric field may be expressed as

$$\boldsymbol{E}(\boldsymbol{x}, t) = (\mathscr{E}_1\hat{\boldsymbol{\varepsilon}}_1 + \mathscr{E}_2\hat{\boldsymbol{\varepsilon}}_2)\, e^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)}, \tag{13.45}$$

[6] For example, the polarization of the cosmic microwave background (CMB) radiation that is observed coming from all directions in the sky by satellite observatories provides a key test for cosmological theories of the origin of the Universe since the CMB is thought to be the cooling remnant of the big bang itself. It has also been proposed that the CMB may carry a polarization imprint of perturbations due to gravitational waves produced in the big bang.

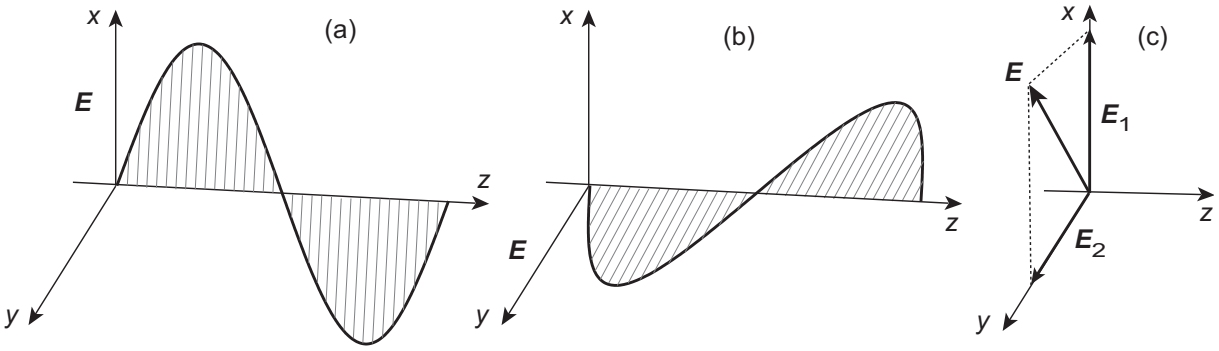


Fig. 13.4 Polarization of transverse electromagnetic wave propagating in the positive z direction. (a) "Vertical" polarization of electric field in the $x-z$ plane. (b) "Horizontal" polarization of electric field in the $y-z$ plane. These are examples of linear or plane polarization. (c) General polarization defined by plane of the electric field vibrations, which can be resolved into two mutually perpendicular components, $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$.

where the complex numbers $\mathscr{E}_1$ and $\mathscr{E}_2$ may be parameterized as

$$\mathscr{E}_1 = Ae^{i\delta_1} \qquad \mathscr{E}_2 = Be^{i\delta_2}, \tag{13.46}$$

in terms of the real numbers A, B, δ_1, and δ_2. Then

$$\mathrm{Re}\,\boldsymbol{E} = A\cos(\phi+\delta_1)\hat{\boldsymbol{\varepsilon}}_1 + B\cos(\phi+\delta_2)\hat{\boldsymbol{\varepsilon}}_2 \equiv E_1\hat{\boldsymbol{\varepsilon}}_1 + E_2\hat{\boldsymbol{\varepsilon}}_2, \tag{13.47}$$

where we have defined

$$\phi \equiv \boldsymbol{k}\cdot\boldsymbol{x} - \omega t, \tag{13.48}$$

and it may be shown (Problem 13.1) that Eqs. (13.45)-(13.48) imply that

$$\left(\frac{E_1}{A}\right)^2 + \left(\frac{E_2}{B}\right)^2 - 2\left(\frac{E_1}{A}\right)\left(\frac{E_2}{B}\right)\cos\delta = \sin^2\delta, \tag{13.49}$$

where we define

$$\delta \equiv \delta_2 - \delta_1. \tag{13.50}$$

Equation (13.49) specifies an ellipse in the $\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2$ plane, with the eccentricity and orientation of the ellipse determined by the parameters δ and the ratio B/A. All possible polarizations of a monochromatic, plane electromagnetic wave in vacuum are then described by the polarization ellipse (13.49) with arbitrary values of parameters.

There are some special cases of the general polarization ellipse of Eq. (13.49) corresponding to two limits of the eccentricity e of an ellipse (which lies between 0 and 1, by definition) that are of note. The first is that the limit of an ellipse as $e \to 1$ is a straight line. The second is that an ellipse becomes a circle in the limit that $e \to 0$. Let us now discuss these two special cases.

13.3.2 Linear (Plane) Polarization

The ellipse described by Eq. (13.49) becomes a straight line if the orthogonal electric field components have the same phase or are 180° out of phase, which corresponds to the restriction

$$\delta = \delta_2 - \delta_1 = m\pi \qquad (m = 0, 1, 2, \ldots). \tag{13.51}$$

This special case corresponds to *linear (or plane) polarization*, because the electric field vector for a linearly polarized electromagnetic wave points in the *same direction for all values of* $\boldsymbol{x}$ *and* t [see Problem 13.2(a)]. The waves in Figs. 13.4(a) and 13.4(b) are examples of such *plane-polarized* waves. Any plane-polarized wave can be decomposed into a sum of two components that are plane-polarized in mutually perpendicular directions and that have the same phase. In Fig. 13.4(c) $\boldsymbol{E}$ has been resolved into two mutually perpendicular x and y components, $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$, respectively, having the same phase.

13.3.3 Circular Polarization

The other special case of Eq. (13.49) of specific interest is that an ellipse becomes a circle in the limit that the eccentricity tends to zero. If two plane-polarized waves that differ in phase are added, the maxima of $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ occur at different times and the tip of the vector $\boldsymbol{E}$ describes an ellipse as the phase of the wave advances by 2π; such a wave is described by the general ellipse (13.49) with arbitrary parameter values and is said to have *elliptical polarization*. As we have seen, if the phases of the two components are made equivalent the ellipse degenerates to a straight line corresponding to plane polarization. However, if $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ have the same amplitudes but are out of phase by $\pm\frac{\pi}{2}$, the ellipse becomes a circle and the wave is said to be *circularly polarized*. If the orthogonal field components are 90° out of phase but of equal amplitude,

$$\mathcal{A} \equiv \sqrt{2}A = \sqrt{2}B \qquad \delta = \delta_2 - \delta_1 = \frac{m\pi}{2} \qquad (m = \pm 1, \pm 3, \ldots), \tag{13.52}$$

Then, as you are asked to show in Problem 13.2(b), the wave is described in a plane containing $\boldsymbol{x}$ by

$$\mathrm{Re}\,[\boldsymbol{E}_{\pm}(0,t)] = \frac{\mathcal{A}}{\sqrt{2}}(\hat{\boldsymbol{\varepsilon}}_1 \cos\omega t \pm \hat{\boldsymbol{\varepsilon}}_2 \sin\omega t). \tag{13.53}$$

As illustrated in Fig. 13.5, for the corresponding circularly polarized light the tip of the electric field vector executes a spiraling motion as the wave advances. Because of the $\pm$ ambiguity in Eq. (13.53), for a circularly polarized wave the tip of the electric field vector rotates in a circle as the wave propagates that can be clockwise or counterclockwise. For an *observer looking toward the source*, we define the wave to be

1. *right circularly polarized (RCP)* if the vector $\boldsymbol{E}$ rotates clockwise as the wave propagates, and
2. *left circularly polarized (LCP)* if it rotates counterclockwise.[7]

[7] In some subfields of physics such as optics a convention for LCP/RCP may be employed that reverses these

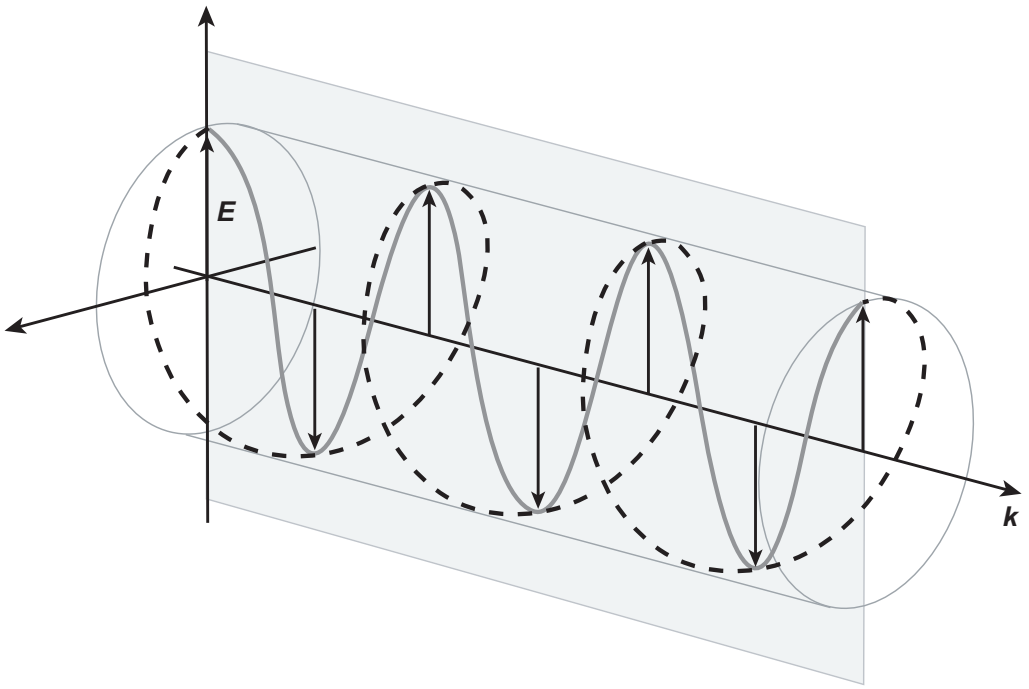


Fig. 13.5

Snapshot at fixed time of $\boldsymbol{E}(x,0)$ for a monochromatic, circularly-polarized, plane wave propagating in the direction given by $\boldsymbol{k}$. The magnitude of the electric field is indicated by the solid curve. The tip of the $\boldsymbol{E}$ vector executes a spiral motion around the $\boldsymbol{k}$ axis as the wave propagates, as indicated by the dashed curve. This wave is left circularly polarized (LCP), since the sense of rotation is counterclockwise, as viewed in the direction of the source.

The wave in Fig. 13.5 is thus left circularly polarized (LCP), since for an observer looking toward the source $\boldsymbol{E}$ appears to be rotating counterclockwise. The corresponding RCP wave looks the same as Fig. 13.5 except that the sense of the spiral would be opposite because it would appear to be rotating clockwise for the same observer.

13.3.4 Circular Polarization Basis

It can be useful to combine the cartesian basis vectors $\hat{\boldsymbol{\varepsilon}}_1$ and $\hat{\boldsymbol{\varepsilon}}_2$ into the complex conjugate pairs

$$\hat{\boldsymbol{\varepsilon}}_+ = \frac{1}{\sqrt{2}}(\hat{\boldsymbol{\varepsilon}}_1 + i\hat{\boldsymbol{\varepsilon}}_2) \qquad \hat{\boldsymbol{\varepsilon}}_- = \frac{1}{\sqrt{2}}(\hat{\boldsymbol{\varepsilon}}_1 - i\hat{\boldsymbol{\varepsilon}}_2), \tag{13.54}$$

which satisfy the relations

$$\hat{\boldsymbol{\varepsilon}}_\pm^* \hat{\boldsymbol{\varepsilon}}_\pm = 1 \qquad \hat{\boldsymbol{\varepsilon}}_\pm^* \cdot \hat{\boldsymbol{\varepsilon}}_\mp = 0.$$

definitions. Generally, in discussing LCP/RCP one must specify whether one is defining the sense of rotation by looking upstream at the source, or downstream toward the destination of the wave, since these points of view reverse the definitions. Here we define LCP/RCP by looking upstream at the source, with RCP corresponding to clockwise rotation and LCP to counterclockwise rotation of $\boldsymbol{E}$.

Equation (13.53) is the real part of the complex field (evaluated at $\boldsymbol{x} = 0$),

$$\boldsymbol{E}_{\pm}(\boldsymbol{x},t) = \mathcal{A}\left(\frac{\hat{\boldsymbol{\varepsilon}}_1 \pm i\hat{\boldsymbol{\varepsilon}}_2}{\sqrt{2}}\right) e^{i(\boldsymbol{k}\cdot\boldsymbol{x}-\omega t)}, \tag{13.55}$$

so $\hat{\boldsymbol{\varepsilon}}_+$ is a pure LCP wave and $\hat{\boldsymbol{\varepsilon}}_-$ is a pure RCP wave. Thus the expansion

$$\boldsymbol{E}(\boldsymbol{x},t) = (\mathscr{E}_+\hat{\boldsymbol{\varepsilon}}_+ + \mathscr{E}_-\hat{\boldsymbol{\varepsilon}}_-)\, e^{i(\boldsymbol{k}\cdot\boldsymbol{x}-\omega t)}, \tag{13.56}$$

is just as legitimate as the expansion (13.45), and either of the orthogonal bases $(\hat{\boldsymbol{\varepsilon}}_1, \hat{\boldsymbol{\varepsilon}}_2)$ or $(\hat{\boldsymbol{\varepsilon}}_+, \hat{\boldsymbol{\varepsilon}}_-)$ may be used to expand monochromatic plane waves in states of definite linear or circular polarization.

13.3.5 Measuring Polarization: Stokes Parameters

At low frequencies the polarization parameters A, B, and δ defining the polarization ellipse (13.49) can be extracted rather directly from data. At optical and higher frequencies it is usual to instead extract the polarization parameters from the *Stokes parameters* s_i, which are defined as follows.

$$\begin{aligned} s_0 &= A^2 + B^2 \qquad & s_1 &= A^2 - B^2, \\ s_2 &= 2AB\cos\delta \qquad & s_3 &= 2AB\sin\delta, \end{aligned} \tag{13.57}$$

with the constraint $s_0^2 = s_1^2 + s_2^2 + s_3^2$, so only three of the s_i are independent. The utility of the Stokes parameters is that each parameter can be related directly to wave intensities measured in different orthogonal bases. Consider the $\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2$ coordinate system used in Eq. (13.45).

1. Let I_0 and I_{90} be intensities measured in detectors sensitive to horizontal and vertical linear polarizations, respectively, for the $\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2$ coordinate system.
2. Let I_{+45} and I_{-45} be intensities measured in detectors sensitive to linear polarizations measured at angles ± 45 degrees, respectively, relative to the $\hat{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}_2$ axes.
3. Let $I_{\rm RCP}$ and $I_{\rm LCP}$ be intensities measured by detectors sensitive to right and left circular polarizations, respectively.

Then, as you are asked to prove in Problem 13.3,

$$\begin{aligned} s_0 &= I_0 + I_{90} \qquad & s_1 &= I_0 - I_{90}, \\ s_2 &= I_{+45} - I_{-45} \qquad & s_3 &= I_{\rm RCP} - I_{\rm LCP}, \end{aligned} \tag{13.58}$$

which shows that *six intensity measurements* are sufficient to determine the amplitudes and phases of the complex electric field components E_1 and E_2 appearing in Eqs. (13.44) - (13.49).

13.3.6 The Poincaré Sphere

The Stokes parameters define a one-to-one relationship between each possible polarization state and a unique point on the surface of a 2-sphere of radius s_0 called the *Poincaré sphere*, which is illustrated in Fig. 13.6. The Stokes parameters map the polarization ellipse

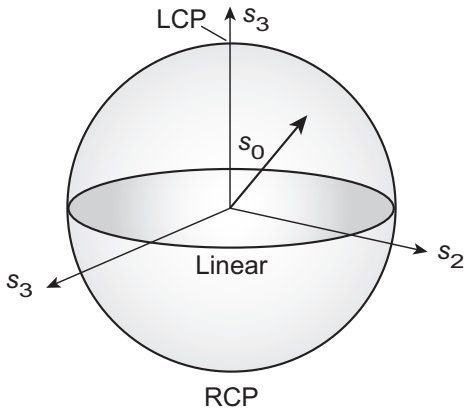


Fig. 13.6 A Poincaré sphere of radius s_0 that is in one-to-one correspondence with values of the Stokes parameters defined in Eqs. (13.57) and (13.58).

parameters A, B, and δ into a unique point on the Poincaré sphere. All states of linear polarization lie on the equator, RCP states lie at the south pole, LCP states lie at the north pole, and all other points correspond to general elliptical polarization in Fig. 13.6 [42].

13.4 Electromagnetic Waves in Simple Matter

Let us now consider the propagation of electromagnetic waves in *simple matter*, where we may assume that the electrical permittivity ε, magnetic permeability μ, and ohmic conductivity σ are all constant. Realistic matter is *frequency dispersive*, meaning that the permitivitty ε is a function of frequency $\varepsilon = \varepsilon(\omega)$, and plane waves propagate with different phase velocities if they have different frequencies. Nevertheless, it is useful to first consider simple matter, where if $\sigma = 0$ plane waves are *non-dispersive* and all frequencies travel at the same phase velocity, because many essential features of real matter are captured in a non-dispersive model. This is particularly true with respect to reflection, refraction, and interference effects occurring at the boundaries separating different types of matter.

Wave propagation in linear, isotropic, non-dispersive matter looks almost like propagation in vacuum. The Maxwell equations are

$$\nabla \cdot \boldsymbol{D} = \rho_{\mathrm{f}}, \tag{13.59a}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{13.59b}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{13.59c}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}_{\mathrm{f}} + \frac{\partial \boldsymbol{D}}{\partial t}. \tag{13.59d}$$

For the present discussion we assume that there is no free charge $\rho_{\mathrm{f}} = 0$ and no free current

$J_f = 0$, and furthermore assume the medium is linear and homogeneous,

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} \qquad \boldsymbol{B} = \mu \boldsymbol{H}. \tag{13.60}$$

The constituitive relations (13.60) imply that there are only two independent fields, which we choose to be $\boldsymbol{E}$ and $\boldsymbol{H}$, with the polarization and magnetization of the medium then given by

$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E} \qquad \boldsymbol{M} = \left(\frac{\mu}{\mu_0} - 1\right)\boldsymbol{H}, \tag{13.61}$$

and the source-free Maxwell equations in terms of $\boldsymbol{E}$ and $\boldsymbol{H}$ are

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = 0 \qquad \boldsymbol{\nabla} \cdot \boldsymbol{H} = 0, \tag{13.62a}$$

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\mu \frac{\partial \boldsymbol{H}}{\partial t} \qquad \boldsymbol{\nabla} \times \boldsymbol{H} = \varepsilon \frac{\partial \boldsymbol{E}}{\partial t}. \tag{13.62b}$$

As in the discussion of waves in vacuum in Section 13.2.3, we assume monochromatic plane waves,

$$\boldsymbol{E}(\boldsymbol{x},t) = \boldsymbol{E} e^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)} \qquad \boldsymbol{H}(\boldsymbol{x},t) = \boldsymbol{H} e^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)}, \tag{13.63}$$

which produces four constraints upon substitution in Eqs. (13.62),

$$\boldsymbol{k} \cdot \boldsymbol{E} \qquad \boldsymbol{k} \cdot \boldsymbol{H} = 0, \tag{13.64a}$$

$$\boldsymbol{k} \times \boldsymbol{E} = \omega \mu \boldsymbol{H} \qquad \boldsymbol{k} \times \boldsymbol{H} = -\omega \varepsilon \boldsymbol{E}. \tag{13.64b}$$

These imply that Eq. (13.63) corresponds to a transverse electromagnetic wave, so our discussion of polarization in Section 13.3 applies without modification. From Problem 13.4, the wave frequency ω is related to the wave vector $\boldsymbol{k}$ by a *dispersion relation*,

$$\omega(\boldsymbol{k}) = \frac{c}{n} k, \tag{13.65}$$

where c is the speed of light and

$$n = c\sqrt{\mu\varepsilon}. \tag{13.66}$$

is the *index of refraction*.

13.4.1 Reflection and Refraction at Interfaces

From our general experience with waves, we expect that a plane electromagnetic wave incident on a sharp boundary between two different dielectric materials produces a reflected and refracted (transmitted) wave if the wavelength is small compared with the curvature of the interface; Fig. 13.7 illustrates. In this figure $z = 0$ is the boundary between Medium 1 characterized by parameters (ε_1, μ_1) and Medium 2 characterized by parameters (ε_1, μ_1). We may expect that at the boundary the incident wave splits into a reflected wave propagating in Medium 1 and a refracted wave propagating in Medium 2. Therefore, the electric fields in the two media are given by

$$\boldsymbol{E}_2(\boldsymbol{x},t) = E_T e^{i(\boldsymbol{k}_T\cdot\boldsymbol{x} - \omega_T t)} \qquad \text{(Medium 1)}, \tag{13.67a}$$

$$\boldsymbol{E}_1(\boldsymbol{x},t) = E_I e^{i(\boldsymbol{k}_I\cdot\boldsymbol{x} - \omega_I t)} + E_R e^{i(\boldsymbol{k}_R\cdot\boldsymbol{x} - \omega_R t)} \qquad \text{(Medium 2)}. \tag{13.67b}$$

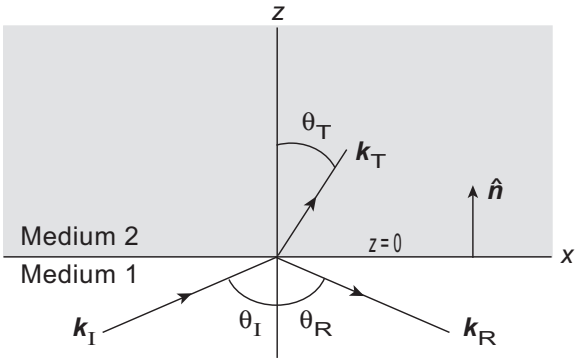


Fig. 13.7 Propagation of an monochromatic electromagnetic plane wave at a straight boundary between two dielectric materials, with $\hat{\boldsymbol{n}}$ the unit normal at the boundary. The wavevector is $\boldsymbol{k}_{\rm I}$ for the incident wave, $\boldsymbol{k}_{\rm R}$ for the reflected wave, and $\boldsymbol{k}_{\rm T}$ for the transmitted (refracted) wave.

At the interface between Medium 1 and Medium 2, the matching conditions in the absence of charge or current are

$$\hat{\boldsymbol{n}}\cdot(\boldsymbol{D}_1-\boldsymbol{D}_2)=0 \qquad \hat{\boldsymbol{n}}\cdot(\boldsymbol{B}_1-\boldsymbol{B}_2)=0, \tag{13.68a}$$

$$\hat{\boldsymbol{n}}\times(\boldsymbol{E}_1-\boldsymbol{E}_2)=0 \qquad \hat{\boldsymbol{n}}\times(\boldsymbol{H}_1-\boldsymbol{H}_2)=0, \tag{13.68b}$$

which may be used to determine the properties of the transmitted and reflected waves, as we shall now discuss.

The matching conditions (13.68) cannot be satisfied at all points and all times on the $z=0$ plane unless the phases of the incident, transmitted, and reflected waves are the same, which requires that

$$\omega_{\rm I}=\omega_{\rm T}=\omega_{\rm R}\equiv\omega \qquad \boldsymbol{k}_{\rm I}\cdot\boldsymbol{x}|_{z=0}=\boldsymbol{k}_{\rm T}\cdot\boldsymbol{x}|_{z=0}=\boldsymbol{k}_{\rm R}\cdot\boldsymbol{x}|_{z=0}\,, \tag{13.69}$$

and implies the conditions

$$k_{\rm Ix}=k_{\rm Tx}=k_{\rm Rx}, \tag{13.70a}$$

$$k_{\rm Iy}=k_{\rm Ty}=k_{\rm Ry}. \tag{13.70b}$$

In Fig. 13.7, $\boldsymbol{k}_{\rm I}$ and $\hat{\boldsymbol{n}}$ lie in the $x-z$ plane, so $k_{\rm Iy}=0$ and thus from Eq. (13.70b) $k_{\rm Ty}=k_{\rm Ry}=0$ and we deduce that the three wavevectors $\boldsymbol{k}_{\rm I}$, $\boldsymbol{k}_{\rm T}$, and $\boldsymbol{k}_{\rm R}$ are all coplanar.

The remainder of this section is under construction.

13.4.2 Absorption and Dispersion

This section is under construction.

Background and Further Reading

The discussion of polarization for electromagnetic waves in this chapter has followed Zangwill [42] closely.

Problems

13.1 Beginning from Eqs. (13.45)-(13.48), prove that the most general polarization of a monochromatic plane electromagnetic wave propagating in the vacuum is described by the polarization ellipse of Eq. (13.49).

13.2 (a) Show that for linearly (plane) polarized monochromatic plane electromagnetic waves the electric field vector points in the same fixed direction for all values of the coordinates $\boldsymbol{x}$ and t.

(b) Show that for circularly polarized monochromatic plane electromagnetic waves the electric field vector satisfies

$$\mathrm{Re}\,[\boldsymbol{E}_{\pm}(0,t)] = \frac{\mathcal{A}}{\sqrt{2}}(\hat{\boldsymbol{\varepsilon}}_1 \cos\omega t \pm \hat{\boldsymbol{\varepsilon}}_2 \sin\omega t),$$

where for circular polarization,

$$\mathcal{A} \equiv \sqrt{2}A = \sqrt{2}B \qquad \delta = \delta_2 - \delta_1 = \frac{m\pi}{2} \qquad (m = \pm 1, \pm 3, \ldots),$$

which corresponds to the spiral motion illustrated in Fig. 13.5.

13.3 Prove the relationships between measured intensities and the Stokes parameters given in Eq. (13.58).

13.4 Prove the dispersion relation (13.65). *Hint*: Take the curl of the relation $\boldsymbol{k} \times \boldsymbol{E} = \omega\mu\boldsymbol{H}$ given in Eq. (13.64b).

14 Radiation

This section is under construction.

Problems

14.1 *No problems yet for this chapter.*

15 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 theory (except for a few quantum concepts required for discussions 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. 16 we will introduce the special theory of relativity and demonstrate explicitly the Lorentz invariance of Maxwell's equations. Finally, Ch. 18 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. At the level of expertise required for our discussion 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 the discussion with an overview of 4-dimensional spacetime.

15.1 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.

15.1.1 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

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad t' = t, \tag{15.1}$$

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 correspond to 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 electromagnetic waves 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]

15.1.2 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 first concluded that the Maxwell theory was correct and that it was the *Galilean transformations* that required modification to reconcile mechanics and electromagnetism. Thus he put forth the bold and unqualified assertions that

1. physical law is the same in all coordinate systems, so there are no preferred inertial frames, and
2. the speed of light is *constant in all inertial frames*

that are the foundation of the 1905 *special theory of relativity*. Requiring the speed of light c be an invariant independent of inertial frame dictated replacement of Galilean transformations (15.1) with *Lorentz transformations,*

$$\begin{aligned} x' &= \gamma(x - vt) \qquad y' = y \qquad z' = z, \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right) \qquad \gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}}, \end{aligned} \tag{15.2}$$

where a boost along the x axis is assumed and 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 (15.2) become equivalent to the Galilean transformations (15.1). 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*

$$(x^0, x^1, x^2, x^3) \equiv (ct, x, y, z), \tag{15.3}$$

[1] Nevertheless, 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.

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:

> *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 rather simple in 4-dimensional spacetime, but becomes complicated when projected onto 3-dimensional space.[2]

15.1.3 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

$$\begin{aligned} ds^2 = \sum_{\mu\nu} \eta_{\mu\nu} dx^\mu dx^\nu &= -c^2 dt^2 + dx^2 + dy^2 + dz^2, \\ &= -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2, \end{aligned} \tag{15.4}$$

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. (15.4) may be expressed as the diagonal matrix

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{15.5}$$

The line element (15.4) 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. (15.4), or on the diagonal of the matrix in Eq. (15.5), defines the *signature of the metric.* For Minkowski space the

[2] An English translation of the full presentation may be found at `https://en.wikisource.org/wiki/Translation:Space_and_Time`. Minkowski undoubtedly would have made further contributions to the development of special and general relativity, but he died unexpectedly of peritonitis only months after his famous Space and Time lecture.

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 (15.5) 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 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", ...) follow directly from this difference in geometry relative to euclidean space implied by the Minkowski indefinite metric.

15.2 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.

15.3 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 coor-

dinates and one timelike coordinate. However, to introduce these concepts it is useful to begin with the simpler case of vector fields in euclidean space [10].

15.3.1 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

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w). \tag{15.6}$$

Further, we assume the transformations to be invertible so that we can solve for (u,v,w) in terms of (x,y,z).

Example 15.1 Let (u,v,w) correspond to spherical coordinates (r,θ,ϕ), so that Eq. (15.6) takes the familiar form

$$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\cos\theta, \tag{15.7}$$

with $r \geq 0$ and $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$.

It will prove useful to combine Eqs. (15.6) into an equation giving a position vector $\boldsymbol{r}$ for a point in terms of the (u,v,w) coordinates:

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k}. \tag{15.8}$$

For example, in terms of the spherical coordinates (r,θ,ϕ),

$$\boldsymbol{r} = (r\sin\theta\cos\phi)\,\boldsymbol{i} + (r\sin\theta\sin\phi)\,\boldsymbol{j} + (r\cos\theta)\,\boldsymbol{k}. \tag{15.9}$$

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.

15.3.2 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 (15.8) and (15.9) 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.

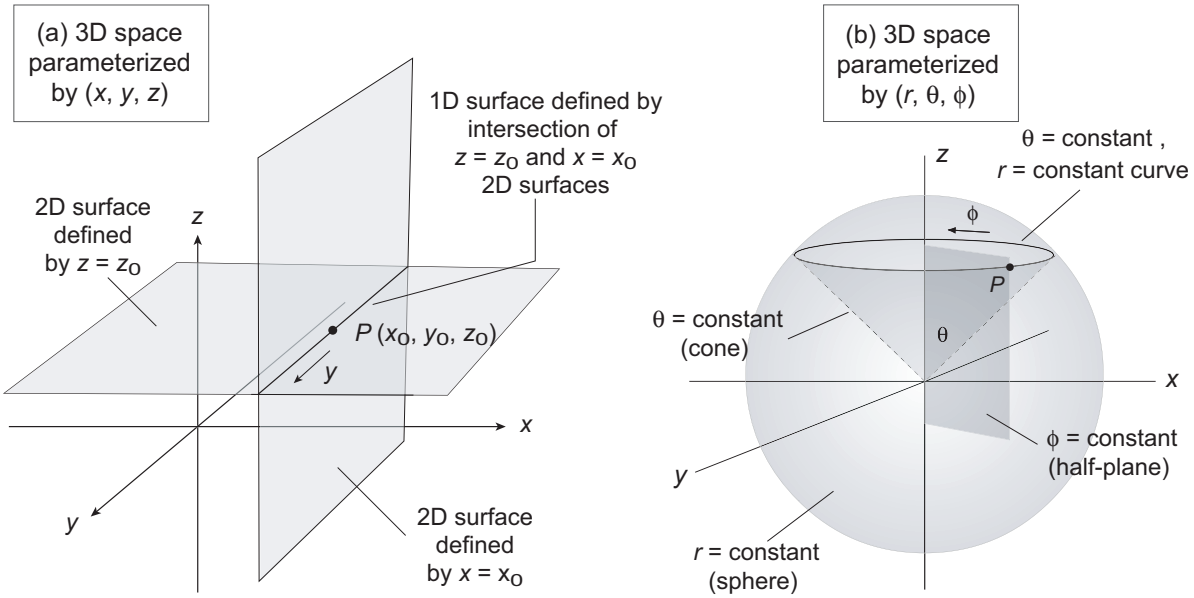


Fig. 15.1 Examples of surfaces and curves arising from constraints. (a) In 3D euclidean space parameterized by cartesian coordinates (x,y,z), the constraints $x = x_0$ and $z = z_0$ define 2D planes and the intersection of these planes defines a 1D surface parameterized by the variable y. (b) In 3D space described in spherical coordinates (r,θ,ϕ), the constraint $r =$ constant defines a 2D sphere, the constraint $\theta =$ constant defines a cone, and the constraint $\phi =$ constant defines a half-plane. The intersection of any two of these surfaces defines a curve parameterized by the variable not being held constant.

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 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. (15.8) 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 $w = w_0$,

$$\boldsymbol{r}(u) = x(u,v_0,w_0)\,\boldsymbol{i} + y(u,v_0,w_0)\,\boldsymbol{j} + z(u,v_0,w_0)\,\boldsymbol{k}, \tag{15.10}$$

where u acts as a coordinate along the resulting curve. Figure 15.1 illustrates for a space parameterized by cartesian and spherical coordinate systems. For example, in Fig. 15.1(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 $\phi =$ 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 (15.8) 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

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w}, \tag{15.11}$$

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 15.2 illustrates construction of the tangent basis for the example of a spherical coordinate system.

Example 15.2 Consider the spherical coordinate system defined in Eq. (15.7) and illustrated in Fig. 15.1(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. (15.11) 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. 15.2.

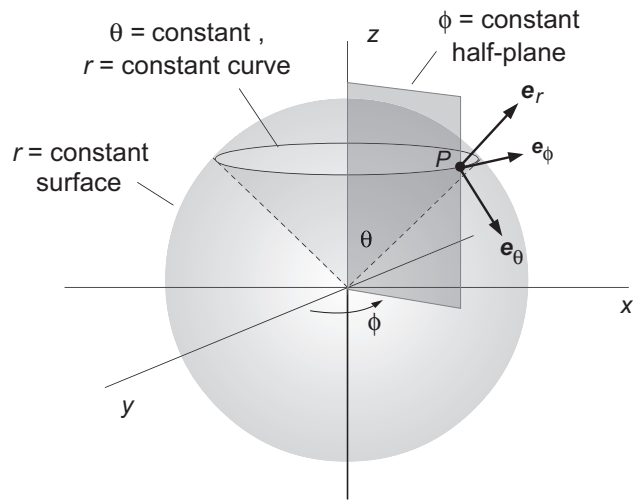


Fig. 15.2 Unit vectors in the tangent basis at point P for Example 15.2. The three basis vectors are tangents to the curves passing through P that are defined by setting any two of the three variables to a constant.

In elementary physics it is common to use an orthogonal coordinate system so that the basis 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. (15.11) 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

$$\begin{aligned} \boldsymbol{e}^u &\equiv \boldsymbol{\nabla} u = \frac{\partial u}{\partial x}\boldsymbol{i} + \frac{\partial u}{\partial y}\boldsymbol{j} + \frac{\partial u}{\partial z}\boldsymbol{k}, \\ \boldsymbol{e}^v &\equiv \boldsymbol{\nabla} v = \frac{\partial v}{\partial x}\boldsymbol{i} + \frac{\partial v}{\partial y}\boldsymbol{j} + \frac{\partial v}{\partial z}\boldsymbol{k}, \\ \boldsymbol{e}^w &\equiv \boldsymbol{\nabla} w = \frac{\partial w}{\partial x}\boldsymbol{i} + \frac{\partial w}{\partial y}\boldsymbol{j} + \frac{\partial w}{\partial z}\boldsymbol{k}, \end{aligned} \tag{15.12}$$

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 (15.11), 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 (15.12) and *subscript indices* on the basis vectors (15.11).

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 15.3 illustrates.

Example 15.3 Define a coordinate system (u,v,w) in terms of cartesian coordinates (x,y,z) through [10]

$$x = u+v \qquad y = u-v \qquad z = 2uv+w.$$

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. (15.11) 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. (15.12) 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. (15.11) and (15.12) 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.

15.3.3 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 $\{\boldsymbol{e}^i\}$:

$$\boldsymbol{V} = V^1\boldsymbol{e}_1 + V^2\boldsymbol{e}_2 + V^3\boldsymbol{e}_3 = \sum_i V^i\boldsymbol{e}_i \equiv V^i\boldsymbol{e}_i, \tag{15.13}$$

$$\boldsymbol{\omega} = \omega_1\boldsymbol{e}^1 + \omega_2\boldsymbol{e}^2 + \omega_3\boldsymbol{e}^3 = \sum_i \omega_i\boldsymbol{e}^i \equiv \omega_i\boldsymbol{e}^i, \tag{15.14}$$

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. (15.13) 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. (15.14) 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.

15.3.4 Vector Scalar Product and the Metric Tensor

Utilizing Eq. (15.13), the scalar product of two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ may be expressed as

$$\boldsymbol{A}\cdot\boldsymbol{B} = (A^i\boldsymbol{e}_i)\cdot(B^j\boldsymbol{e}_j) = \boldsymbol{e}_i\cdot\boldsymbol{e}_j A^iB^j = g_{ij}A^iB^j, \tag{15.15}$$

where the components of the *metric tensor* g_{ij} in this basis are given by

$$g_{ij} \equiv \boldsymbol{e}_i\cdot\boldsymbol{e}_j. \tag{15.16}$$

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. (15.15) illustrates. Equivalently, the scalar product of dual vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may be expressed as

$$\boldsymbol{\alpha}\cdot\boldsymbol{\beta} = \alpha_i\boldsymbol{e}^i\cdot\beta_j\boldsymbol{e}^j = g^{ij}\alpha_i\beta_j, \tag{15.17}$$

where the metric tensor components g^{ij} with two upper indices are defined by

$$g^{ij} \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}^j, \tag{15.18}$$

and the scalar product of dual vectors and vectors as

$$\boldsymbol{\alpha} \cdot \boldsymbol{B} = \alpha_i \boldsymbol{e}^i \cdot B^j \boldsymbol{e}_j = g^i_j \alpha_i B^j, \tag{15.19}$$

where the metric tensor components g^i_j with mixed upper and lower indices are defined by

$$g^i_j \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}_j. \tag{15.20}$$

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.

15.3.5 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 15.4.6, but we summarize here the salient points.

1. Vectors are not defined directly in the manifold, but instead are defined in a euclidean vector space (see Box 15.3) 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 15.3) 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

$$\boldsymbol{\omega}(\boldsymbol{V}) = \boldsymbol{V}(\boldsymbol{\omega}) = V^i\omega_i \in \mathbb{R}. \tag{15.21}$$

Notice further that the expression $\boldsymbol{A}\cdot\boldsymbol{B} = g_{ij}A^iB^j$ from Eq. (15.15) 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

$$\boldsymbol{A}(\boldsymbol{B}) = \boldsymbol{A}\cdot\boldsymbol{B} \equiv A_iB^i = g_{ij}A^iB^j. \tag{15.22}$$

But since in $A_iB^i = g_{ij}A^iB^j$ the vector B is arbitrary, in general

$$A_i = g_{ij}A^j, \tag{15.23}$$

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. (15.23) can be inverted using that the inverse of g_{ij} is g^{ij} (see Section 15.3.6) to give

$$A^i = g^{ij}A_j. \tag{15.24}$$

Hence, using the metric tensor to raise and lower indices by summing over a repeated index (an operation called *contraction*) as in Eqs. (15.23) and (15.24), we see that the vector and dual vector components are related through contraction with the metric tensor.[4]

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. (15.13) and (15.14)

$$\boldsymbol{e}^i(\boldsymbol{e}_j) = \boldsymbol{e}^i\cdot\boldsymbol{e}_j = \delta^i_j, \tag{15.25}$$

[4] A metric is a function for measuring distance between points in a manifold. Mathematically, a respectable manifold need not have a metric defined, and the absence of a metric would complicate the present discussion. However, these lectures are about physics, and physics cannot get very far without the notion of measuring distance between points, so it is typically formulated in manifolds that do have a metric defined (such manifolds are termed *metric spaces*). Thus, we will always assume a metric to be at our disposal in these discussions.

where the Kronecker delta is defined by

$$\delta^i_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} . \tag{15.26}$$

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,

$$V^i = \boldsymbol{e}^i \cdot \boldsymbol{V} \qquad V_i = \boldsymbol{e}_i \cdot \boldsymbol{V}. \tag{15.27}$$

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. (15.21)–(15.27) 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.

15.3.6 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:

$$g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}. \tag{15.28}$$

From Eqs. (15.23) and (15.24)

$$g_{ij}A^j = A_i \qquad g^{ij}A_j = A^i. \tag{15.29}$$

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

$$\boldsymbol{A} \cdot \boldsymbol{B} = g_{ij}A^iB^j = g^{ij}A_iB_j = g^i_jA_iB^j = A^iB_i = A_iB^i. \tag{15.30}$$

From the two expressions in (15.29), $A^i = g^{ij}A_j = g^{ij}g_{jk}A^k$, and since this is valid for arbitrary components A^i it follows that the metric tensor obeys

$$g^{ij}g_{jk} = g_{kj}g^{ji} = \delta^i_k. \tag{15.31}$$

Viewing g^{ij} as the elements of a matrix G and g_{ij} as the elements of a matrix $\tilde{G}$, Eqs. (15.28) are equivalent to the matrix equations

$$G = G^{\mathrm{T}} \qquad \tilde{G} = \tilde{G}^{\mathrm{T}}, \tag{15.32}$$

Box 15.1 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: $\boldsymbol{A} \cdot \boldsymbol{B} = g_{ij} A^i B^j = g^{ij} A_i B_j = g^i_j A_i B^j = A^i B_i = A_i B^i$

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 : $\tilde{G} G = G \tilde{G} = I \qquad G \equiv [g^{ij}] \qquad \tilde{G} \equiv [g_{ij}]$

where T denotes the transpose of the matrix. The Kronecker delta is just the unit matrix I, implying that Eq. (15.31) may be written as the matrix equations

$$\tilde{G} G = G \tilde{G} = I. \tag{15.33}$$

Therefore, we note the useful property that

> 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 15.1.

15.3.7 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. (15.8), 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

$$\frac{d\boldsymbol{r}}{dt} = \frac{\partial \boldsymbol{r}}{\partial u^1}\frac{du^1}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^2}\frac{du^2}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^3}\frac{du^3}{dt} = \dot{u}^1 \boldsymbol{e}_1 + \dot{u}^2 \boldsymbol{e}_2 + \dot{u}^3 \boldsymbol{e}_3, \tag{15.34}$$

where (15.11) was used and $du^i/dt \equiv \dot{u}^i$. The squared infinitesimal distance along the curve is then given by

$$\begin{aligned} ds^2 &= d\boldsymbol{r} \cdot d\boldsymbol{r} = du^i \boldsymbol{e}_i \cdot du^j \boldsymbol{e}_j \\ &= \boldsymbol{e}_i \cdot \boldsymbol{e}_j du^i du^j \\ &= g_{ij} du^i du^j, \end{aligned} \tag{15.35}$$

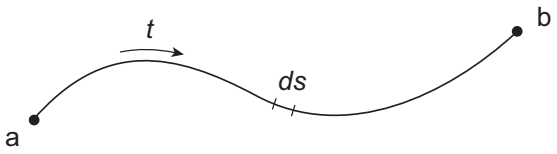


Fig. 15.3 Distance ds between points a and b along a curve parameterized by t.

where Eq. (15.16) was used. [Notice the notational convention $d\alpha^2 \equiv (d\alpha)^2$.] Thus $ds^2 = g_{ij}du^i du^j$ is the infinitesimal line element implied by the metric g_{ij}, and the length d of a finite segment of the curve between points a and b is obtained integration,

$$d = \int_a^b \left(g_{ij} \frac{du^i}{dt} \frac{du^j}{dt} \right)^{1/2} dt, \tag{15.36}$$

where t parameterizes the position along the curve, as illustrated in Fig. 15.3.

15.3.8 Euclidean Line Element

The line element for 2D euclidean space in cartesian coordinates (x,y) is

$$ds^2 = dx^2 + dy^2, \tag{15.37}$$

which is just the Pythagorean theorem in differential form. The corresponding line element in plane polar coordinates (r,ϕ) is then

$$ds^2 = dr^2 + r^2 d\phi^2, \tag{15.38}$$

as worked out in Example 15.4.

Example 15.4 For plane polar coordinates (r,ϕ)

$$x = r\cos\phi \qquad y = r\sin\phi,$$

so the position vector (15.8) is

$$\boldsymbol{r} = (r\cos\phi)\,\boldsymbol{i} + (r\sin\phi)\,\boldsymbol{j}.$$

Then from Eq. (15.11) 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. (15.16),

$$g_{11} = \cos^2\phi + \sin^2\phi = 1 \qquad g_{22} = r^2(\cos^2\phi + \sin^2\phi) = r^2$$

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. (15.35),

$$ds^2 = g_{11}(du^1)^2 + g_{22}(du^2)^2 = dr^2 + r^2 d\phi^2,$$

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.*[5] 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 (15.37) or (15.38).

15.3.9 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

$$dA = \sqrt{\det g}\, dx^1 dx^2, \tag{15.39}$$

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 $\boldsymbol{e}_i$,

$$\boldsymbol{V} = V^i \boldsymbol{e}_i. \tag{15.40}$$

[5] 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.

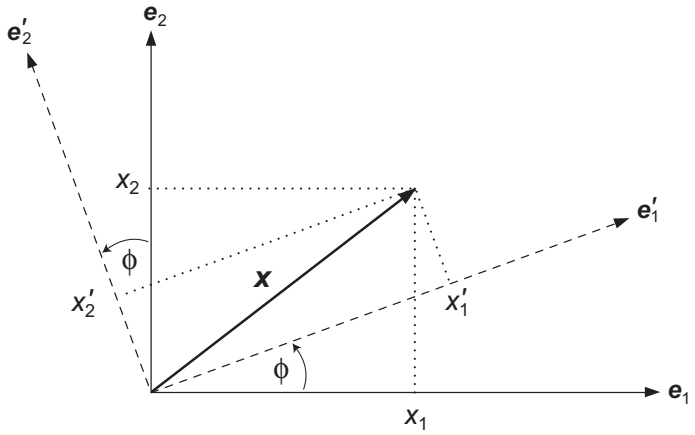


Fig. 15.4 Rotation of the coordinate system for a vector $\boldsymbol{x}$. The vector is invariant under the rotation but its components in the original and rotated coordinate systems are different.

Applying the usual (Leibniz) rule for the derivative of a product, the partial derivative is

$$\frac{\partial \boldsymbol{V}}{\partial x^j} = \frac{\partial V^i}{\partial x^j}\boldsymbol{e}_i + V^i \frac{\partial \boldsymbol{e}_i}{\partial x^j}, \tag{15.41}$$

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 (15.40),

$$\frac{\partial \boldsymbol{e}_i}{\partial x^j} = \Gamma^k_{ij}\boldsymbol{e}_k. \tag{15.42}$$

The Γ^k_{ij} appearing in Eq. (15.42) 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.

15.3.10 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. 15.4. 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:

$$\boldsymbol{x} = x^i \boldsymbol{e}_i = x'^i \boldsymbol{e}'_i, \tag{15.43}$$

and from the geometry of Fig. 15.4 the components in the two bases are related by (assuming a clockwise rotation)

$$\begin{pmatrix} x'^1 \\ x'^2 \\ x'^3 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix}, \tag{15.44}$$

which may be written compactly as

$$x'^i = R^i_j x^j, \tag{15.45}$$

where the R^i_j are the elements of the matrix in Eq. (15.44). 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

$$\boldsymbol{x}' = \boldsymbol{x}'(\boldsymbol{x},t) = \boldsymbol{x} - \boldsymbol{v}t \qquad t' = t'(\boldsymbol{x},t) = t. \tag{15.46}$$

"Newtonian relativity" asserts that the laws of physics are invariant under Galilean transformations. The laws of mechanics at low velocity are approximately invariant under (15.46), 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.

15.4 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.[6] 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.

15.4.1 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 15.2 (adapted from Ref. [16]). 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:

$$x \equiv x^\mu = (x^0, x^1, x^2, x^3) = (ct, \boldsymbol{x}), \tag{15.47}$$

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, $x^\mu \to x'^\mu$, with x^μ and x'^μ labeling the *same point* but in two different coordinate systems. Our

[6] 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 16.8, it isn't clear whether the Maxwell equations written in their usual form (1.1) are invariant under Lorentz transformation (because they are not formulated in standard Lorentz 4-vector notation), but they are, and they can be re-written so that the Lorentz invariance is manifest.

Box 15.2 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.

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

$$x'^{\mu} = x'^{\mu}(x) \qquad (\mu = 0,1,2,3), \tag{15.48}$$

where $x = x^{\mu}$ denotes the original (untransformed) coordinates. This notation is an economical form of

$$x'^{\mu} = f^{\mu}(x^0,x^1,x^2,x^3) \qquad (\mu = 0,1,\ldots), \tag{15.49}$$

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. (15.48). The coordinates in Eq. (15.48) are just labels so the laws of physics cannot depend on them. Hence the system x'^{μ} is not privileged and Eq. (15.48) 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.

15.4.2 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. 15.5 for 1D and 2D spheres.[7] 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.

[7] The idea conveyed by Fig. 15.5 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.

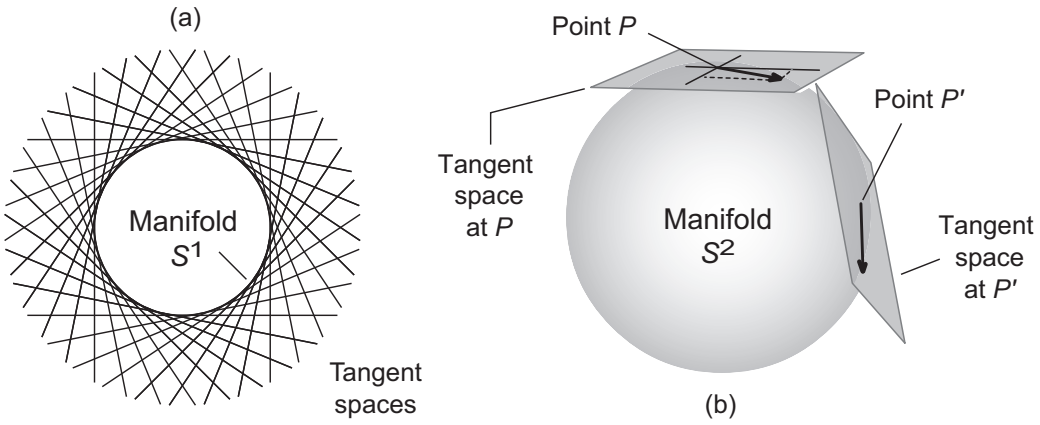


Fig. 15.5 Tangent spaces in curved manifolds, illustrated (a) for the manifold S^1 and (b) for the manifold S^2. As illustrated for S^2, vectors (indicated by arrows) are defined in the tangent spaces at each point, not in the curved manifold. Embedding the 1D tangent-space manifold in 2D euclidean space 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.

15.4.3 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 *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.*

15.4.4 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

$$e_\mu = \underset{\delta x^\mu \to 0}{\text{Lim}} \frac{\delta s}{\delta x^\mu}, \tag{15.50}$$

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 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

$$\frac{df}{d\lambda} \equiv \underset{\varepsilon \to 0}{\text{Lim}} \left[\frac{f(x^\mu(\lambda+\varepsilon)) - f(x^\mu(\lambda))}{\varepsilon} \right] = \frac{dx^\mu}{d\lambda} \frac{\partial f}{\partial x^\mu} = t^\mu \frac{\partial f}{\partial x^\mu},$$

and since $f(x)$ is arbitrary this implies the operator relation

$$\frac{d}{d\lambda} = \frac{dx^\mu}{d\lambda} \frac{\partial}{\partial x^\mu} = t^\mu \frac{\partial}{\partial x^\mu}. \tag{15.51}$$

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,

$$e_\mu = \frac{\partial}{\partial x^\mu} \equiv \partial_\mu, \tag{15.52}$$

which permits an arbitrary vector to be expanded as

$$V = V^\mu e_\mu = V^\mu \frac{\partial}{\partial x^\mu} = V^\mu \partial_\mu. \tag{15.53}$$

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. 15.6 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. (15.50) 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,

$$e_\mu(x) \cdot e_\nu(x) \equiv g_{\mu\nu}(x). \tag{15.54}$$

The scalar product of vectors A and B in a coordinate basis is given by

$$A \cdot B = (A^\mu e_\mu) \cdot (B^\nu e_\nu) = g_{\mu\nu} A^\mu B^\nu. \tag{15.55}$$

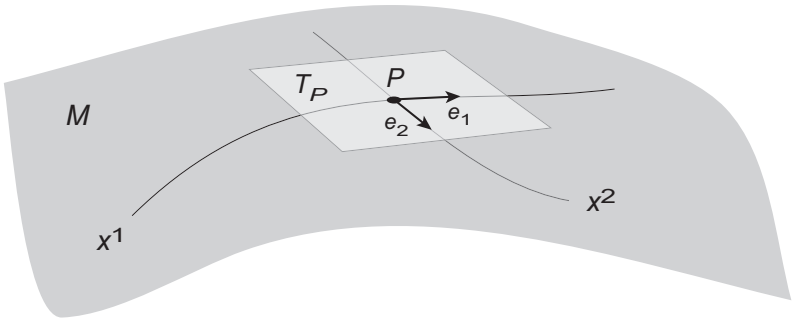


Fig. 15.6 Tangent space T_P at a point P for a curved 2D manifold M. The vectors tangent to the coordinate curves at each point define a coordinate or holonomic basis. This figure is a generalization of Fig. 15.5 to an arbitrary curved 2D manifold with a position-dependent, non-orthogonal (coordinate) 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. (15.50)–(15.53).

Equation (15.54) 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 (15.54)–(15.55), 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[8]

$$\omega = \omega_\mu e^\mu, \tag{15.56}$$

allowing the metric tensor with upper indices to be defined through

$$e^\mu(x) \cdot e^\nu(x) \equiv g^{\mu\nu}(x), \tag{15.57}$$

with the scalar product of arbitrary dual vectors α and β given by

$$\alpha \cdot \beta = g^{\mu\nu}\alpha_\mu\beta_\nu. \tag{15.58}$$

Equation (15.57) may be taken as a definition of a dual-vector coordinate basis $\{e^\mu\}$.

Just as Eqs. (15.54) or (15.57) are characteristic of a coordinate basis, an orthonormalized non-coordinate basis is specified by the requirement

$$e_{\hat{\mu}}(x) \cdot e_{\hat{\nu}}(x) = \eta_{\hat{\mu}\hat{\nu}}, \tag{15.59}$$

where $\eta = \mathrm{diag}\{-1,1,1,1\}$. In this expression, hats on indices indicate explicitly that this is an orthonormal and not coordinate basis. Our discussion will seldom require display of

[8] 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 15.3.5.

explicit basis vectors but usually we will assume implicitly the use of a coordinate basis such that Eqs. (15.50)–(15.58) are valid.

15.4.5 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. (15.48). 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.[9]

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^{ν}_{μ} is a mixed tensor of type $(1,1)$ and rank 2.

15.4.6 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 15.5 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.

[9] 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.

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.

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

$$\omega(aA+bB) = a\omega(A)+b\omega(B) \in \mathbb{R}, \tag{15.60}$$

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 15.5 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 15.3. 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,

$$V = V^\mu e_\mu, \tag{15.61}$$

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,

$$\omega = \omega_\mu e^\mu, \tag{15.62}$$

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 15.4.2. 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 15.3.5, the vector spaces for V and ω are said to be *dual* in the following sense.

Box 15.3 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.

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.

This *duality* of vector and dual vector spaces can be implemented systematically by requiring the basis vectors to satisfy

$$e^\mu(e_\nu) = e^\mu \cdot e_\nu = \delta^\mu_\nu, \tag{15.63}$$

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. (15.63) is also commonly written as $\langle e^\mu, e_\nu\rangle = \delta^\mu_\nu$. From Eqs. (15.60)–

(15.63) it follows that a dual vector ω acts on a vector V in the manner

$$\begin{aligned}\omega(V) = \langle \omega, V \rangle &= \omega_\mu e^\mu (V^\nu e_\nu) \\ &= \omega_\mu V^\nu e^\mu (e_\nu) \\ &= \omega_\mu V^\nu \delta^\mu_\nu \\ &= \omega_\mu V^\mu \in \mathbb{R}, \end{aligned} \tag{15.64}$$

where $\mathbb{R}$ denotes the real numbers.[10] This illustrates clearly that a dual vector is an operator 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 15.6 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. [37]). 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.

15.4.7 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

[10] 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 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".

corresponding basis vectors and dual basis vectors; for example,

$$V^\mu = V(e^\mu) = e^\mu \cdot V \qquad \omega_\mu = \omega(e_\mu) = e_\mu \cdot \omega, \tag{15.65}$$

which follows from Eq. (15.63). Equations such as (15.65) 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.

Example 15.7 The validity of Eqs. (15.65) 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 (15.61)–(15.62), the linearity requirement (15.60), and the orthogonality condition (15.63) were employed.

Vector and dual vector components as in Eq. (15.65), 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 15.8 illustrates for dual vectors and vectors.

Example 15.8 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. (15.71). By a similar proof, vector components V^μ may be shown to have the transformation law (15.73).

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 15.5–15.8 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.

15.4.8 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,

$$\begin{aligned} V_\mu \equiv \tilde{V}(e_\mu) &= g(V, e_\mu) \\ &= g(V^\nu e_\nu, e_\mu) \\ &= V^\nu g(e_\nu, e_\mu) \\ &= g_{\mu\nu} V^\nu. \end{aligned} \tag{15.66}$$

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,

$$V_\mu = g_{\mu\nu} V^\nu \qquad V^\mu = g^{\mu\nu} V_\nu. \tag{15.67}$$

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. (15.67), 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:

$$g(U,V) = g_{\mu\nu}U^\mu V^\nu = U_\nu V^\nu. \tag{15.68}$$

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.

15.4.9 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 15.5 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.

15.5 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 15.4.7) 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 15.3 for simpler euclidean spaces; you are urged to review that material if any conceptual difficulties are encountered in the following material.

15.5.1 Scalar Transformation Law

Consider the behavior of fields under the coordinate transformations introduced in Section 15.4.1. 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 (15.48),

$$\phi'(x') = \phi(x). \tag{15.69}$$

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.[11] 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.

15.5.2 Dual Vector Transformation Law

In parallel with the discussion of vectors defined in the tangent and dual bases for euclidean space in Section 15.6, 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 15.3, and because of the duality discussed in Section 15.3.5. By the argument in Section 15.4.8, 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

$$\frac{\partial \phi(x)}{\partial x'^{\mu}} = \frac{\partial \phi(x)}{\partial x^{\nu}} \frac{\partial x^{\nu}}{\partial x'^{\mu}}. \tag{15.70}$$

Now consider a vector having a transformation law mimicking that of the scalar field gradient (15.70),

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)}. \tag{15.71}$$

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 (15.71) 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

[11] 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.

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 using. For example, Eq. (15.71) 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.

15.5.3 Vector Transformation Law

Now consider application of the rules of partial differentiation to transformation of the differential,

$$dx'^{\mu} = \frac{\partial x'^{\mu}}{\partial x^{\nu}}\, dx^{\nu}. \tag{15.72}$$

This suggests a second vector transformation rule,

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}. \tag{15.73}$$

A quantity behaving in this way is termed a *vector*.[12] 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 (15.73).

Example 15.9 Equations (15.71) and (15.73) may be viewed as matrix equations,

$$A'_{\mu}(x') = \hat{U}^{\nu}_{\mu} A_{\nu}(x) \qquad A'^{\mu}(x') = U^{\mu}_{\nu} A^{\nu}(x), \tag{15.74}$$

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*.

15.5.4 Duality of Vectors and Dual Vectors

The preceding discussion distinguishes two kinds of "vectors": dual vectors, which carry a lower index and transform like (15.71), and vectors, which carry an upper index and

[12] 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.

transform like (15.73). This distinction is analogous to that introduced in Section 15.6 for vectors and dual vectors in euclidean spaces. In particular instances vectors and dual vectors may 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 15.4.8 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.

15.6 Scalar Product of Vectors

Just as was found in Section 15.3.4 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

$$A \cdot B \equiv A_\mu B^\mu, \tag{15.75}$$

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. (15.71) and (15.73),

$$\begin{aligned} A' \cdot B' = A'_\mu B'^\mu &= \frac{\partial x^\nu}{\partial x'^\mu} A_\nu \frac{\partial x'^\mu}{\partial x^\alpha} B^\alpha = \frac{\partial x^\nu}{\partial x'^\mu}\frac{\partial x'^\mu}{\partial x^\alpha} A_\nu B^\alpha \\ &= \frac{\partial x^\nu}{\partial x^\alpha} A_\nu B^\alpha = \delta^\nu_\alpha A_\nu B^\alpha = A_\alpha B^\alpha = A \cdot B, \end{aligned} \tag{15.76}$$

where the Kronecker delta δ^ν_μ is given by

$$\delta^\mu_\nu = \frac{\partial x'^\mu}{\partial x'^\nu} = \frac{\partial x^\mu}{\partial x^\nu} = \begin{cases} 1 & \mu = \nu \\ 0 & \mu \neq \nu \end{cases}, \tag{15.77}$$

which is a rank-2 tensor with the unusual property that its components take the same value in all coordinate systems.

15.7 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

$$T'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta}, \tag{15.78}$$

$$T'^\nu{}_\mu = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x'^\nu}{\partial x^\beta} T^\beta{}_\alpha, \tag{15.79}$$

$$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}. \tag{15.80}$$

Table 15.1 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	$T'^\nu{}_\mu = \dfrac{\partial x^\alpha}{\partial x'^\mu} \dfrac{\partial x'^\nu}{\partial x^\beta} T^\beta{}_\alpha$

These transformation rules may be generalized easily to tensors of any rank. In such a generalization,

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 15.1.

15.8 The Metric Tensor

By analogy with the discussion in Section 15.3.7, a rank-2 tensor of special importance is the metric tensor $g_{\mu\nu}$, because it is associated with the line element

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{15.81}$$

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

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta}. \tag{15.82}$$

The contravariant form of the metric tensor $g^{\mu\nu}$ is defined by the requirement

$$g_{\mu\alpha} g^{\alpha\nu} = \delta^\nu_\mu, \tag{15.83}$$

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,

$$A^\mu = g^{\mu\nu}A_\nu \qquad A_\mu = g_{\mu\nu}A^\nu \qquad T^\mu{}_\nu = g_{\nu\alpha}T^{\mu\alpha} \qquad T^\alpha{}_{\beta\gamma} = g^{\alpha\mu}g_{\gamma\epsilon}T_{\mu\beta}{}^\epsilon. \quad (15.84)$$

Thus, the scalar product of vectors may be expressed as

$$A\cdot B = g_{\mu\nu}A^\mu B^\nu \equiv A_\nu B^\nu = g^{\mu\nu}A_\mu B_\nu \equiv A^\nu B_\nu. \quad (15.85)$$

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. (15.81).

In mixed-tensor expressions like the third or fourth ones in Eq. (15.84) 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 15.9): $T^\mu{}_\nu = g_{\nu\alpha}T^{\mu\alpha}$ 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 15.4.2). However, Eqs. (15.83)–(15.85) and the discussion in Section 15.3.5, and Section 15.4.8 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. (15.84). 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 15.5.2 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.

15.9 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

$$T_{\alpha\beta} = \tfrac{1}{2}(T_{\alpha\beta} + T_{\beta\alpha}) + \tfrac{1}{2}(T_{\alpha\beta} - T_{\beta\alpha}), \tag{15.86}$$

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.

15.10 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.

15.11 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.

15.11.1 Invariant Integration

Under a change of coordinates the volume element for integration over the spacetime coordinates changes according to

$$d^4x' = \det\left(\frac{\partial x'}{\partial x}\right) d^4x = J d^4x, \tag{15.87}$$

where $d^4x \equiv dx^0 dx^1 dx^2 dx^3$ 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 15.9). Notice that the right side of Eq. (15.82) may be viewed as a triple matrix product and recall that the determinant of a matrix product is the product of determinants. Thus Eq. (15.82) 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. (16.6)], and this result may be substituted into Eq. (15.87) to give $\sqrt{|g'|}\, d^4x' = \sqrt{|g|}\, d^4x$. This implies that an *invariant volume element,*

$$dV = \sqrt{|g|}\, d^4x, \tag{15.88}$$

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 15.10.

Example 15.10 The metric for a 2-dimensional spherical surface (the 2-sphere S^2) is specified by the line element $d\ell^2 = R^2 d\theta^2 + R^2 \sin^2\theta d\phi^2$, 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. (15.88).

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. (15.88). What about the derivatives of tensor quantities? This is a more complicated issue that we must now address.

15.11.2 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.[13]

$$\partial_\mu \phi = \phi_{,\mu} \equiv \frac{\partial \phi(x)}{\partial x^\mu} \qquad \partial'_\mu \phi' = \phi'_{,\mu} \equiv \frac{\partial \phi'(x')}{\partial x'^\mu} \qquad \partial^\mu \phi = \phi^{,\mu} \equiv \frac{\partial \phi(x)}{\partial x_\mu}. \tag{15.89}$$

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. (15.70), which is just the transformation law (15.71); 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,

$$\begin{aligned} A'_{\mu,\nu} \equiv \frac{\partial A'_\mu}{\partial x'^\nu} &= \frac{\partial}{\partial x'^\nu}\left(A_\alpha \frac{\partial x^\alpha}{\partial x'^\mu}\right) \\ &= \frac{\partial A_\alpha}{\partial x'^\nu}\frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu} \\ &= \frac{\partial A_\alpha}{\partial x^\beta}\frac{\partial x^\beta}{\partial x'^\nu}\frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu} \\ &= A_{\alpha,\beta}\frac{\partial x^\beta}{\partial x'^\nu}\frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu}. \end{aligned} \tag{15.90}$$

[13] 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. (15.90). 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 15.11.3

15.11.3 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

$$\Gamma'^{\lambda}_{\alpha\beta} = \Gamma^{\kappa}_{\mu\nu}\frac{\partial x^{\mu}}{\partial x'^{\alpha}}\frac{\partial x^{\nu}}{\partial x'^{\beta}}\frac{\partial x'^{\lambda}}{\partial x^{\kappa}} + \frac{\partial^2 x^{\mu}}{\partial x'^{\alpha}\partial x'^{\beta}}\frac{\partial x'^{\lambda}}{\partial x^{\mu}}, \tag{15.91}$$

it may be shown that

$$\left(A'_{\mu,\nu} - \Gamma'^{\lambda}_{\mu\nu}A'_{\lambda}\right) = \left(A_{\alpha,\beta} - \Gamma^{\kappa}_{\alpha\beta}A_{\kappa}\right)\frac{\partial x^{\alpha}}{\partial x'^{\mu}}\frac{\partial x^{\beta}}{\partial x'^{\nu}}. \tag{15.92}$$

Comparing with Eq. (15.78), 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

$$A_{\mu;\nu} \equiv A_{\mu,\nu} - \Gamma^{\lambda}_{\mu\nu}A_{\lambda}, \tag{15.93}$$

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,

$$\nabla_\nu A_\mu = A_{\mu;\nu} \equiv \partial_\nu A_\mu - \Gamma^\lambda_{\mu\nu} A_\lambda . \tag{15.94}$$

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

$$A^\lambda{}_{;\mu} = A^\lambda{}_{,\mu} + \Gamma^\lambda_{\alpha\mu} A^\alpha \qquad \nabla_\mu A^\lambda = \partial_\mu A^\lambda + \Gamma^\lambda_{\alpha\mu} A^\alpha , \tag{15.95}$$

where the result of (15.95) is a mixed rank-2 tensor, and the covariant derivatives of the three possible rank-2 tensors through

$$A_{\mu\nu;\lambda} = A_{\mu\nu,\lambda} - \Gamma^\alpha_{\mu\lambda} A_{\alpha\nu} - \Gamma^\alpha_{\nu\lambda} A_{\mu\alpha} , \tag{15.96}$$

$$A^\mu{}_{\nu;\lambda} = A^\mu{}_{\nu,\lambda} + \Gamma^\mu_{\alpha\lambda} A^\alpha{}_\nu - \Gamma^\alpha_{\nu\lambda} A^\mu{}_\alpha , \tag{15.97}$$

$$A^{\mu\nu}{}_{;\lambda} = A^{\mu\nu}{}_{,\lambda} + \Gamma^\mu_{\alpha\lambda} A^{\alpha\nu} + \Gamma^\nu_{\alpha\lambda} A^{\mu\alpha} , \tag{15.98}$$

or in alternative notation

$$\nabla_\lambda A_{\mu\nu} = \partial_\lambda A_{\mu\nu} - \Gamma^\alpha_{\mu\lambda} A_{\alpha\nu} - \Gamma^\alpha_{\nu\lambda} A_{\mu\alpha} , \tag{15.99}$$

$$\nabla_\lambda A^\mu{}_\nu = \partial_\lambda A^\mu{}_\nu + \Gamma^\mu_{\alpha\lambda} A^\alpha{}_\nu - \Gamma^\alpha_{\nu\lambda} A^\mu{}_\alpha , \tag{15.100}$$

$$\nabla_\lambda A^{\mu\nu} = \partial_\lambda A^{\mu\nu} + \Gamma^\mu_{\alpha\lambda} A^{\alpha\nu} + \Gamma^\nu_{\alpha\lambda} A^{\mu\alpha} , \tag{15.101}$$

where the derivatives in Eqs. (15.96)–(15.101) 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 15.11).

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. (15.93)–(15.101) 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,

$$(A_\mu B_\nu)_{;\lambda} = A_{\mu;\lambda} B_\nu + A_\mu B_{\nu;\lambda} . \tag{15.102}$$

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 15.11 The Leibniz differentiation rule for the product of two vectors may be used to derive the expressions (15.96)–(15.98) for the covariant derivatives of rank-2 tensors from those given for vectors in Eqs. (15.93) and (15.95). 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. (15.95) 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. (15.98) 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. (15.101): 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,[14]

$$\nabla_\alpha g_{\mu\nu} = g_{\mu\nu;\alpha} = 0. \tag{15.103}$$

This means that in manipulating tensor equations the operation of covariant differentiation commutes with the operation (15.84) 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.

[14] The validity of Eq. (15.103) is a consequence of particular assumptions concerning the nature of parallel transport in curved spacetime.

15.12 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. (15.79) 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 15.11.1.
3. Replace all partial derivatives with the covariant derivatives introduced in Section 15.11.3.
4. Take care to remember that a covariant differentiation generally does not commute with a second covariant differentiation, so the order matters.

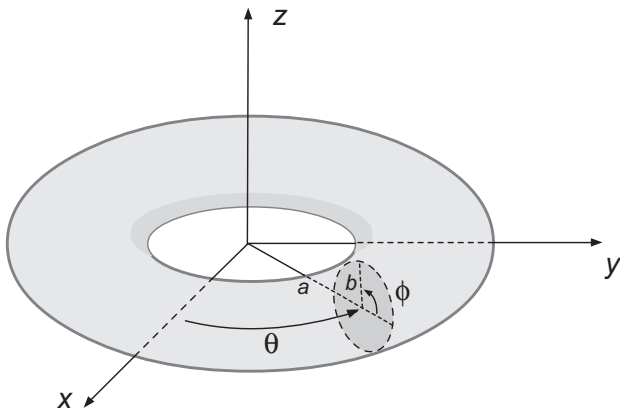
This prescription in terms of tensors will provide a powerful formalism for dealing with mathematical relations that would be much more formidable in standard notation.

Background and Further Reading

Significant parts of this chapter have been adapted from Refs. [16] and [17]. The discussion of coordinate transformations in euclidean space is based on the presentation of Ref. [10].

Problems

15.1 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.

15.2 (a) Verify explicitly that the Lorentz transformation of Eq. (16.25)

$$\begin{aligned} cdt' &= c\cosh\xi\, dt + \sinh\xi\, dx \\ dx' &= c\sinh\xi\, dt + \cosh\xi\, dx \\ dy' &= dy \\ dz' &= dz \end{aligned}$$

leaves invariant the Minkowski line element ds^2.

(b) Use the Lorentz transformations (16.31) 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.

15.3 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 15.3. Verify that the matrices g^{ij} and g_{ij} found for this problem are inverses of each other. ***

15.4 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.

15.5 A theorem useful in various contexts states that if a set of quantities produces a tensor when contracted with a tensor, then that set of quantities is necessarily a tensor. This is called the *quotient theorem.* (a) Use the definition of the scalar product of vectors and the quotient theorem to argue that the metric $g_{\mu\nu}$ is a tensor. (b) Use the quotient theorem to show explicitly that if for an arbitrary vector V^γ

$$T^\alpha{}_{\beta\gamma}V^\gamma = \frac{\partial x'^\alpha}{\partial x^\delta}\frac{\partial x^\varepsilon}{\partial x'^\beta}T^\delta{}_{\varepsilon\phi}V^\phi,$$

then $T^\alpha{}_{\beta\gamma}$ is necessarily a tensor of type $(1,2)$. ***

16 Special Relativity

In this chapter we shall build on the mathematical foundation of Ch. 15 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.

16.1 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.

16.2 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.[2]

[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. The interested reader will find a tractable introduction to general relativity and its relationship to special relativity in Ref. [15].

[2] Physicists almost always work in metric spaces and usually take existence of a metric for granted (it is hard to do physics without measurement of distance). However, mathematicians are quite familiar with respectible manifolds that need not include a metric among their attributes.

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 manifolds, fundamentally new features will enter. Many of these new features are associated with the *indefinite metric* of Minkowski space.

16.2.1 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. 15 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.

16.2.2 Scalar Products and the Metric Tensor

In a particular inertial frame, introduce unit vectors e_0, e_1, e_2, and e_3 that point along the t, x, y, and z axes, respectively, allowing any 4-vector A to be expressed as

$$A = A^0 e_0 + A^1 e_1 + A^2 e_2 + A^3 e_3. \tag{16.1}$$

Thus (A^0, A^1, A^2, A^3) are the (contravariant) components of the 4-vector A.[3] The scalar product of 4-vectors is given by

$$A \cdot B = B \cdot A = (A^\mu e_\mu) \cdot (B^\nu e_\nu) = e_\mu \cdot e_\nu A^\mu B^\nu. \tag{16.2}$$

Introducing the definition

$$\eta_{\mu\nu} \equiv e_\mu \cdot e_\nu, \tag{16.3}$$

we may express the scalar product as

$$A \cdot B = \eta_{\mu\nu} A^\mu B^\nu. \tag{16.4}$$

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 standard to give it the special symbol $\eta_{\mu\nu}$.

16.2.3 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

$$ds^2 = -c^2 d\tau^2 = \eta_{\mu\nu} dx^\mu dx^\nu = -c^2 dt^2 + dx^2 + dy^2 + dz^2, \tag{16.5}$$

[3] Recall that non-bold symbols denote 4-vectors, bold symbols denote 3-vectors, and that a notation such as A^μ may stand for the full 4-vector, or for a component of it, depending on 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

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \equiv \text{diag } (-1,1,1,1). \tag{16.6}$$

Then Eq. (16.5) for the line element may be written as the matrix equation

$$ds^2 = (cdt \;\; dx \;\; dy \;\; dz) \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cdt \\ dx \\ dy \\ dz \end{pmatrix}, \tag{16.7}$$

where ds^2 represents the square of the spacetime interval between x and $x+dx$ with

$$x = (x^0, x^1, x^2, x^3) = (ct, x^1, x^2, x^3). \tag{16.8}$$

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 16.1 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. (15.84) 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 (16.6) 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.

16.2.4 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. (16.5) 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 $c^2 dt^2$; it is true only for the particular combination of spatial and time intervals defined by Eq. (16.5). Because of this invariance, Minkowski space is the natural manifold for the formulation of special relativity.

Example 16.2 The metric can be used to determine the relationship between the time coordinate t and the proper time τ. From Eq. (16.5)

$$\begin{aligned} d\tau^2 = \frac{-ds^2}{c^2} &= \frac{1}{c^2}(c^2dt^2 - dx^2 - dy^2 - dz^2) \\ &= dt^2\left\{1 - \frac{1}{c^2}\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2\right]\right\} \\ &= \left(1 - \frac{v^2}{c^2}\right)dt^2. \end{aligned} \tag{16.9}$$

where v is the magnitude of the velocity. Therefore, the proper time that elapses between coordinate times t_1 and t_2 is

$$\tau_{12} = \int_{t_1}^{t_2}\left(1 - \frac{v^2}{c^2}\right)^{1/2} dt\,. \tag{16.10}$$

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. (16.10) is always less than one. This is the special-relativistic *time dilation effect,* stated in general form. For the special case of constant velocity, (16.10) yields

$$\Delta\tau = \left(1 - \frac{v^2}{c^2}\right)^{1/2}\Delta t, \tag{16.11}$$

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 (16.5) under transformations between inertial frames. As was suggested by the discussion in Ch. 15, 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.

16.3 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 15.1 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

$$\varphi' = \varphi \qquad \text{Scalar} \tag{16.12}$$

$$A'^{\mu} = \Lambda^{\mu}{}_{\nu} A^{\nu} \qquad \text{Vector} \tag{16.13}$$

$$A'_{\mu} = \Lambda_{\mu}{}^{\nu} A_{\nu} \qquad \text{Dual vector} \tag{16.14}$$

$$T'^{\mu\nu} = \Lambda^{\mu}{}_{\gamma} \Lambda^{\nu}{}_{\delta} T^{\gamma\delta} \qquad \text{Contravariant rank-2 tensor} \tag{16.15}$$

$$T'_{\mu\nu} = \Lambda_{\mu}{}^{\gamma} \Lambda_{\nu}{}^{\delta} T_{\gamma\delta} \qquad \text{Covariant rank-2 tensor} \tag{16.16}$$

$$T'^{\mu}{}_{\nu} = \Lambda^{\mu}{}_{\gamma} \Lambda_{\nu}{}^{\delta} T^{\gamma}{}_{\delta} \qquad \text{Mixed rank-2 tensor} \tag{16.17}$$

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. (15.90) can be transformed away so *covariant derivatives are equivalent to partial derivatives* in Minkowski space. In the transformation laws (16.17) the $\Lambda^{\mu}{}_{\nu}$ are elements of *Lorentz transformations* that we will now discuss in more detail.

16.4 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.

16.4.1 Rotations in Euclidean Space

First we consider a rotation of the coordinate system in euclidean space, as illustrated in Fig. 15.4. 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. (15.45)] act on the metric tensor g_{ij} in the following way

$$R g_{ij} R^{\mathrm{T}} = g_{ij}, \tag{16.18}$$

where R^{T} denotes the transpose of R. For euclidean space the metric tensor is just the unit matrix so the requirement (16.18) 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. (16.18) is valid generally, not just for euclidean spaces. Therefore, it may be used as guidance for constructing more general rotations in Minkowski space.

16.4.2 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,

$$dx'^{\mu} = \Lambda^{\mu}{}_{\nu} dx^{\nu}, \tag{16.19}$$

where the transformation matrix $\Lambda^{\mu}{}_{\nu}$ is expected to satisfy the analog of Eq. (16.18) for the Minkowski metric $\eta_{\mu\nu}$,

$$\Lambda \eta_{\mu\nu} \Lambda^{\mathrm{T}} = \eta_{\mu\nu}, \tag{16.20}$$

or explicitly in terms of components, $\Lambda_{\mu}{}^{\rho} \Lambda^{\sigma}{}_{\nu} \eta_{\rho\sigma} = \eta_{\mu\nu}$.[4] 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 16.1). 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 16.1. Consider first the simple case of rotations about the z axis.

16.4.3 Lorentz Spatial Rotations

Rotations about a single spatial axis in Minkowski space correspond to a 2-dimensional problem with euclidean metric, so the condition (16.18) may be written as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{16.21}$$

[4] 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 (16.20) 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 15.1).

Box 16.1 Symmetries and groups [16]

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).

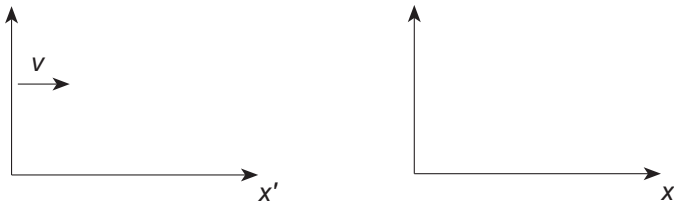


Fig. 16.1 A Lorentz boost along the positive x axis by a velocity v.

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

$$a=\cos\theta \qquad b=\sin\theta \qquad c=-\sin\theta \qquad d=\cos\theta.$$

This leads to the standard result

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = R\begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x^1 \\ x^2 \end{pmatrix}, \tag{16.22}$$

which is Eq. (15.44) restricted to rotations about a single axis. Now we shall apply this same technique to determine the elements of a Lorentz boost transformation.

16.4.4 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. 16.1. 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

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} cdt \\ dx \end{pmatrix}, \tag{16.23}$$

and the condition (16.20) can be written out explicitly as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{16.24}$$

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 16.2 Minkowski rotations

The respective derivations make clear that the appearance of hyperbolic functions in Eq. (16.25) instead of trigonometric functions as in Eq. (16.22) traces to the role of the indefinite metric diag$(-1,1)$ in Eq. (16.24) relative to the positive-definite metric diag$(1,1)$ in Eq. (16.21). 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 (16.5) with indefinite metric (16.6). 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 $-\infty \leq \xi \leq \infty$. Therefore, we may write the boost transformation as

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix} \begin{pmatrix} cdt \\ dx \end{pmatrix}. \tag{16.25}$$

A geometrical interpretation of this result is discussed in Box 16.2.

The Lorentz boost transformation (16.25) 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. (16.25) 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. (16.25), 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

$$\beta \equiv \frac{v}{c} = \frac{x}{ct} = -\frac{\sinh\xi}{\cosh\xi} = -\tanh\xi. \tag{16.26}$$

This relationship between ξ and β is plotted in Fig. 16.2. Utilizing the identity $1 = \cosh^2\xi - \sinh^2\xi$ and the definition

$$\gamma \equiv \left(1 - \frac{v^2}{c^2}\right)^{-1/2} \tag{16.27}$$

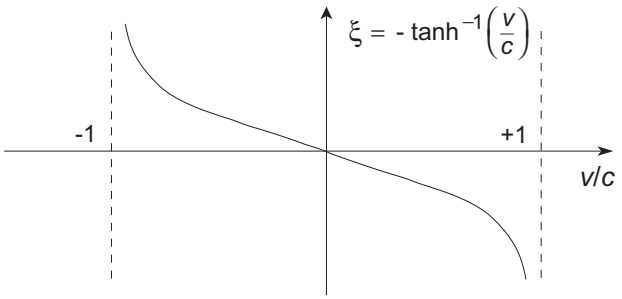


Fig. 16.2 Dependence of the Lorentz parameter ξ on $\beta = v/c$.

of the *Lorentz γ-factor,* we may write

$$\cosh\xi = \sqrt{\frac{\cosh^2\xi}{1}} = \frac{1}{\sqrt{1-\sinh^2\xi/\cosh^2\xi}} = \frac{1}{\sqrt{1-v^2/c^2}} = \gamma, \tag{16.28}$$

and from this result and (16.26),

$$\sinh\xi = -\beta\cosh\xi = -\beta\gamma. \tag{16.29}$$

Inserting (16.28)–(16.29) into (16.25) for finite intervals gives

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} \tag{16.30}$$

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,

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) \qquad x' = \gamma(x - vt) \qquad y' = y \qquad z' = z, \tag{16.31}$$

with the inverse transformation corresponding to the replacement $v \to -v$. By inspection, these reduce to the Galilean boost equations (15.46) if $v/c \to 0$ and it is easy to verify that the Lorentz transformations leave invariant the spacetime interval ds^2.

16.5 Lightcone Diagrams

The line element (16.5) defines a cone, implying that Minkowski spacetime may be classified according to the *lightcone diagram* exhibited in Fig. 16.3. 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*).

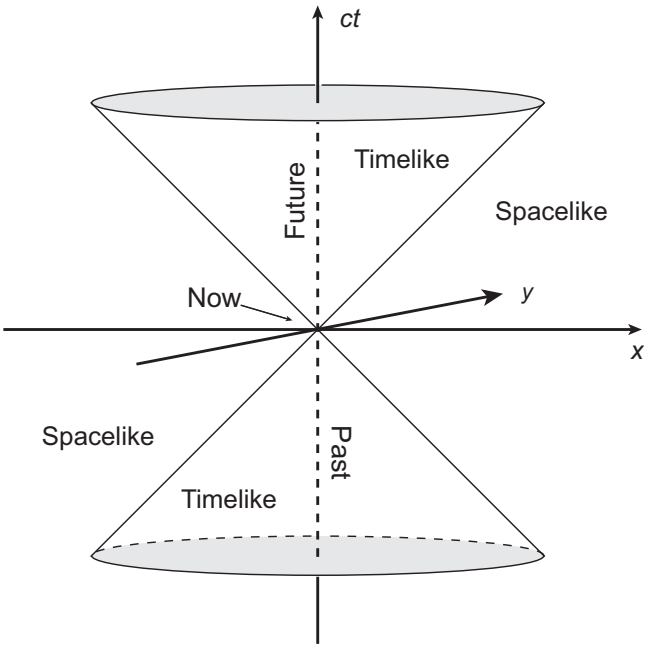


Fig. 16.3 Lightcone diagram for flat spacetime in two space and one time dimensions. The future lightcone is the surface swept out by a spherical light pulse emitted from the origin.

These regions for flat spacetime are labeled in Fig. 16.3. 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 16.3 The Minkowski line element (16.5) 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 $\pm 45^\circ$ angles in the lightcone diagram. By similar

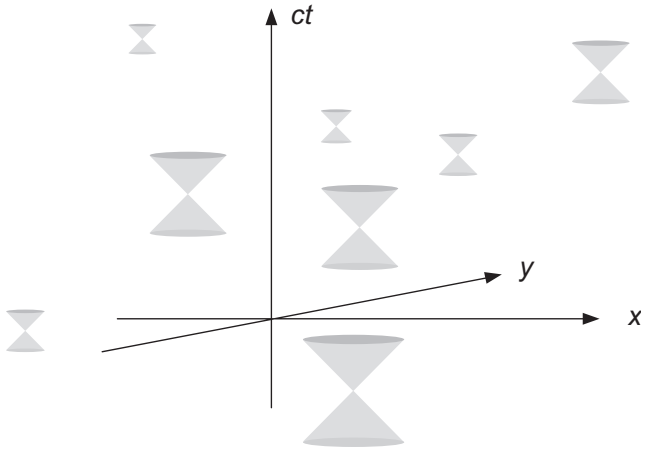


Fig. 16.4 Lightcones are local concepts. Each point of spacetime should be imagined to have its own lightcone.

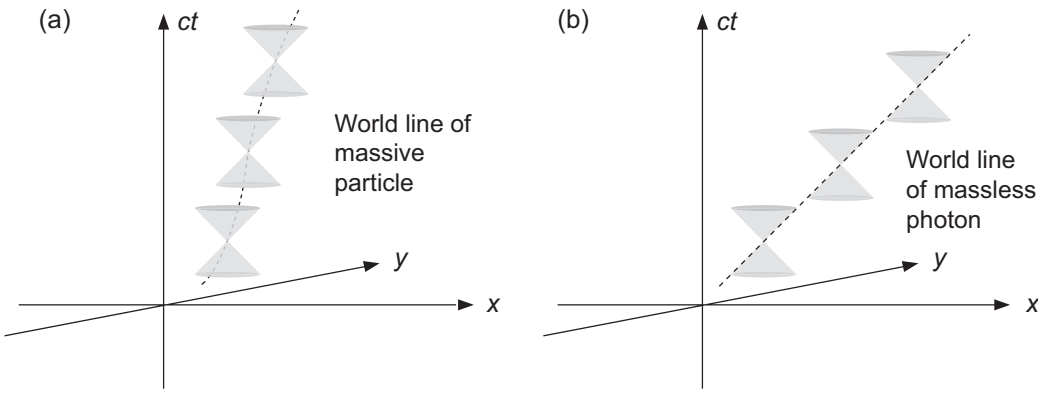


Fig. 16.5 Worldlines for (a) massive (timelike) particles and (b) massless (lightlike) particles. For massive particles the trajectory must always lie inside the local light cone at each point; for massless particles it must always lie on the lightcone at each point.

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. 16.3 was placed at the origin for illustration but each point in the spacetime has its own lightcone, as illustrated in Fig. 16.4. From Example 16.3, 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. 16.5(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. 16.5(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.

16.6 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. 16.3), as illustrated in Fig. 16.4. From Example 16.3, 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. 16.3) through signals propagating at $v < c$. Likewise, the event at the origin may be influenced by events within its backward lightcone ("Past" in Fig. 16.3) 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.

Box 16.3 Time Machines and Causal Paradoxes

Discussions of time travel are usually about going *backward* in time; it requires no special talent to travel *forward* in time, and various scenarios can be arranged that are 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!). For example, in the twin paradox it is possible to arrange for the traveler to arrive back at Earth centuries in the future relative to clocks that remain on Earth, which is a kind of time travel. Similar possibilities exist using the gravitational time dilation in strong gravity. However, the real question is, could you go *back in time* to explore your earlier history and potentially influence it?

Our current understanding of special and general relativity, and of the nature of the Universe, says no! Bending a forward-going timelike worldline continuously into a backward-going one *requires going outside the local lightcone,* implying a violation of the relativity constraint $v \leq c$. If *closed timelike loops* were permitted in general relativity, travel to earlier times might be possible. However, they are forbidden in our present understanding if there are no negative energy densities and the Universe has the topology in evidence. Thus, the committed time traveler must 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, and there is no evidence at present for exotic spacetime topologies with closed timelike loops, which bodes ill for time travelers. The reader is warned that these statements are based entirely on classical gravity considerations (general relativity). It is unclear at present whether some future understanding of quantum gravity could alter the prospects for time travel.

Example 16.4 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 special relativity, and general relativity inherits the local lightcone structure of Minkowski space. Therefore, as explored further in Box 16.3, 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

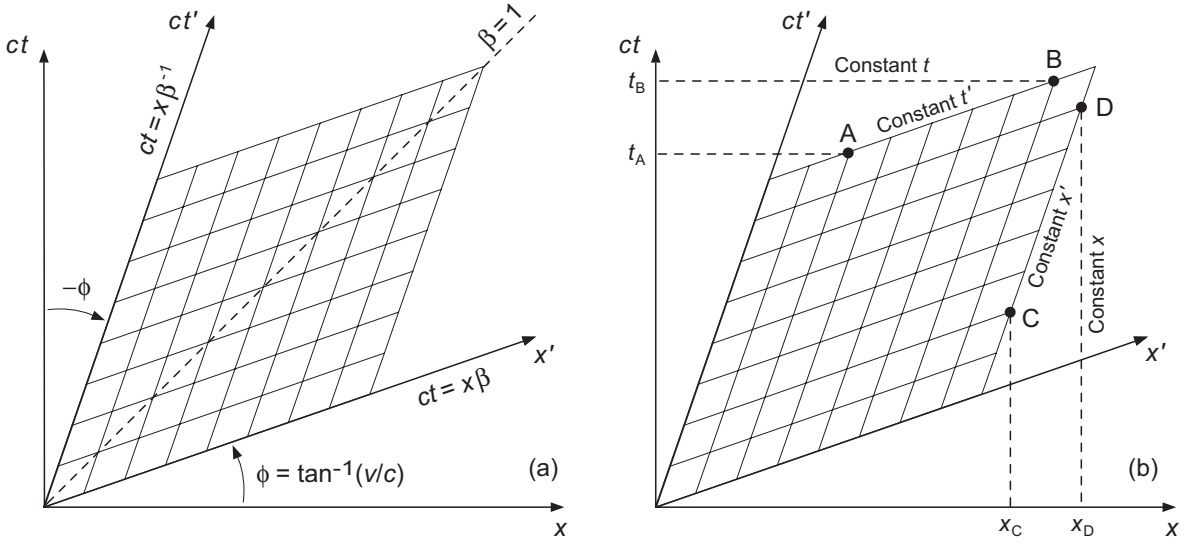


Fig. 16.6 (a) Lorentz boost transformation in a spacetime diagram. (b) Comparison of events in boosted and unboosted reference frames.

absolute simultaneity introduces *no causal ambiguity,* because all observers agree on the lightcone structure of spacetime and hence all observers agree that event A can cause event B only if A lies in the past lightcone of B, for example.

16.7 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. 16.6.

16.7.1 Lorentz Boosts and the Lightcone

What happens to the axes in Fig. 16.6 under a Lorentz boost? From the first two of Eqs. (16.31),

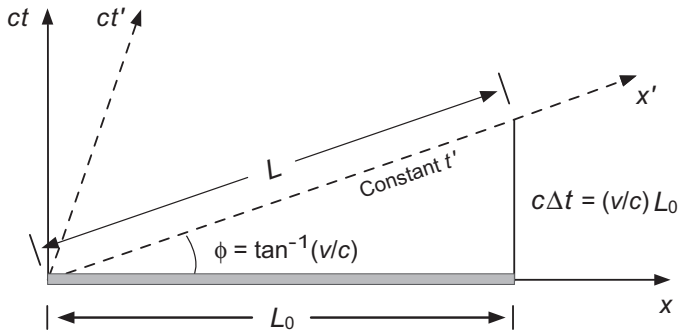
$$ct' = c\gamma\left(t - \frac{vx}{c^2}\right) \qquad x' = \gamma(x - vt). \tag{16.32}$$

The t' axis corresponds to $x' = 0$, which implies from the second of Eqs. (16.32) 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. (16.32) that $ct = x\beta$. The x' and ct' axes for the boosted system are also shown in Fig. 16.6(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. 16.6. For example, relativity of simultaneity follows directly, as illustrated in Fig. 16.6(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.

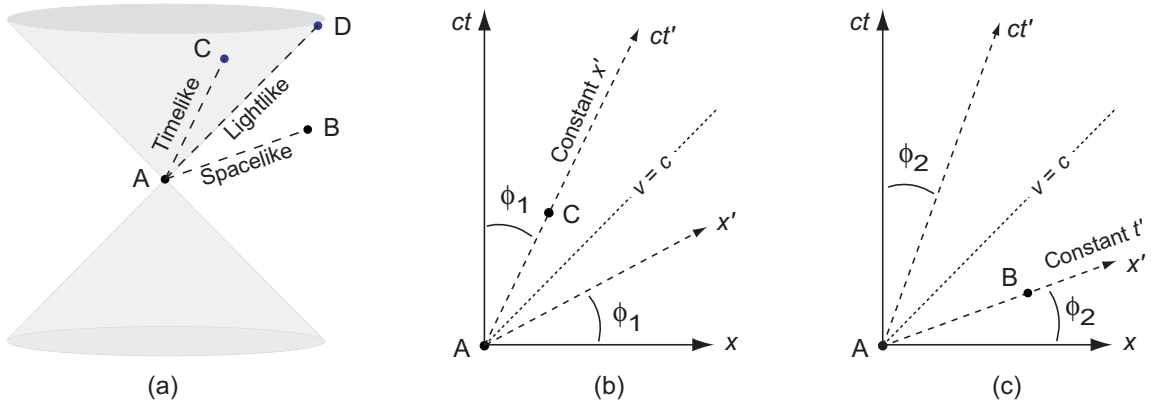
Example 16.5 Consider the following schematic representation of the spacetime diagram illustrated in Fig. 16.6, where a rod of length L_0, as measured in its own rest frame (t,x), is oriented along the x axis.



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. 16.6(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. (16.32) 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.



(a) (b) (c)

Fig. 16.7 (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).

16.7.2 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 light-cone at one of the points, as illustrated in Fig. 16.7(a). This geometry and that of Fig. 16.6 then suggest another important distinction between events at spacelike separations [the line AB in Fig. 16.7(a)] and timelike separations [the line AC in Fig. 16.7(a)]:

(i) If two events have a timelike separation, a Lorentz transformation exists that can bring them into spatial congruence. Figure 16.7(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 16.7(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).

16.8 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.

1. It provides a nice example of how useful Lorentz invariance and Lorentz tensors can be in application to a physical problem.
2. It is of *considerable historical interest* because the beauty and properties of the Maxwell equations had strong influence on Einstein in his development of the special theory of relativity.
3. There are many useful parallels between general relativity and the Maxwell theory, particularly for weak gravity where the Einstein field equations may be linearized.[5]

In the following sections we write the Maxwell equations in standard form, introduce a covariant notation for them, and finally rewrite them in a way that is manifestly Lorentz invariant because all quantities appearing in the equations will be Lorentz tensors.

16.8.1 Maxwell Equations in Non-Covariant Form

Let us first write the Maxwell equations in the usual non-covariant way. Departing from our normal SI units, we choose to do this in Heaviside–Lorentz units (see Appendix B.3), which will be recognized as very similar to SI units but with the speed of light set to $c = 1$. The free-space Maxwell equations may be written in Heaviside–Lorentz units as

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho \tag{16.33a}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} + \boldsymbol{\nabla} \times \boldsymbol{E} = 0 \tag{16.33b}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \tag{16.33c}$$

$$\boldsymbol{\nabla} \times \boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{J}, \tag{16.33d}$$

where $\boldsymbol{E}$ is the 3-vector electric field, $\boldsymbol{B}$ is the 3-vector magnetic field, ρ is the scalar charge density, and $\boldsymbol{J}$ is the current 3-vector, with the density and current required to satisfy the usual equation (1.3) of continuity

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{J} = 0.$$

The Maxwell equations are consistent with the special theory of relativity. However, in the form of Eqs. (16.33) 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.

[5] For example, this is particularly useful in understanding the propagation of gravitational waves far from strong sources. Then the highly-nonlinear Einstein equations that describe strong gravity can be linearized and the resulting formalism exhibits many similarities to the Maxwell equations of classical electromagnetism (see Section 22.2 of Ref. [16]).

16.8.2 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

$$\boldsymbol{B} \equiv \boldsymbol{\nabla} \times \boldsymbol{A} \qquad \boldsymbol{E} \equiv -\boldsymbol{\nabla}\Phi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{16.34}$$

The vector identities

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = 0 \qquad \boldsymbol{\nabla} \times \boldsymbol{\nabla}\Phi = 0, \tag{16.35}$$

may then be used to show that the second and third Maxwell equations are satisfied identically, and the identity

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \boldsymbol{\nabla}^2 \boldsymbol{A}, \tag{16.36}$$

may be used to write the remaining two Maxwell equations as the coupled second-order equations

$$\boldsymbol{\nabla}^2 \Phi + \frac{\partial}{\partial t} \boldsymbol{\nabla} \cdot \boldsymbol{A} = -\rho \tag{16.37}$$

$$\boldsymbol{\nabla}^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{\partial \Phi}{\partial t}\right) = -\boldsymbol{J}. \tag{16.38}$$

These equations may then be decoupled by exploiting a fundamental symmetry of electromagnetism termed *gauge invariance*.

16.8.3 Gauge Transformations

Because of the identity $\boldsymbol{\nabla} \times \boldsymbol{\nabla}\Phi = 0$, the simultaneous transformations

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \to \Phi - \frac{\partial \chi}{\partial t} \tag{16.39}$$

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 (16.39) are termed (classical) *gauge transformations*. This freedom of gauge transformation may be used to decouple Eqs. (16.37)–(16.38). For example, if a set of potentials $(\boldsymbol{A}, \Phi)$ that satisfy

$$\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{\partial \Phi}{\partial t} = 0, \tag{16.40}$$

is chosen, the equations decouple to yield

$$\boldsymbol{\nabla}^2 \Phi - \frac{\partial^2 \Phi}{\partial t^2} = -\rho \qquad \boldsymbol{\nabla}^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\boldsymbol{J}, \tag{16.41}$$

which may be solved independently for $\boldsymbol{A}$ and Φ.

A constraint of the sort (16.40) 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 (16.41) is termed the *Lorentz gauge*. Another common gauge is the *Coulomb gauge*, with a gauge-fixing condition

$$\boldsymbol{\nabla} \cdot \boldsymbol{A} = 0, \tag{16.42}$$

which leads to the equations

$$\boldsymbol{\nabla}^2\Phi = -\rho \qquad \boldsymbol{\nabla}^2\boldsymbol{A} - \frac{\partial^2\boldsymbol{A}}{\partial^2 t} = \boldsymbol{\nabla}\frac{\partial\Phi}{\partial t} - \boldsymbol{J}. \tag{16.43}$$

Let us utilize the shorthand notation for derivatives introduced in Eq. (15.89):

$$\begin{aligned} \partial^\mu &\equiv \frac{\partial}{\partial x_\mu} = (\partial^0, \partial^1, \partial^2, \partial^3) = \left(-\frac{\partial}{\partial x^0}, \boldsymbol{\nabla}\right), \\ \partial_\mu &\equiv \frac{\partial}{\partial x^\mu} = (\partial_0, \partial_1, \partial_2, \partial_3) = \left(\frac{\partial}{\partial x^0}, \boldsymbol{\nabla}\right), \end{aligned} \tag{16.44}$$

where, for example, $\partial^1 = \partial/\partial x_1$ and

$$\boldsymbol{\nabla} \equiv (\partial^1, \partial^2, \partial^3) \tag{16.45}$$

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

$$A^\mu \equiv (\Phi, \boldsymbol{A}) = (A^0, \boldsymbol{A}) \qquad J^\mu \equiv (\rho, \boldsymbol{J}) \qquad \Box \equiv \partial_\mu\partial^\mu. \tag{16.46}$$

Thus the time-like component of the 4-vector potential A^μ is the scalar potential Φ and the spacelike components are the 3-vector potential $\boldsymbol{A}$, while the timelike component of the 4-current J^μ is the charge density $\rho$and the spacelike components are the 3-vector current $\boldsymbol{J}$. Then a gauge transformation takes the form

$$A^\mu \to A^\mu - \partial^\mu\chi \equiv A'^\mu \tag{16.47}$$

and the preceding examples of gauge-fixing constraints become

$$\partial_\mu A^\mu = 0 \quad \text{(Lorentz gauge)} \qquad \boldsymbol{\nabla}\cdot\boldsymbol{A} = 0 \quad \text{(Coulomb gauge)}, \tag{16.48}$$

with the notation in Eq. (16.48) indicating explicitly that

1. the *Lorentz condition is a covariant constraint* because it is formulated in terms of 4-vectors, but
2. the *Coulomb gauge condition is not covariant* because it is formulated in terms of only three of the components of a 4-vector.

The d'Alembertian operator $\Box$ is Lorentz invariant because under Lorentz transformations between coordinate systems,

$$\Box' = \partial'_\mu\partial'^\mu = \Lambda_\mu{}^\nu\Lambda^\mu{}_\lambda\partial_\nu\partial^\lambda = \partial_\mu\partial^\mu = \Box.$$

Thus, the Lorentz-gauge wave equation may be expressed in the compact and manifestly covariant form

$$\Box A^\mu = J^\mu \tag{16.49}$$

and the continuity equation (1.3) becomes becomes

$$\partial_\mu J^\mu = 0, \tag{16.50}$$

when written in covariant form.

The covariance of the Maxwell wave equation (16.49) in the Lorenz gauge, coupled with the general gauge invariance of electromagnetism, ensures that *the Maxwell equations are covariant in all gauges*.

However, as was seen in our original formulation of he Maxwell equations Eq. (1.1) and in the example of the Coulomb gauge, this covariance may not be manifest for a particular choice of notation or gauge. In Section 16.8.4 we shall address this issue by constructing a manifestly (Lorentz) covariant form of the Maxwell equations.

16.8.4 Maxwell Equations in Manifestly Covariant Form

The Maxwell equations may be cast in a form that is manifestly covariant by appealing to Eqs. (16.34) 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*

$$F^{\mu\nu} = -F^{\nu\mu} = \partial^\mu A^\nu - \partial^\nu A^\mu, \tag{16.51}$$

which may be expressed in matrix form as

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & -B^3 & B^2 \\ E^2 & B^3 & 0 & -B^1 \\ E^3 & -B^2 & B^1 & 0 \end{pmatrix}. \tag{16.52}$$

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

$$\mathscr{F}^{\mu\nu} = \tfrac{1}{2}\varepsilon^{\mu\nu\gamma\delta}F_{\gamma\delta} = \begin{pmatrix} 0 & -B^1 & -B^2 & -B^3 \\ B^1 & 0 & E^3 & -E^2 \\ B^2 & -E^3 & 0 & E^1 \\ B^3 & E^2 & -E^1 & 0 \end{pmatrix}, \tag{16.53}$$

the Maxwell equations (16.33a) and (16.33d) may be written

$$\partial_\mu F^{\mu\nu} = J^\nu, \tag{16.54}$$

and the Maxwell equations (16.33b) and (16.33c) may be expressed as

$$\partial_\mu \mathscr{F}^{\mu\nu} = 0. \tag{16.55}$$

The Maxwell equations in this form are manifestly covariant because they are formulated exclusively in terms of Lorentz tensors.

Background and Further Reading

For an introduction to special relativity, see Kogut [23]. Special and general relativity are introduced at an accessible level in Guidry [16]. The use of linearized gravity in describing gravitational waves far from strong sources, and the similarity of the linearized Einstein equations to the Maxwell equations, is discussed in Ch. 22 of Ref. [16]. Manifest Lorentz covariance of the Maxwell equations is discussed in Jackson [19].

Problems

16.1 The primed axes (x', t') plotted in the coordinate system of the unprimed axes in Fig. 16.6 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. ***

16.2 If a Lorentz transformation matrix 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.

16.3 (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 (16.34), show that the Maxwell field tensor $F^{\mu\nu}$ has components given by Eqs. (16.51) and (16.52).

16.4 Show that the Maxwell equations written in the covariant form (16.54) and (16.55) are equivalent to the non-covariant form of the Maxwell equations given by Eqs. (16.33a)–(16.33d).

16.5 Two events in Minkowski space have a timelike separation. Use the invariance of the interval ds^2 to show that the time between the events as measured by any inertial observer is always greater than or equal to the proper time between the events. ***

17 Action Formulation of Classical Electromagnetism

In this chapter we will demonstrate that classical electromagnetism can be formulated in terms of a least-action principle such that the Lagrangian and Hamiltonian methods of classical mechanics can be applied to the electromagnetic field.

17.1 A Lagrangian Formulation of Electromagnetism

Since the starting point for a field theory is typically a Lagrangian from which the fields can be derived through a variational principle, it is useful write down a *Lagrangian density* (a Lagrangian evaluated at a single spacetime point) appropriate for the electromagnetic field.

17.1.1 Principle of Extremal Proper Time

The motion of free particles in Minkowski spacetime is governed by a simple principle: the worldline for free particles between points separated by timelike intervals extremizes the proper time. This is called the *principle of extremal proper time* and is illustrated in Fig. 17.1. The proper time between the points A and B in Fig. 17.1 can be computed from Eq. (16.5) as,

$$\tau_{AB} = \int_A^B (dt^2 - dx^2 - dy^2 - dz^2)^{1/2}. \tag{17.1}$$

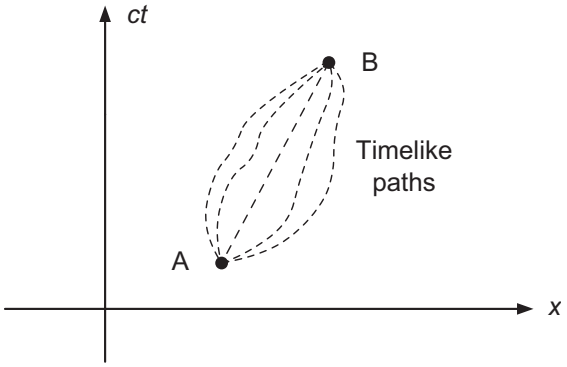


Fig. 17.1 *Principle of Extremal Proper Time:* of all the possible classical paths between two timelike-separated points A and B in spacetime, a free particle follows one that *extremizes the proper time*. Adapted from Ref. [16].

Parameterizing a path by a variable σ that varies continuously from 0 to 1 as the particle moves from A to B, the path is specified by $x^\mu = x^\mu(\sigma)$ and

$$\tau_{AB} = \int_0^1 \left[\left(\frac{dt}{d\sigma}\right)^2 - \left(\frac{dx}{d\sigma}\right)^2 - \left(\frac{dy}{d\sigma}\right)^2 - \left(\frac{dz}{d\sigma}\right)^2 \right]^{1/2} d\sigma. \tag{17.2}$$

The solution to finding the path that extremizes the action is described in Box 17.1, where the condition for an extremum is found to be

$$\delta \int d\tau = 0. \tag{17.3}$$

Defining a *Lagrangian* L as

$$L = \left(-g_{\mu\nu} \frac{dx^\mu}{d\sigma} \frac{dx^\nu}{d\sigma} \right)^{1/2}, \tag{17.4}$$

where $g_{\mu\nu}$ is the metric tensor, so that

$$\tau_{AB} = \int_0^1 L d\sigma, \tag{17.5}$$

then results in the *Euler–Lagrange equation of motion*

$$-\frac{d}{d\sigma}\left(\frac{\partial L}{\partial (dx^\mu/d\sigma)} \right) + \frac{\partial L}{\partial x^\mu} = 0. \tag{17.6}$$

Requiring that the Lagrangian L obey the differential equation (17.6) satisfies the variational condition (17.3).

17.1.2 Construction of Classical Fields

The standard approach to classical field theory is to define fields that obey appropriate equations of (classical) motion. The classical action S is defined by

$$S = \int d^4x \mathscr{L} = \int_{t_1}^{t_2} L dt, \tag{17.7}$$

where $\mathscr{L} = \mathscr{L}(x)$ is the Lagrangian defined at a Minkowski spacetime point (the *Lagrangian density*). The total Lagrangian L is then given by

$$L = \int d^3x \mathscr{L}(x). \tag{17.8}$$

We may view the action S as the central quantity for field quantization. It determines the equations of motion for the classical fields to be quantized in the canonical method, and in the path integral method each path in the functional integral is weighted by a factor $\exp(iS)$.

> The formal starting point for developing a field theory is to construct the appropriate Lagrangian density, which then leads to the action through Eq. (17.7), and thus to the equations of motion for the field (17.6).

Thus, we need appropriate Lagrangian densities for the fields of interest.

Box 17.1 The Euler–Lagrange Equations [16]

Consider the paths in spacetime between the fixed points labeled A and B in Fig. 17.1. For a function $L(x^\mu(\sigma), \dot{x}^\mu(\sigma))$, where σ parameterizes the path and $\dot{x}^\mu \equiv dx^\mu/d\sigma$, define the path integral

$$S = \int_A^B L(x^\mu(\sigma), \dot{x}^\mu(\sigma))\, d\sigma.$$

For an arbitrary small variation in the path $x^\mu(\sigma) \to x^\mu(\sigma) + \delta x^\mu(\sigma)$, the corresponding variation in the value of the integral is

$$\delta S \equiv \int_A^B \delta L d\sigma = \int_A^B \left(\frac{\partial L}{\partial \dot{x}^\mu(\sigma)}\, \delta \dot{x}^\mu(\sigma) + \frac{\partial L}{\partial x^\mu(\sigma)}\, \delta x^\mu(\sigma) \right) d\sigma.$$

The first term can be integrated by parts, giving

$$\delta S = \frac{\partial L}{\partial \dot{x}^\mu(\sigma)}\, \delta x^\mu(\sigma) \Big|_A^B + \int_A^B \left[-\frac{d}{d\sigma} \left(\frac{\partial L}{\partial \dot{x}^\mu(\sigma)} \right) + \frac{\partial L}{\partial x^\mu(\sigma)} \right] \delta x^\mu(\sigma) d\sigma,$$

but the variation $\delta x^\mu(\sigma)$ vanishes at the fixed endpoints so the first term is zero and

$$\delta S = \int_A^B \left[-\frac{d}{d\sigma} \left(\frac{\partial L}{\partial \dot{x}^\mu(\sigma)} \right) + \frac{\partial L}{\partial x^\mu(\sigma)} \right] \delta x^\mu(\sigma) d\sigma.$$

For paths that extremize the integral $\delta S = 0$ and for arbitrary variations $\delta x^\mu(\sigma)$ this will be true only if the expression inside the square brackets of the above integrand vanishes. Thus

$$-\frac{d}{d\sigma} \left(\frac{\partial L}{\partial \dot{x}^\mu(\sigma)} \right) + \frac{\partial L}{\partial x^\mu(\sigma)} = 0,$$

which is the *Euler–Lagrange equation* (17.6). Satisfying the variational condition $\delta S = 0$ characterizing an extremal path is equivalent to requiring that the function L satisfy the Euler–Lagrange equation.

17.1.3 Lagrangian Densities for Photons and Charged Fields

The Lagrangian densities that we require are constructed precisely to yield the appropriate fields when the variational principle (17.3) is applied. Let's summarize a few important wave equations entering into formulations of a field theory for electromagnetism and for charged particles coupled to electromagnetism.

1. The *Maxwell wave equation* for massless vector fields $A^\mu(x)$ is, in Lorentz gauge,

$$\Box A^\mu(x) = j^\mu \qquad \text{(Lorenz gauge)}, \tag{17.9}$$

which is quadratic in ∂^μ.

2. The *Klein–Gordon equation* for scalar (single-component) fields $\Phi(x)$

$$(\Box + m^2)\Phi(x) = 0 \qquad \Box \equiv \partial_\mu \partial^\mu, \tag{17.10}$$

   which is quadratic in ∂_μ and carries no Lorentz index.
3. The *Dirac equation* for 4-component Lorentz spinor fields $\Psi(x)$

$$(-i\gamma^\mu \partial_\mu + m)\Psi(x) = 0, \tag{17.11}$$

   which is linear in ∂^μ.
4. The *massive vector* or *Proca field* $A^\mu(x)$, with an equation

$$(\Box + m^2)A^\mu(x) = 0 \tag{17.12}$$

that is quadratic in ∂^μ.

The corresponding Lagrangian densities that yield the wave equations given above are summarized in the following.

Massless Vector Field: In terms of the rank-2 *electromagnetic field tensor* $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$ introduced in Eq. (16.51) through the 4-vector potential A^μ defined in Eq. (16.46), the free-field Lagrangian density appropriate for the massless vector (photon) field is

$$\mathscr{L}_0 = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \qquad \text{(Massless vector field).} \tag{17.13}$$

This Lagrangian density describes massless vector particles like the photon and satisfies the wave equation (17.9) in Lorentz gauge.

Complex Scalar or Pseudoscalar Fields: A complex scalar (or pseudoscalar) field $\phi(x)$ may be expressed in terms of two real scalar fields $\phi_1(x)$ and $\phi_2(x)$ having the same mass ($m_1 = m_2 \equiv m$),

$$\phi = \frac{1}{\sqrt{2}}(\phi_1 + i\phi_2) \qquad \phi^\dagger = \frac{1}{\sqrt{2}}(\phi_1 - i\phi_2), \tag{17.14}$$

where the corresponding free Lagrangian density is

$$\mathscr{L}_0 = (\partial^\mu \phi)^\dagger (\partial_\mu \phi) - m^2 \phi^\dagger \phi \qquad \text{(Complex scalar field).} \tag{17.15}$$

Separate variations of ϕ and $\phi^\dagger$ then lead to two independent Klein–Gordon equations (17.10),

$$(\Box + m^2)\phi(x) = 0 \qquad (\Box + m^2)\phi^\dagger(x) = 0. \tag{17.16}$$

Charged scalar or pseudoscalar particles are associated with complex scalar fields.

Dirac Spinor Field: For a Dirac field $\psi(x)$ and a conjugate (adjoint spinor) field $\overline{\psi}(x)$ defined by

$$\overline{\psi}(x) \equiv \psi^\dagger(x)\gamma^0, \tag{17.17}$$

where the matrix γ_0 is defined through the 4×4 matrices γ^μ satisfying the anticommutator

$$\{\gamma_\mu, \gamma_\nu\} \equiv \gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2\eta_{\mu\nu} \tag{17.18}$$

Table 17.1 Properties of the Dirac γ-matrices [17]

$$\{\alpha_i,\alpha_j\}=2\delta_{ij} \qquad \{\alpha_i,\beta\}=0 \qquad \alpha_i^2=\beta^2=1$$

$$\gamma_0=\gamma^0\equiv\beta \qquad \gamma_i\equiv\gamma^0\alpha_i \qquad \gamma_5=\gamma^5\equiv i\gamma^0\gamma^1\gamma^2\gamma^3 \qquad \sigma^{\mu\nu}\equiv\tfrac{i}{2}[\gamma^\mu,\gamma^\nu]$$

$$\{\gamma^\mu,\gamma^\nu\}=2\eta^{\mu\nu} \qquad \gamma_0\gamma_\mu^\dagger\gamma_0=\gamma_\mu \qquad \{\gamma_5,\gamma^\mu\}=0 \qquad (\gamma^i)^2=-1$$

$$(\gamma^0)^2=1 \qquad (\gamma^5)^2=1 \qquad (\gamma_0)^\dagger=\gamma_0 \qquad (\gamma^i)^\dagger=-\gamma^i \qquad (\gamma^5)^\dagger=\gamma^5$$

Pauli–Dirac representation (4×4 matrices):

$$\gamma^0=\begin{pmatrix}1&0\\0&-1\end{pmatrix} \quad \gamma^i=\begin{pmatrix}0&\sigma_i\\-\sigma_i&0\end{pmatrix} \quad \gamma^5=\begin{pmatrix}0&1\\1&0\end{pmatrix} \quad \alpha_i=\begin{pmatrix}0&\sigma_i\\\sigma_i&0\end{pmatrix}$$

$$\beta=\begin{pmatrix}1&0\\0&-1\end{pmatrix} \qquad \sigma^{ij}=\varepsilon_{ijk}\begin{pmatrix}\sigma_k&0\\0&\sigma_k\end{pmatrix} \qquad \sigma^{0k}=i\begin{pmatrix}0&\sigma_k\\\sigma_k&0\end{pmatrix}$$

Weyl (chiral) representation (4×4 matrices):

$$\gamma^0=\begin{pmatrix}0&1\\1&0\end{pmatrix} \qquad \gamma^i=\begin{pmatrix}0&-\sigma_i\\\sigma_i&0\end{pmatrix} \qquad \gamma^5=\begin{pmatrix}1&0\\0&-1\end{pmatrix}$$

Majorana representation (4×4 matrices):

$$\gamma^0=\begin{pmatrix}0&\sigma_2\\\sigma_2&0\end{pmatrix} \quad \gamma^1=\begin{pmatrix}i\sigma_1&0\\0&i\sigma_1\end{pmatrix} \quad \gamma^2=\begin{pmatrix}0&\sigma_2\\-\sigma_2&0\end{pmatrix}$$

$$\gamma^3=\begin{pmatrix}-i\sigma_3&0\\0&i\sigma^3\end{pmatrix} \qquad \gamma^5=\begin{pmatrix}\sigma_2&0\\0&\sigma_2\end{pmatrix}$$

Standard Pauli matrices (2×2 matrices):

$$\sigma_1=\begin{pmatrix}0&1\\1&0\end{pmatrix} \qquad \sigma_2=\begin{pmatrix}0&-i\\i&0\end{pmatrix} \qquad \sigma_3=\begin{pmatrix}1&0\\0&-1\end{pmatrix}$$

that are termed the (Dirac) γ-*matrices*. Explicit forms in different representations and general properties of the gamma-matrices are summarized in Table 17.1. The Lagrangian density for a free Dirac spinor field is

$$\mathscr{L}_0=\overline{\psi}(i\partial\!\!\!/-m)\psi \qquad \text{(Dirac spinor field)}, \tag{17.19}$$

where m is the mass and *Feynman slash notation* has been used: a slash through a quantity indicates a 4-vector contracted completely with the γ-matrices: $\partial\!\!\!/\equiv\gamma^\mu\partial_\mu$. The Lagrangian density (17.19) then yields the Dirac equation (17.11),

$$(i\partial\!\!\!/-m)\psi=0, \tag{17.20}$$

as the equation of motion for a charged fermion spinor field.

Massive Vector Field: For a vector field of finite mass m, the appropriate free-field La-

grangian density is obtained from Eq. (17.13) by adding a Lorentz-invariant mass term,

$$\mathscr{L}_0 = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}m^2 A^\mu A_\mu \qquad \text{(Massive vector field).} \tag{17.21}$$

This implies an equation of motion $(\Box + m^2)A^\mu = 0$ corresponding to the wave equation (17.12) for a massive vector field.

Background and Further Reading

See Wald [40], Zangwill [42], and Garg [11].

Problems

17.1 *No problems yet for this chapter.*

18 Electromagnetism as an Abelian Gauge Field Theory

As we have alluded to in various places, modern relativisitic quantum field theory, in particular the Standard Model of elementary particle physics, is a quantized generalization of the gauge invariance principle of classical electromagnetism. In this chapter we provide some remarks intended to outline the relationship of classical gauge invariance to the more complex gauge symmetries prominent in modern gauge field theories. These more complex gauge symmetries are typically implemented in (usually relativistic) quantum field theories, so our discussion will occasionally touch on a quantum description. Therefore, the reader will be assumed to have an understanding of basic concepts from quantum mechanics.

18.1 Symmetry and Gauge Invariance

Symmetries and broken symmetries are of fundamental importance in the generalization of the classical gauge invariance exhibited by the Maxwell equations to the varied manifestation of gauge invariance in modern quantum field theory. The natural language for describing gauge symmetries in their most general form is the theory of Lie groups and Lie algebras. The rudiments are described in Box 18.1. The essential difference between the gauge symmetry of electromagnetism and the gauge symmetries associated with the electromagnetic, weak, and strong interactions in the Standard Model is that electromagnetism is governed by a symmetry called U(1) (described in Box 18.2) that is *abelian*, but the weak and strong interactions are governed by gauge symmetries that correspond to larger and more complex *nonabelian* gauge groups.

18.2 The Minimal Coupling Prescription

Gauge invariance is of fundamental importance in classical electromagnetism and becomes even more so in quantum field theory. In either case it is important to ask what coupling terms between fields in Lagrangians or Hamiltonians are permitted that preserve gauge invariance. Note first that a classical particle carrying charge q and moving with 3-velocity $\boldsymbol{v}$ in an electromagnetic field experiences the *Lorentz force* of classical electromagnetism,

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}), \tag{18.1}$$

Box 18.1 **Symmetries and Groups [16]**

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*).[a] Because boost velocities can approach c asymptotically but never reach it, the Lorentz-group parameter space is bounded by $v < c$ but not closed (the limit $v = c$ is not part of the set since Lorentz transformations diverge for $v = c$) 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).

[a] A closed set contains all its limit points. In a bounded set all point lie within some maximum fixed distance from each other. Example: $[0, a)$ (where "$[$" denotes a closed interval and "$)$" an open interval) is not a closed subset of the real numbers $\mathbb{R}$ because it contains 0 but doesn't contain a.

Box 18.2 The Gauge Group U(1)

A gauge transformation in electromagnetism corresponds to simultaneous transformations on the scalar potential Φ and vector potential A_i,

$$\Phi \to \Phi' = e^{iq\chi}\Phi \qquad A_i \to A_i' = A_i + \partial_\mu \chi$$

(in $\hbar = 1$ units), where q is the electrical charge and $\chi = \chi(x)$ is an arbitrary scalar function depending on the spacetime coordinates x. Gauge symmetry then implies that classically (Φ, A_i) and (Φ', A_i') are *physically equivalent*.[a] Thus, the gauge transformation acts on the charged scalar field Φ by multiplying it by a spacetime dependent phase factor. Mathematically, multiplication by a phase factor corresponds to a unitary map $U\colon \mathbb{C} \to \mathbb{C}$ on the complex numbers $z \in \mathbb{C}$ that is defined in this case by $Uz = e^{iq\chi}z$, where U implements a *unitary transformation*.[b] The set of all unitary maps on $\mathbb{C}$ constitute a group (see Box 18.1) with the "multiplication" operation defined by composition (successive application) of maps. This group is commonly labeled U(1) in physics, where U denotes "unitary" and 1 denotes a 1D complex vector space $\mathbb{C}$. Since the complex vector space is one-dimensional, the group elements $e^{iq\chi}$ of U(1) are parameterized by a single parameter χ and the group is abelian. The unitary maps $Uz = e^{iq\chi}z$ are in one-to-one correspondence with the real numbers χ modulo $2\pi/q$, so the manifold of U(1) (the space traced out by allowing the group parameter χ to vary over all possible values) is represented naturally by a circle S^1 of circumference $2\pi/q$.

[a] They lead to the same classical electric and magnetic fields, and thus to the same solutions of the Maxwell equations. Since $\chi(x)$ is a local function of the coordinates, this is termed a *local gauge transformation* (a *global gauge transformation* corresponds to a constant χ). As discussed in Section 11.3 and Box 11.2, gauges related by the same gauge transformation constitute an *equivalence class*.

[b] A unitary transformation Uz is a linear map with the property $|Uz^2| = |z^2|$), so it preserves the modulus $|z^2|$ for all $z \in \mathbb{C}$). Such transformations are said to be *unimodular*.

provided that the classical Hamiltonian function is given by

$$H = \frac{1}{2m}(\boldsymbol{p} - q\boldsymbol{A})^2 + q\phi, \tag{18.2}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, $\boldsymbol{p}$ is the 3-momentum, $\boldsymbol{A}$ is the 3-vector potential, and ϕ is the scalar potential. Quantizing this Hamiltonian assuming nonrelativistic quantum mechanics gives the modified Schrödinger equation

$$\left[\frac{1}{2m}(-i\boldsymbol{\nabla} - q\boldsymbol{A})^2 + q\phi\right]\psi(\boldsymbol{x},t) = i\frac{\partial\psi(\boldsymbol{x},t)}{\partial t}. \tag{18.3}$$

Comparing this expression with the free-particle Schrödinger equation

$$-\frac{1}{2m}\boldsymbol{\nabla}^2\psi(\boldsymbol{x},t) = i\frac{\partial\psi(\boldsymbol{x},t)}{\partial t},$$

suggests that a Schrödinger equation accounting for gauge-invariant coupling of charged particles to an electromagnetic field can be constructed from the free-particle Schrödinger equation by making the operator substitutions

$$\boldsymbol{\nabla} \to \boldsymbol{D} \equiv \boldsymbol{\nabla} - iq\boldsymbol{A} \qquad \frac{\partial}{\partial t} \to D^0 \equiv \frac{\partial}{\partial t} + iq\phi. \tag{18.4}$$

In covariant notation this becomes

$$\partial^\mu \to D^\mu \equiv \partial^\mu + iqA^\mu, \tag{18.5}$$

where we have defined the 4-vectors $A^\mu = (\phi, \boldsymbol{A})$ and $D^\mu = (D^0, \boldsymbol{D})$. The replacement (18.5) is just the replacement of the partial derivative by the covariant derivative as in Eq. (15.95), with the notational change $\boldsymbol{\nabla}_\mu \to D^\mu$. Equation (18.5) is termed the *minimal coupling prescription* or the *minimal substitution*. The Schrödinger equation (18.3) is invariant under the *local gauge transformation* (11.10),

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \to \Phi - \frac{\partial \chi}{\partial t}, \tag{18.6}$$

provided that the wavefunction is transformed simultaneously as

$$\psi(\boldsymbol{x}, t) \to e^{iq\chi(x)}\psi(\boldsymbol{x}, t), \tag{18.7}$$

where χ is an arbitrary scalar function that is a *local function of the spacetime coordinates*.

For completeness we note that for relativisitic wave equations, minimal substitution in the Klein–Gordon equation (17.10) (appropriate for charged scalar or pseudoscalar particles if Φ is complex) gives the wave equation

$$\left(D^\mu D_\mu + m^2\right)\psi(x) = 0, \tag{18.8}$$

while minimal substitution in the Dirac equation (17.11) (appropriate for charged fermions if Ψ is complex) gives

$$\left(i\gamma_\mu D^\mu - m\right)\psi(x) = 0. \tag{18.9}$$

Both Eqs. (18.8) and (18.9) are invariant under a gauge transformation

$$A^\mu \to A^\mu - \partial^\mu \chi(x) \qquad \psi \to e^{iq\chi(x)}\psi, \tag{18.10}$$

where we note again that these are *local gauge transformations* since χ is a function of the spacetime coordinate x. Summarizing:

Minimal Substitution: In a free charged-particle wave equation, replace all derivatives with covariant derivatives

$$\partial^\mu \to D^\mu \equiv \partial^\mu + iqA^\mu.$$

The resulting wave equation couples the electromagnetic and charged-particle fields and is *invariant under a local gauge transformation*

$$A^\mu \to A^\mu - \partial^\mu \chi(x) \qquad \psi \to e^{iq\chi(x)}\psi,$$

where q is charge, A^μ is the 4-vector potential, and $\chi(x)$ is a *local* scalar field.

This is called *minimal substitution* or the *minimal coupling prescription* because it, and not a more complicated coupling, is adequate for implementing gauge-invariant coupling of charged particles to the electromagnetic field.

18.3 Gauge Invariance and the Photon Mass

The Lagrangian density for a *massless* vector field is given by Eq. (17.13). Since $F^{\mu\nu}F_{\mu\nu}$ is gauge invariant, the massless vector field (photon field) is gauge invariant. A *massive* vector field has a Lagrangian density given by Eq. (17.21). The first term for (17.21) is gauge invariant but the second (mass) term is not, since under the gauge transformation $A^\mu \to A^\mu - \partial^\mu \chi$,

$$A^\mu A_\mu \to (A^\mu - \partial^\mu \chi)(A_\mu - \partial_\mu \chi) \neq A^\mu A_\mu .$$

We reach a conclusion that has far-reaching implications for gauge symmetry and its influence on modern physics:

> Gauge invariance of the electromagnetic field is tied directly both to charge conservation and to masslessness of the vector boson (the photon) that mediates the electromagnetic interaction. If the photon had a finite restmass, the corresponding vector field would break gauge symmetry.

That local gauge symmetry implies massless gauge bosons is crucial in understanding abelian and non-abelian gauge theories in the Standard Model. However, that is beyond the scope of our discussion here.

18.4 Conserved Currents and Charges

Observables in quantum field theory are often formulated through generalizations of electromagnetic charges and currents. Field-theory symmetries may be expressed in terms of conservation of those charges and currents.

18.4.1 Noether's Theorem

Conservation laws for field theories formulated in terms of Lagrangian densities often are expressed in terms of *Noether's theorem:*

> ***Noether's Theorem:*** For each continuous symmetry of a field theory Lagrangian there is a corresponding conserved quantity.

Conserved quantities may involve spacetime or internal symmetries and are termed *conserved Noether tensors.* Let's illustrate by deriving conservation of 4-momentum from invariance under spacetime translations.

18.4.2 Conservation of Energy and Momentum

This solution follows a discussion in Quigg [33]. Poincaré invariance[1] requires that the action be unchanged by a spacetime transformation $x_\mu \to x'_\mu = x_\mu + a_\mu$, where the infinitesimal displacement a_μ is independent of x_μ. The corresponding change in a scalar field Lagrangian density is

$$\delta\mathscr{L} = \mathscr{L}(x') - \mathscr{L}(x) = a^\mu \partial_\mu \mathscr{L}. \tag{18.11}$$

Assuming the Lagrangian density to be translationally invariant in spacetime (independent of spacetime coordinates x), but to depend on the field ϕ and $\partial_\mu\phi$, we may also write

$$\begin{gathered}\delta\mathscr{L} = \frac{\partial\mathscr{L}}{\partial\phi}\delta\phi + \frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi)}\delta(\partial_\mu\phi) \qquad \delta\phi \equiv \phi(x') - \phi(x) = a^\mu\partial_\mu\phi, \\ \delta(\partial_\mu\phi) \equiv \partial_\mu\phi(x') - \partial_\mu\phi(x) = a^\nu\partial_\nu(\partial_\mu\phi).\end{gathered} \tag{18.12}$$

But from the Euler–Lagrange equations (17.6),

$$\frac{\partial\mathscr{L}}{\partial\phi} = \partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi)}\right), \tag{18.13}$$

and from Eqs. (18.13) and (18.12),

$$\delta\mathscr{L} = \left(\partial_\nu\frac{\partial\mathscr{L}}{\partial(\partial_\nu\phi)}\right)a^\mu\partial_\mu\phi + \frac{\partial\mathscr{L}}{\partial(\partial_\nu\phi)}a^\mu\partial_\mu(\partial_\nu\phi) = \partial_\nu\frac{\partial\mathscr{L}}{\partial(\partial_\nu\phi)}a^\mu\partial_\mu\phi,$$

where the last step may be checked by taking the derivative of the product. Equating this expression with Eq. (18.11) and rearranging (remember that repeated indices are dummy indices that can be changed without affecting the validity of the equation) leads to

$$a_\nu\partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi)}\partial^\nu\phi - \eta^{\mu\nu}\mathscr{L}\right) = 0,$$

where $\eta^{\mu\nu}$ is the metric tensor of flat spacetime. This relation must evaluate to zero for arbitrary displacements a_ν, which implies that

$$\partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi)}\partial^\nu\phi - \eta^{\mu\nu}\mathscr{L}\right) = 0.$$

Let us define the *stress–energy–momentum tensor*

$$\Theta^{\mu\nu} \equiv \frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi)}\partial^\nu\phi - \eta^{\mu\nu}\mathscr{L}, \tag{18.14}$$

[1] The Poincaré group appends the generators of spacetime translations to the Lorentz transformations; see Box 18.1. Thus Poincaré symmetry means that the system is invariant under Lorentz transformations (which preserve the invariant spacetime interval) and also under uniform translations in spacetime (which implies conservation of energy and momentum).

for which it may be shown that Θ^{00} is the energy density and the Θ^{0k} are momentum densities. This leads to a local conservation law

$$\partial_\mu \Theta^{\mu\nu}(x) = 0,$$

for energy and momentum. The quantity $\Theta^{\mu\nu}$ is an example of a conserved Noether tensor that in this case links 4-momentum conservation with spacetime translational invariance. In this example the symmetry is a spacetime symmetry, but probably the most important applications of Noether's theorem are to conserved charges and currents associated with internal degrees of freedom for elementary particles, particularly the gauge degrees of freedom.

18.4.3 Internal Noether Currents and Charges

For an internal symmetry with a symmetry group described by a set of matrix generators T^a and a Lagrangian density $\mathscr{L}$, the *conserved Noether currents* J^a_μ are given by

$$J^a_\mu = -i\,\frac{\partial \mathscr{L}}{\partial(\partial^\mu \phi_i)} T^a_{ij}\phi_j, \tag{18.15}$$

where a labels an internal symmetry generator, μ is a spacetime index, and i and j label entries in the matrix T^a. Given a local 4-current

$$J^\mu = \left(J^0(x), J^1(x), J^2(x), J^3(x)\right),$$

as in Eq. (18.15), an associated charge $Q(t)$ may be defined by

$$Q(t) \equiv \int J^0(x)\, d^3x. \tag{18.16}$$

For example, the electromagnetic 4-current $j = (\rho, \boldsymbol{j})$ implies an electrical charge

$$Q_{\rm e} = \int \rho(x)\, d^3x = \int j^0(x)\, d^3x. \tag{18.17}$$

The charge operator satisfies the *Heisenberg equation of motion* for quantum operators,

$$\dot{Q} \equiv \frac{dQ(t)}{dt} = i\,[H, Q(t)], \tag{18.18}$$

where H is the Hamiltonian operator and the commutator of operators A and B is denoted by $[A,B] \equiv AB - BA$. If J^μ has vanishing 4-divergence,

$$\partial_\mu J^\mu \equiv \partial_0 J^0 + \boldsymbol{\nabla}\cdot\boldsymbol{J} = \partial_0 J^0 + \partial_k J^k = 0. \tag{18.19}$$

Then $[H,Q] = 0$ and Q is a constant of motion.

Conserved Charges: If a 4-current $J^\mu(x)$ satisfies $\partial_\mu J^\mu = 0$, the charge Q defined by the spatial integral of $J^0(x)$ is conserved.

This establishes a formal link between conserved Noether currents in a Lagrangian formulation of a problem (which have vanishing 4-divergence) and conserved charges that are constants of motion in a Hamiltonian formulation of the same problem (the charge operator commutes with the Hamiltonian).

Example 18.1 The electromagnetic 4-current $j^\mu = (\rho, \boldsymbol{j})$ has $\partial_\mu j^\mu = 0$, implying that the electrical charge Q_e of Eq. (18.17) is conserved.

Example 18.1 is well known for electrical charges and currents. However, the preceding derivation implies that *any charge* defined through Eq. (18.16) is conserved (it is a constant of motion) if the 4-divergence of the associated current J^μ vanishes. In the generalizations of the gauge symmetry of classical electromagnetism that underly the Standard Model of elementary particle physics, one encounters gauge charges and associated currents that generalize the electromagnetic charge and current. The U(1) symmetry of electromagnetism (see Box 18.2) has generators that can be represented as 1×1 matrices and the symmetry is abelian. For Standard Model gauge symmetries many of the symmetry groups are non-abelian and the matrices entering Eq. (18.15) are $n \times n$ matrices with $n > 1$.

18.5 The Aharonov–Bohm Effect

The Aharonov–Bohm effect [1] demonstrates that the electromagnetic 4-vector potential can have measurable consequences in a quantum theory, even for regions having no electric or magnetic fields. This contrasts with classical electromagnetism, where electric and magnetic fields are fundamental and vector potentials are only a mathematical convenience.

18.5.1 Experimental Setup

An experimental setup to test the Aharonov–Bohm effect is illustrated in Fig. 18.1(a). It is a two-slit, electron-scattering interference experiment, with a long, thin solenoid is placed behind the screen and between the two classical paths that electrons passing through the slits would follow to reach the screen. The solenoid confines the magnetic field to regions that that classically the electrons cannot pass through, but experiments show that the solenoid alters the interference pattern for the electrons scattering through the two slits [6]. We now show that this results from the gauge symmetry of the electromagnetic field and the non-trivial topology of the scattering plane due to the solenoid [35].

18.5.2 Analysis of Magnetic Fields

From Fig. 18.1(a), the phase difference between waves of de Broglie wavelength λ associated with paths 1 and 2 is $\delta = 2\pi a/\lambda$, implying that $x \simeq L\lambda\delta/2\pi d$, where we've assumed

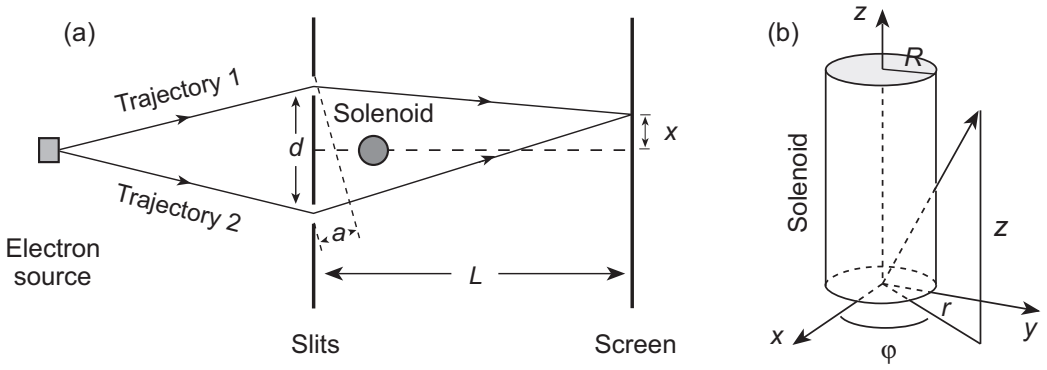


Fig. 18.1 (a) Aharonov–Bohm experiment. (b) Cylindrical coordinates for the Aharonov–Bohm effect.

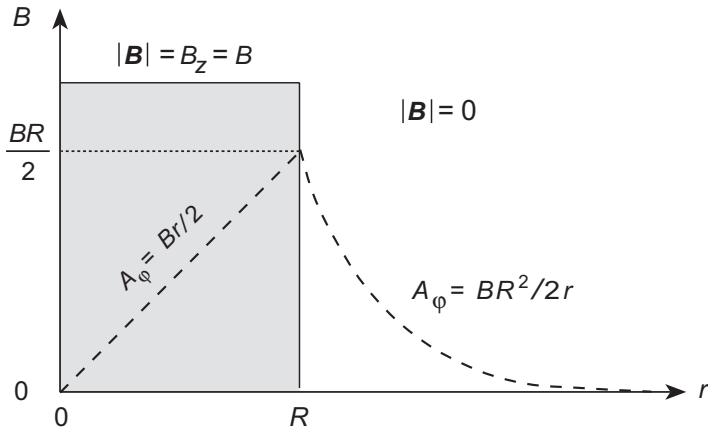


Fig. 18.2 Magnetic field B and vector potential A_ϕ for a solenoid of radius R in Fig. 18.1(a).

$a \simeq (x/L)d$ for $x \ll L$. In the cylindrical coordinate system of Fig. 18.1(b) the vector potential takes the form

$$\begin{aligned} &\text{Inside solenoid:} \quad A_r = A_z = 0 \qquad A_\phi = \frac{Br}{2}, \\ &\text{Outside solenoid:} \quad A_r = A_z = 0 \qquad A_\phi = \frac{BR^2}{2r}, \end{aligned} \tag{18.20}$$

so that the components of the magnetic field $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$ are

$$\begin{aligned} &\text{Inside solenoid:} \quad B_r = 0 \qquad B_\phi = 0 \qquad B_z = B, \\ &\text{Outside solenoid:} \quad B_r = 0 \qquad B_\phi = 0 \qquad B_z = 0. \end{aligned} \tag{18.21}$$

Outside the solenoid the magnetic field $\boldsymbol{B}$ vanishes but the vector potential A_ϕ remains finite, falling off as $1/r$ for $r > R$ (Fig. 18.2). Experimentally, one can use a very thin solenoid and shielding to ensure that the electrons never enter the solenoid and therefore never encounter a finite magnetic field. Classically, a particle of charge q in an electromagnetic field experiences a Lorentz force, $\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$, where $\boldsymbol{v}$ the velocity, $\boldsymbol{E}$ the electric field and $\boldsymbol{B}$ the magnetic field. Thus classically electrons experience no Lorentz

force in the Aharonov–Bohm experiment. However, the electrons pass through regions where vector potential is finite, which has non-trivial implications in quantum theory.

18.5.3 Phase of the Electron Wavefunction

The free-electron wavefunction is $\psi = |\psi| e^{i\alpha}$, where $\alpha \equiv \boldsymbol{p}\cdot\boldsymbol{r}/\hbar$. Gauge invariance requires the minimal substitution $\boldsymbol{p} \to \boldsymbol{p} - e\boldsymbol{A}$, where $\boldsymbol{A}$ is the 3-vector potential, altering the electron phase to,

$$\alpha \to \frac{1}{\hbar}(\boldsymbol{p}\cdot\boldsymbol{r} - e\boldsymbol{A}\cdot\boldsymbol{r}) = \alpha - \frac{e}{\hbar}\boldsymbol{A}\cdot\boldsymbol{r}.$$

Then over a trajectory τ the change in phase is

$$\Delta\alpha = -\frac{e}{\hbar}\int_\tau \boldsymbol{A}\cdot d\boldsymbol{r}, \tag{18.22}$$

and the shift in phase between trajectories 1 and 2 is

$$\Delta\delta = \frac{e}{\hbar}\left(\int_1 \boldsymbol{A}\cdot d\boldsymbol{r} - \int_2 \boldsymbol{A}\cdot d\boldsymbol{r}\right) = \frac{e}{\hbar}\oint \boldsymbol{A}\cdot d\boldsymbol{r}. \tag{18.23}$$

Now apply Stokes' theorem:

$$\oint_C \boldsymbol{A}\cdot d\boldsymbol{r} = \int_S (\boldsymbol{\nabla}\times\boldsymbol{A})\cdot\boldsymbol{n}\,dS \equiv \int_S (\boldsymbol{\nabla}\times\boldsymbol{A})\cdot d\boldsymbol{S}, \tag{18.24}$$

where $\boldsymbol{S}$ is the surface enclosed by the contour C and $\boldsymbol{n}$ is an outward normal to the surface, giving

$$\Delta\delta = \frac{e}{\hbar}\Phi \qquad \Phi \equiv \pi R^2 B, \tag{18.25}$$

where Φ is the magnetic flux through the solenoid. Thus, the solenoid shifts the interference pattern of Fig. 18.1(a) by

$$\Delta x = \frac{L\lambda}{2\pi d}\Delta\delta = \frac{L\lambda}{2\pi d}\frac{e}{\hbar}\Phi, \tag{18.26}$$

even though the electrons never encounter the magnetic field $\boldsymbol{B}$. This predicted shift has been verified experimentally.

18.5.4 Topological Origin of the Aharonov–Bohm Effect

The magnetic field is related to the vector potential by $\boldsymbol{B} = \boldsymbol{\nabla}\times\boldsymbol{A}$. Outside the solenoid there is an electromagnetic vacuum with $\boldsymbol{B} = 0$. This is satisfied by the obvious choice $\boldsymbol{A} = 0$, but by gauge invariance $\boldsymbol{B}$ is invariant under the transformation $A_\mu \to A_\mu + \partial_\mu\chi \equiv A'_\mu$, where χ is an arbitrary scalar function. Thus, a gauge transform of the trivial vacuum $A_\mu = 0$, which is given by $A_\mu = \partial_\mu\chi$, *also satisfies the boundary condition*. It follows that in the cylindrical coordinates of Fig. 18.1(b),

$$\boldsymbol{A} = (A_r, A_\phi, A_z) = \boldsymbol{\nabla}\chi \tag{18.27}$$

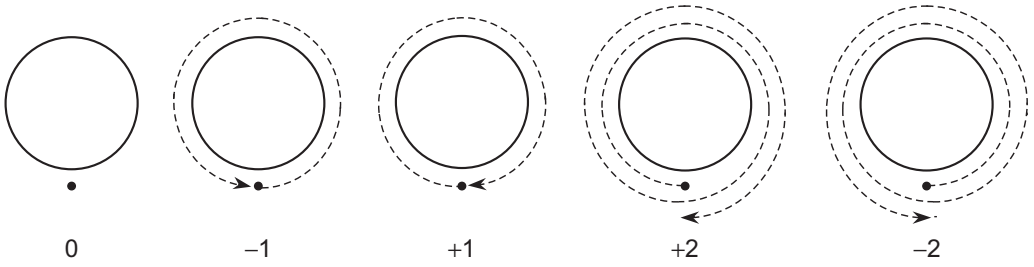


Fig. 18.3 Some winding numbers characterizing the topology of the U(1) manifold S^1. Intuitively the winding number is the number of times the dashed curve is wrapped around the solid circle, clockwise (+) or counterclockwise (−).

is the most general vector potential giving a vanishing electromagnetic field. The scalar function χ is then

$$\chi = \frac{1}{2} BR^2 \phi, \tag{18.28}$$

and χ increases by πBR^2 as ϕ increases by 2π and *isn't single-valued.* This is possible because *the vacuum configuration of the Aharonov–Bohm experiment isn't simply connected.*

From Fig. 18.1, the Aharonov–Bohm experiment corresponds to a plane pierced by a hole corresponding to the solenoid; topologically this is the product of the real numbers and a circle, $\mathbb{R} \times S^1$. As discussed in Box 18.2, the gauge group for electromagnetism is U(1), which has a manifold that is S^1 topologically. Thus, we must consider mappings between $\mathbb{R} \times S^1$ and S^1. But $\mathbb{R}$ is simply connected so $S^1 \to S^1$ is the only part of the mapping that matters topologically, and this mapping is characterized by an infinity of integer winding numbers giving how many times one circle is wrapped around the other, as illustrated in Fig. 18.3. Therefore, *the gauge function χ is multivalued* because the mappings are not all deformable to the trivial mapping $\chi =$ constant, which would give $A_\mu = 0$ and no Aharonov–Bohm effect, by virtue of Eq. (18.27).

18.6 Primacy of the Potentials in Electromagnetism

Two pieces of evidence that become completely clear only in a full quantum treatment of electromagnetism demonstrate that it is the magnetic vector potential potential $\boldsymbol{A}$ and the scalar potential Φ that are fundamental in electromagnetism, *not* the magnetic field $\boldsymbol{B}$ and the electric field $\boldsymbol{E}$:

1. The gauge invariant coupling of charged particles and the electromagnetic field can be implemented only through the *minimal coupling prescription* described in Section 18.2, which requires the 4-vector potential A_μ and cannot be formulated in terms of the electric and magnetic fields.
2. The *Aharonov–Bohm effect* described in Section 18.5 can be accounted for only through

the influence of the magnetic vector potential $\boldsymbol{A}$, not through the magnetic field $\boldsymbol{B}$, which the test particle never encounters.

This primacy of the scalar and vector potentials is contrary to what one would infer from classical electromagnetism alone, where the Maxwell equations and Lorentz force implementing the full scope of classical electromagnetism are defined in terms of the electric and magnetic fields, and the vector and scalar potentials seem to be only optional mathematical conveniences.

Background and Further Reading

Chapter 9 of Wald [40] elaborates on some of the themes of this chapter. Gauge field theories in elementary particle physics are described in Guidry [14] and gauge fields in modern quantum field theory are discussed extensively from the perspective of symmetry and broken symmetry in Guidry and Sun [17].

Problems

18.1 The Euler–Lagrange equation (17.6) is used in a field-theory context in this chapter, but it is applicable to a broad range of problems. Show that inserting a Lagrangian $L(x,\dot{x}) = \frac{1}{2}\dot{m}^2 - V(x)$ in Eq. (17.6) leads to Newton's 2nd law of motion.

18.2 Show that if Eq. (18.19) is true, the charge Q defined in Eq. (18.16) is conserved. *Hint*: Use the Heisenberg equation of motion (18.18).

18.3 Show that the Lagrangian density of Eq. (17.13) leads to the field equation (16.54) for a free field in electromagnetism.

18.4 Use Eqs. (17.19) and (17.13) to write the Lagrangian density corresponding to non-interacting Dirac and electromagnetic fields. Introduce a gauge-invariant coupling between the fields by the minimal substitution, $\partial_\mu \to \partial_\mu + ieA_\mu$, where e is the charge of the particle described by the Dirac field and A_μ is the vector potential field associated with electromagnetism, to give a Lagrangian density

$$\mathscr{L} = \overline{\psi}(i\partial\!\!\!/ - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - e\overline{\psi}A\!\!\!/\psi$$

(where $A\!\!\!/ \equiv \gamma_\mu A^\mu$ and $\partial\!\!\!/ \equiv \gamma_\mu \partial^\mu$), indicating that the minimal substitution has introduced an interaction term $\mathscr{L}_{\text{int}} = -e\overline{\psi}A\!\!\!/\psi$ between the fields. Show that independent variation of this Lagrangian density with respect to the fields leads to the coupled equations of motion

$$(i\partial\!\!\!/ - m)\psi(x) = eA\!\!\!/(x)\psi(x),$$
$$\partial_\nu F^{\mu\nu}(x) = e\overline{\psi}(x)\gamma^\mu\psi(x),$$

for the charged fermion and electromagnetic fields. ***

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. Books such as Griffiths [13] or Schey [38] 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

$$\boldsymbol{A} \cdot \boldsymbol{B} = AB\cos\theta \qquad \text{(Scalar product)}, \tag{A.2}$$

where θ is the angle between the vectors $\boldsymbol{A}$ and $\boldsymbol{B}$. The *vector product* or *cross product* of two vectors is

$$\boldsymbol{A} \times \boldsymbol{B} = AB\sin\theta\hat{\boldsymbol{n}} = \begin{vmatrix} \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}} \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix} \qquad \text{(Vector or cross product)}, \tag{A.3}$$

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. [16].

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 [13].

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

$$\begin{aligned} \varepsilon_{ijk}\varepsilon_{lmn} = {} & \delta_{il}\delta_{jm}\delta_{kn} + \delta_{jl}\delta_{km}\delta_{in} + \delta_{kl}\delta_{im}\delta_{jn} \\ & - \delta_{jl}\delta_{im}\delta_{kn} - \delta_{kl}\delta_{jm}\delta_{in} - \delta_{il}\delta_{km}\delta_{jn}, \end{aligned} \tag{A.13}$$

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 appear commonly in realistic problems and product rules for vector derivatives can be proved that are similar 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

$$\int_a^b \left(\frac{\partial f}{\partial x} \right) dx = \int_a^b F(x)\, dx = f(b) - f(a) \qquad \text{(Fundamental theorem of calculus),} \tag{A.26}$$

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.

Let us consider the fundamental theorems of vector calculus for gradients, divergences, and curls.

A.5.1 Fundamental Theorem of Gradients

For *gradients*, the fundamental theorem takes the form

$$\int_{\boldsymbol{A}}^{\boldsymbol{B}} (\boldsymbol{\nabla} F)\cdot d\boldsymbol{l} = F(\boldsymbol{B}) - F(\boldsymbol{A}) \qquad \text{(Fundamental theorem of gradients)}, \tag{A.27}$$

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).

A.5.2 Fundamental Theorem of Divergences

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 it is commonly termed *the divergence theorem* in the literature and we will typically adopt that terminology.

A.5.3 Fundamental Theorem of Curls

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 indi-

[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.

cated, it is commonly called *Stokes' theorem* in the literature and 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.

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. 2.4).

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

$$\text{Gradient:}\quad \boldsymbol{\nabla}F = \frac{\partial F}{\partial r}\hat{\boldsymbol{r}} + \frac{1}{r}\frac{\partial F}{\partial\theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial F}{\partial\theta}\hat{\boldsymbol{\phi}}, \tag{A.47}$$

$$\text{Divergence:}\quad \boldsymbol{\nabla}\cdot\boldsymbol{V} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2V_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta V_\theta) + \frac{1}{r\sin\theta}\frac{\partial V_\phi}{\partial\phi}, \tag{A.48}$$

$$\begin{aligned}\text{Curl:}\quad \boldsymbol{\nabla}\times\boldsymbol{V} = {} & \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta V_\phi) - \frac{\partial V_\theta}{\partial\phi}\right]\hat{\boldsymbol{r}} \\ & + \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial V_r}{\partial\phi} - \frac{\partial}{\partial r}(rV_\phi)\right]\hat{\boldsymbol{\theta}} + \frac{1}{r}\left[\frac{\partial}{\partial r}(rV_\theta - \frac{\partial V_r}{\partial\theta}\right]\hat{\boldsymbol{\phi}},\end{aligned} \tag{A.49}$$

$$\begin{aligned}\text{Laplacian:}\quad \nabla^2F = {} & \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial F}{\partial r}\right) \\ & + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial F}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2F}{\partial\phi^2},\end{aligned} \tag{A.50}$$

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 Taylor Series Expansion

In one dimension a Taylor series expansion may be written for infinitesimal ε as

$$\begin{aligned} f(x+\varepsilon) &= f(x) + \varepsilon\frac{df}{dx} + \frac{1}{2!}e^2\frac{d^2 f}{dx^2} + \cdots \\ &= \left[1 + \varepsilon\frac{d}{dx} + \frac{1}{2!}\left(\varepsilon\frac{d}{dx}\right)^2 + \cdots\right] f(x) = \exp\left(\varepsilon\frac{d}{dx}\right) f(x). \end{aligned} \tag{A.58}$$

For a function of three variables this generalizes to

$$f(x+\varepsilon_x, y+\varepsilon_y, z+\varepsilon_z) = \exp\left(\varepsilon_x\frac{\partial}{\partial x}\right)\exp\left(\varepsilon_y\frac{\partial}{\partial y}\right)\exp\left(\varepsilon_z\frac{\partial}{\partial z}\right) f(x,y,z), \tag{A.59}$$

which may be written in vector notation as

$$f(\boldsymbol{r}+\boldsymbol{\varepsilon}) = \exp(\boldsymbol{\varepsilon}\cdot\boldsymbol{\nabla}) f(\boldsymbol{r}) = \left[1 + \boldsymbol{\varepsilon}\cdot\boldsymbol{\nabla} + \frac{1}{2!}(\boldsymbol{\varepsilon}\cdot\boldsymbol{\nabla})^2 + \cdots\right] f(\boldsymbol{r}). \tag{A.60}$$

A.9 Binomial Expansion

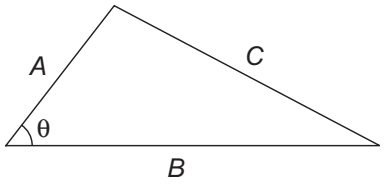
The binomial expansion is

$$(1+a)^n = 1 + na + \frac{n(n-1)}{2!}a^2 + \frac{n(n-1)(n-2)}{3!}a^3 + \cdots \tag{A.61}$$

It is particularly useful when $a \ll 1$, since then $(1+a)^n \simeq 1 + na$.

A.10 Miscellaneous Equations

LAW OF COSINES:



$$C^2 = A^2 + B^2 - 2AB\cos\theta \qquad \text{(Law of cosines).} \tag{A.62}$$

A.11 Integration by Parts

For scalar functions u and v, by basic calculus manipulations

$$\int_a^b \frac{d}{dx}(uv)dx = uv|_a^b = \int_a^b u\left(\frac{dv}{dx}\right)dx + \int_a^b v\left(\frac{du}{dx}\right)dx,$$

from which we may write

$$\int_a^b u\left(\frac{dv}{dx}\right)dx = uv|_a^b - \int_a^b v\left(\frac{du}{dx}\right)dx, \tag{A.63}$$

which is often expressed more compactly as

$$\int udv = uv - \int vdu. \tag{A.64}$$

This is called *integration by parts*, and is often useful when it is necessary to itegrate a function multiplied by a derivative of another function. The net effect of integration by parts is to transfer the derivative operation from one function to the other, at the expense of adding a boundary term $uv|_a^b$. For vector calculus this generalizes to equations of the form

$$\int_V \phi(\boldsymbol{\nabla}\cdot\boldsymbol{A})d\tau = -\int_V \boldsymbol{A}\cdot(\boldsymbol{\nabla}\phi)d\tau + \oint_S \phi\boldsymbol{A}\cdot d\boldsymbol{a}, \tag{A.65}$$

$$\int_S \phi(\boldsymbol{\nabla}\times\boldsymbol{A})\cdot d\boldsymbol{a} = \int_S [\boldsymbol{A}\times(\boldsymbol{\nabla}\phi)]\cdot d\boldsymbol{a} + \oint_P \phi\boldsymbol{A}\cdot d\boldsymbol{l}, \tag{A.66}$$

$$\int_V \boldsymbol{B}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A})d\tau = \int_V \boldsymbol{A}\cdot(\boldsymbol{\nabla}\times\boldsymbol{B})d\tau + \oint_S (\boldsymbol{A}\times\boldsymbol{B})\cdot d\boldsymbol{a}, \tag{A.67}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are arbitrary vectors, ϕ is an arbitrary scalar function, a subscript V indicates a volume integral, a subscript S indicates a surface integral, and a subscript P indicates a line integral. Notice that in these expressions also one is integrating a function times a vector derivative (divergence or curl) of a function, and the operation transfers the derivative from one function to the other (as a gradient or curl), at the expense of an additional boundary term. Such equations can be derived by using the product rules and fundamental theorems of vector calculus.

A.12 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

$$\delta(x-a) = \begin{cases} 0 & (\text{if } x \neq a), \\ \infty & (\text{if } x = a), \end{cases} \tag{A.68}$$

with the normalization

$$\int_{-\infty}^{+\infty} \delta(x-a)dx = 1. \tag{A.69}$$

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 $f(x)$,

$$f(x)\delta(x-a) = f(a)\delta(x-a) \tag{A.70}$$

and insertion of a Dirac delta function in an integral over a function picks out the value of the integrand at $x = a$,

$$\int_{-\infty}^{+\infty} f(x)\delta(x-a)dx = f(a). \tag{A.71}$$

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 $\boldsymbol{x} = (x_1, x_2, x_3)$,

$$\delta^3(\boldsymbol{x}-\boldsymbol{X}) \equiv \delta(x_1 - X_1)\delta(x_2 - X_2)\delta(x_3 - X_3). \tag{A.72}$$

This vanishes everywhere except at $\boldsymbol{x} = \boldsymbol{X}$, and generalizing Eqs. (A.68) and (A.69) to 3D,

$$\int_{\Delta V} \delta^3(\boldsymbol{x}-\boldsymbol{X})\, d^3x = \begin{cases} 1 & (\text{If } \Delta V \text{ contains } \boldsymbol{x} = \boldsymbol{X}), \\ 0 & (\text{If } \Delta V \text{ doesn't contain } \boldsymbol{x} = \boldsymbol{X}), \end{cases} \tag{A.73}$$

where ΔV is the integration volume over d^3x, while Eq. (A.71) generalizes to

$$\int_{\Delta V} f(\boldsymbol{x})\, \delta^3(\boldsymbol{x}-\boldsymbol{a})\, d^3x = f(\boldsymbol{a}). \tag{A.74}$$

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

[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 [25] for a more complete exposition.

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,

$$\nabla^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = -4\pi\delta^3(\boldsymbol{x}-\boldsymbol{x}'), \tag{A.75}$$

which can be of considerable utility in evaluating electrostatic integrals (see Example 2.8). 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,

$$\frac{d}{dx}\delta = -\delta\frac{d}{dx}. \tag{A.76}$$

(One can show that this "parts" operation is legitimate, even though δ is not a true function.)

B Appendix B Electromagnetic Units

We have used SI units systematically throughout this book except for employing Heaviside–Lorentz units in discussing the Lorentz invariance of the Maxwell equations in Section 16.8. This Appendix gives the Maxwell equations and Lorentz force equations expressed in SI units, gaussian or CGS units, and Heaviside–Lorentz units.

B.1 Maxwell Equations in SI Units

In SI units the vacuum Maxwell equations are

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = \frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)}, \tag{B.1a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{B.1b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{B.1c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B} - \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} = \mu_0 \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{B.1d}$$

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) or electrostatic units the vacuum Maxwell equations are

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = 4\pi\rho \qquad \text{(Gauss's law)}, \tag{B.3a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E} + \frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{B.3b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{B.3c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B} - \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} = \frac{4\pi}{c}\boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{B.3d}$$

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

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = \rho \qquad \text{(Gauss's law)}, \tag{B.5a}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E} + \frac{\partial\boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)}, \tag{B.5b}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0 \qquad \text{(No magnetic charges)}, \tag{B.5c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B} - \frac{\partial\boldsymbol{E}}{\partial t} = \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)}, \tag{B.5d}$$

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] Aharonov, Y., and Bohm, D. 1959. *Phys. Rev.*, **115**, 484.
[2] Arfken, G. 1970. *Mathematical Methods for Physicists*. New York: Academic Press.
[3] Ashcroft, N. W., and Mermin, N. D. 1976. *Solid State Physics*. Toronto: Cengage.
[4] Barrera, R. G., Estevez, G. A., and Giraldo, J. 1985. *European J. Physics*, **6**, 287.
[5] Chaichian, M., Merches, I., Radu, D., and Tureanu, A. 2016. *Electrodynamics: An Intensive Course*. Berlin: Springer.
[6] Chambers, R. G. 1960. *Phys. Rev. Lett.*, **5**, 3.
[7] Corson, D., and Lorrain, P. 1962. *Introduction to Electromagnetic Fields and Waves*. San Francisco: W. H. Freeman and Company.
[8] Drude, P. 1900a. *Ann. Physik*, **306**, 566.
[9] Drude, P. 1900b. *Ann. Physik*, **308**, 369.
[10] Foster, J., and Nightingale, J. D. 2006. *A Short Course in General Relativity*. New York: Springer.
[11] Garg, A. 2012. *Classical Electromagnetism in a Nutshell*. Princeton: Princeton University Press.
[12] Goldhaber, A. S., and Nieto, M. M. 1971. *Rev. Mod. Phys.*, **43**, 277.
[13] Griffiths, D. J. 2024. *Introduction to Electrodynamics*. Cambridge: Cambridge University Press.
[14] Guidry, M. W. 1999. *Gauge Field Theories: An Introduction with Application*. New York: Wiley Interscience.
[15] Guidry, M. W. 2018. *Sci. Bull.*, **63**, 2 (DOI: 10.1016/j.scib.2017.11.021).
[16] Guidry, M. W. 2019. *Modern General Relativity: Black Holes, Gravitational Waves, and Cosmology*. Cambridge: Cambridge University Press.
[17] Guidry, M. W., and Sun, Y. 2022. *Symmetry, Broken Symmetry, and Topology in Modern Physics: A First Course*. Cambridge: Cambridge University Press.
[18] Hunt, B. J. 2012 (11). *Phys. Today.*, **65**, 48.
[19] Jackson, J. D. 1999. *Classical Electrodynamics*. New York: Wiley.
[20] Jackson, J. D., and Okun, L. B. 2001. *Phys. Rev. Mod. Phys.*, **75**, 663.
[21] Kamberaj, H. 2022. *Electromagnetism*. Springer.
[22] Kobzarev, I. Yu., and Okun, L. B. 1968. *Sov. Phys. Usp.*, **11**, 338.
[23] Kogut, John B. 2001. *Introduction to Relativity*. San Diego: Academic Press.
[24] Kragh, H. 1991. *Applied Optics*, **30**, 4688.
[25] Lighthill, M. J. 1958. *Introduction to Fourier Analysis and Generalised Functions*. Cambridge: Cambridge University Press.
[26] Lorrain, P., and Corson, D. R. 1978. *Electromagnetism: Principles and Applications*. New York: W. H. Freeman and Company.

[27] 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).
[28] Maxwell, J. C. 1890a. https://archive.org/details/scientificpapers01maxwuoft/page/498/mode/2up.
[29] Maxwell, J. C. 1890b. https://archive.org/details/scientificpapers01maxwuoft/page/500/mode/2up.
[30] Maxwell, James Clerk. 2005. *A Treatise on Electricity and Magnetism*. Oxford: Oxford University Press.
[31] Nolte, D. D. 2025. *Maxwellian Steampunk and the Origins of Maxwell's Equations*. https://galileo-unbound.blog/2025/03/05/maxwellian-steampunk-and-the- origins-of-maxwells-equations/.
[32] Purcell, E. M., and Morin, D. J. 2013. *Electricity and Magnetism*. Cambridge: Cambridge University Press.
[33] Quigg, C. 1983. *Gauge Theories of the Strong, Weak, and Electromagnetic Interactions*. Reading, Massachusetts: Benjamin/Cummings.
[34] Ryder, L. H. 1996a. *Quantum Field Theory*. Cambridge: Cambridge University Press.
[35] Ryder, L. H. 1996b. *Quantum Field Theory*. Cambridge: Cambridge University Press.
[36] Schneider, C. S, and Ertel, J. P. 1998. *Am. J. Phys.*, **66**, 686.
[37] Schutz, Bernard. 1980. *Geometrical Methods of Mathematical Physics*. Cambridge: Cambridge University Press.
[38] Sommerfield, A. 1952. *Electrodynamics*. New York: Academic.
[39] Tjossem, P. J. H., and Brost, E. C. 2011. *Am. J. Phys.*, **79**, 353.
[40] Wald, R. M. 2022. *Advanced Classical Electromagnetism*. Princeton: Princeton University Press.
[41] Williams, E. R, Faller, J. E, and Hill, H. A. 1971. *Phys. Rev. Lett.*, **26**, 721.
[42] Zangwill, A. 2013. *Modern Electrodynamics*. Cambridge: Cambridge University Press.

Index