Lecture Notes: Classical Electromagnetic Theory

Mike Guidry

This document summarizes lecture notes on classical electromagnetic theory in a format suitable for presentation. Sources and references for the material contained here may be found in the typeset lecture notes. Problem references are to problems at the ends of chapters of the typeset notes. These lecture notes are linked from the Physics 541 class homepage.

Contents

1 Overview **1**

1.1 The Synthesis of Classical Electromagnetism 2

1.2 Electromagnetism in Modern Physics 10

1.2.1 Quantization of Electrical Charge 11

1.2.2 Particles in the Universe 12

2 Introduction to Electrostatics **15**

2.1 Coulomb's Law 17

2.2 The Electric Field 21

2.3 Principle of Superposition 28

2.4 Gauss's Law 30

2.5 The Scalar Potential 39

2.6 The Electric Field and the Scalar Potential 44

2.7 Superposition of Scalar Potentials 48

2.8 The Poisson and Laplace Equations 54

2.9 Work and Energy in Electric Fields 57

2.9.1 Work to Move a Test Charge 59

2.9.2 Energy of a Continuous Charge Distribution 63

3 Electrostatic Boundary Value Problems **67**

3.1 Properties of Conductors 68

3.2 Capacitance 73
3.3 Energy Stored in a Capacitor 75
3.4 Boundary Conditions 76
3.5 Properties of Poisson and Laplace Solutions 79
3.5.1 One-Dimensional Laplace Equation 80
3.5.2 2D and 3D Laplace Equations 82
3.6 Uniqueness Theorems 84
3.7 Uniqueness Theorems by Green's Methods 87
3.8 Boundary-Value Problems by Green Functions 95
3.9 Method of Images 100
3.10 Green Function for the Conducting Sphere 106
3.11 Solving the Poisson and Laplace Equations Directly 115
3.11.1 Separation of Variables in Cartesian Coordinates 116
3.11.2 Separation of Variables in Spherical Coordinates 126
3.12 Multipole Expansions 134
3.12.1 Multipole Components of the Electric Field 151
3.12.2 Energy of a Charge Distribution in an External Field 152

4 Electrostatics in Matter **155**
4.1 Dielectrics 158
4.1.1 Capacitors and Dielectrics 158
4.1.2 Dielectric Constants 162
4.1.3 Bound Charge and Free Charge 163
4.2 Moments of a Molecular Charge Distribution 165
4.3 Induced Dipole Moments 170
4.4 Permanent Dipole Moments 174
4.5 Polarization 177
4.6 Field Outside Polarized Dielectric Matter 178
4.7 Field Inside Polarized Dielectric Matter 185

4.8 Displacement and Constituitive Relations 190
4.9 Surface and Volume Bound Charges 193
4.10 Average Electric Fields in Matter 198
4.11 Boundary Conditions at Interfaces 204
4.12 Dielectric Boundary Value Problems 222

5 Magnetostatics 241
5.1 Introduction 242
5.2 Magnetic Forces 247
5.3 The Law of Biot and Savart 252
5.4 Differential Form of the Biot–Savart Law 255
5.5 The Vector Potential and Gauge Invariance 263
5.6 Magnetic Fields of Localized Current Distributions 268

6 Magnetic Fields in Matter 277
6.1 Ampère's Law in Magnetically Polarized Matter 279
6.1.1 Macroscopic Averaging 280
6.1.2 The Auxiliary Field $\boldsymbol{H}$ and Constituitive Relations 283
6.1.3 Magnetic Boundary-Value Conditions 292
6.2 Solving Magnetostatic Boundary-Value Problems 293
6.2.1 Method of the Vector Potential 294
6.2.2 Using a Magnetic Scalar Potential if $\boldsymbol{J} = \mathbf{0}$ 297
6.2.3 Hard Ferromagnets 298
6.2.4 Solve Using the Vector Potential 301
6.2.5 Example: Uniformly Magnetized Sphere 305
6.3 Faraday and the Law of Induction 324

7 Maxwell's Equations 337
7.1 The Almost-but-Not-Quite Maxwell's Equations 338
7.1.1 Ampère's Law and the Displacement Current 339

7.1.2 Implications of the Ampère–Maxwell Law 344
7.2 Vector and Scalar Potentials 346
7.2.1 Exploiting Gauge Symmetry 350
7.2.2 Coulomb Gauge 357
7.3 Retarded Green Function 368

8 Minkowski Spacetime 385
8.1 Maxwell's Equations and Special Relativity 387
8.2 Minkowski Spacetime and Spacetime Tensors 389
8.3 Coordinate Systems and Transformations 390
8.4 Minkowski Space 392
8.4.1 Coordinate Systems in Euclidean Space 396
8.4.2 Basis Vectors 399
8.4.3 Dual Basis 405
8.4.4 Expansion of Vectors 410
8.4.5 Scalar Product of Vectors and the Metric Tensor 411
8.4.6 Vector Scalar Product and the Metric Tensor 412
8.4.7 Relationship of Vectors and Dual vectors 413
8.4.8 Properties of the Metric Tensor 418
8.4.9 Line Elements and Distances 421
8.5 Integration 428
8.6 Differentiation 429
8.7 Transformations 432
8.7.1 Rotational Symmetries 433
8.7.2 Galilean Transformations 435
8.8 Spacetime Tensors and Covariance 436
8.8.1 Spacetime Coordinates and Transformations 437
8.8.2 Covariance and Tensor Notation 441
8.8.3 Tangent and Cotangent Spaces 446

8.8.4 Coordinates in Spacetime 447
8.8.5 Coordinate and Non-Coordinate Bases 448
8.8.6 Tensors and Coordinate Transformations 449
8.8.7 Tensors Specified by Transformation Laws 452
8.8.8 Symmetric and Antisymmetric Tensors 463
8.8.9 Summary of Algebraic Tensor Operations 467
8.8.10 Tensor Calculus in Spacetime 468
8.8.11 Invariant Equations 472

9 Special Relativity **475**
9.1 The Indefinite Metric of Spacetime 476
9.2 Minkowski Space 477
9.2.1 Invariance of the Spacetime Interval 481
9.2.2 Tensors in Minkowski Space 483
9.3 Lorentz Transformations 485
9.3.1 Rotations 488
9.3.2 Lorentz Boosts 489
9.3.3 Lightcone Diagrams 495
9.3.4 Causal Structure of Spacetime 502
9.3.5 Time Machines and Causality Paradoxes 504
9.3.6 Lorentz Transformations in Spacetime Diagrams 507
9.4 Lorentz Invariance of Maxwell's Equations 520
9.4.1 Maxwell Equations in Non-covariant Form 521
9.4.2 Scalar and Vector Potentials 522
9.4.3 Gauge Transformations 523
9.4.4 Maxwell Equations in Manifestly Covariant Form 527

Chapter 1

Overview

The first inkling of electricity and magnetism were when humans began to realize and remark upon

- naturally occurring *electricity in amber*, and
- naturally occurring *magnetism in lodestones*.

These properties were *known to the ancient Greeks*, but the modern quantitative understanding of electricity and magnetism emerged over a *period of only about a century*, beginning in the late 1700s.

1.1 The Synthesis of Classical Electromagnetism

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

1. *Cavendish's pioneering experiments in electrostatics* in the *early 1770s*,
2. *Publication of Coulomb's work on electrostatics* beginning in *1785*, which triggered world-wide interest in the subject.
3. *Faraday's study of time-varying currents and magnetic fields* in the *mid-1800s*,
4. *Maxwell's famous 1865 paper* synthesizing in equations the prior results into a *dynamical theory of the electromagnetic field*,
5. *Heaviside's reformulation of Maxwell's equations in their modern vector calculus form* in the *late 1800s*, and
6. *Herz's publication in 1888* of data demonstrating that *transverse electromagnetic waves propagated at the speed of light,* thereby placing Maxwell's theory on a firm empirical footing.

Thus, by the year 1900 classical electromagnetic theory was in place, and could be summarized concisely in the *Maxwell equations*.

The *Maxwell equations* may be written in *free space using SI units* and modern notation as

$$\begin{aligned} \nabla \cdot \boldsymbol{E} &= \frac{\rho}{\varepsilon_0} && \text{(Gauss's law)}, \\ \nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)}, \\ \nabla \cdot \boldsymbol{B} &= 0 && \text{(No magnetic charges)}, \\ \nabla \times \boldsymbol{B} - \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} &= \mu_0 \boldsymbol{j} && \text{(Ampère–Maxwell law)}, \end{aligned}$$

1. $\boldsymbol{E}$ is the *electric field*,
2. $\boldsymbol{B}$ is *the magnetic field*,
3. ρ is the *charge density*,
4. $\boldsymbol{j}$ is the *current vector*,
5. ε_0 is the *permittivity of free space*,

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

where the farad (F) is the derived SI unit of capacitance,

6. and μ_0 is the *permeability of free space*,

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

For the *Maxwell equations*

$$\begin{aligned} \nabla \cdot \boldsymbol{E} &= \frac{\rho}{\varepsilon_0} && \text{(Gauss's law)}, \\ \nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)}, \\ \nabla \cdot \boldsymbol{B} &= 0 && \text{(No magnetic charges)}, \\ \nabla \times \boldsymbol{B} - \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} &= \mu_0 \boldsymbol{j} && \text{(Ampère–Maxwell law)}, \end{aligned}$$

the *density* ρ and the *current* $\boldsymbol{j}$ satisfy the *continuity equation*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{j} = 0 \qquad \text{(continuity equation)},$$

which ensures *conservation of charge* by requiring that

- *charge variation in some volume* (arbitrarily small)
- is caused by *current through the surface bounding the volume*.

Ampère's original law was *valid only for stationary charge densities* because it lacked the $\partial \boldsymbol{E}/\partial t$ term of the 4th Maxwell equation.

1. Maxwell realized that for *time-dependent densities* Ampère's law was *incompatible with the continuity equation*.
2. Maxwell *added the* $\partial \boldsymbol{E}/\partial t$ *term to Ampère's law*, which brought the Maxwell equations into *harmony with the continuity equation*.

Maxwell's modification of the *Ampére law*

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j} \qquad \text{(Ampère law)},$$

to the *Ampére–Maxwell law*

$$\nabla \times \boldsymbol{B} - \underbrace{\frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t}}_{\text{Maxwell}} = \mu_0 \boldsymbol{j} \qquad \text{(Ampère–Maxwell law)},$$

The term added by Maxwell is called the *displacement current*.

> This *coupling of* $\boldsymbol{E}$ *and* $\boldsymbol{B}$ *unified the previously separate subjects of electricity and magnetism*.

This has a number of far-reaching implications.

1. We may now speak of *electromagnetism*.
2. This will lead eventually to the *interpretation of electromagnetic waves as light*.
3. That Maxwell's equations obey the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{j} = 0$$

   will lead to the idea of *electromagnetic gauge invariance* and a *quantum field theory of electromagnetism (QED)*.
4. Electromagnetic gauge invariance will eventually be generalized to the weak and strong interactions, resulting in the quantum field theory that we call the *Standard Model of elementary particle physics*.

Symmetry plays a fundamental role in physics. The Maxwell equations exhibit some important symmetries.

- *Invariance under Space and Time Translations:* Invariance under translations in the spatial and time axes is associated with *conservation of momentum and energy*. This implies that we can *assign energy and momentum to the electromagnetic field*.

- *Invariance under Rotations:* The vector form of the Maxwell equations ensures *rotational invariance (isotropy of space)*, which in turn implies *conservation of angular momentum* by the electromagnetic field.

  That invariance under space and time translations, and rotations imply conservation of momentum, energy, and angular momentum, respectively, follows from **Noether's theorem:** *For every continuous symmetry of a field theory Lagrangian there is a corresponding conserved quantity.*

- *Symmetry under Space Inversion (Parity) P:* In the presence of rotational invariance, parity is equivalent to mirror reflection. Under parity $\boldsymbol{E} \rightarrow -\boldsymbol{E}$, which is the transformation law for a *polar vector* (normal 3-vector), but under parity $\boldsymbol{B} \rightarrow \boldsymbol{B}$, which is the transformation law for a *pseudovector* or *axial vector*.

- *Symmetry under Time Reversal (T):* A motion picture of electromagnetic events would *look the same if run backwards or forward*.

- *Lorentz Invariance:* Electromagnetism is *Lorentz invariant (consistent with special relativity)*. Einstein was strongly influenced by this property of electromagnetism in formulating the special theory of relativity.

- *Gauge Invariance (Charge Conservation):* That the Maxwell equations are consistent with the continuity equation implies that *charge is conserved locally*. This conservation law is associated with *local gauge invariance* for the electromagnetic field.

  We shall have more to say about the *local gauge symmetry of electromagnetism* and its far-reaching implications in later chapters.

- *Electric and Magnetic Field Asymmetry:* The Maxwell equations exhibit a *large asymmetry between the roles of electric and magnetic fields*. If magnetic charge existed, the Maxwell equations would become highly symmetric under a transformation $\boldsymbol{E} \rightarrow \boldsymbol{B}$ and $\boldsymbol{B} \rightarrow -\boldsymbol{E}$.

  However, *no magnetic charges (no magnetic monopoles)* have ever been observed.

The *four Maxwell equations*, supplemented by *boundary conditions*, may be solved (in principle) for the fields $\boldsymbol{E}$ and $\boldsymbol{B}$, if the charge density ρ and current $\boldsymbol{j}$ are known.

- However, these equations make *no reference to forces*, which are often the *connection to experimental results*.
- This is remedied by introducing the *Lorentz force law*,

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

> The force on a particle with charge q and velocity $\boldsymbol{v}$ at the point $\boldsymbol{x}$ is determined completely by the instantaneous values of the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ at $\boldsymbol{x}$.

- Then *Newton's second law*,

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

determines the motion of charges in the electromagnetic field.

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

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

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

1. The solution of Maxwell's equations with appropriate boundary conditions often *requires considerable mathematical and computational prowess*.

2. Classical electromagnetism itself has changed little since the late 1800s, but *the context in which we view classical electromagnetism has changed dramatically* because of modern advances in quantum field theory.

1. The following chapters will address the first point at a practical level, by providing *mathematical and computational tools to facilitate solution* of Maxwell's equations.

2. The remainder of the current chapter will address the second point at a more philosophical level, by *setting electromagnetism in the context of modern theoretical physics*.

1.2 Electromagnetism in Modern Physics

One of the remarkable findings of modern physics is that

- the already beautiful edifice of Maxwell's equations for electromagnetism
- hides within it a deceptively powerful symmetry called *(local) gauge invariance*—in essence a symmetry under local phase transformations—
- that, with suitable exposition, explains the origin of both classical and quantum theories of electromagnetism.

But that is not all! The local gauge symmetry describing electromagnetism can be generalized mathematically into a more powerful theory

- that can *partially unify* the electromagnetic and weak nuclear forces into a single *electroweak interaction*,
- and this can be generalized into the *Standard Model* that partially unifies the electromagnetic, weak, and strong interactions in a single gauge theory.

1.2.1 Quantization of Electrical Charge

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

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

- The charges on all presently known particles or systems of particles, are found to be *integral multiples of this unit* (with positive or negative signs).
- It is *known experimentally* that the ratios of charges between different particles are *integers to one part in* 10^{20}.
- Indeed, the *stability of the atomic matter all around us would be compromised* by even a tiny difference in the absolute values of the electron and proton charges.

There is presently *no convincing explanation for quantization of charge*, but it is an empirical fact.

> Dirac proposed long ago that, *if magnetic monopoles existed, there could be a topological reason for charge quantization.* However, no magnetic monopoles have been found, so Dirac's idea remains conjecture.

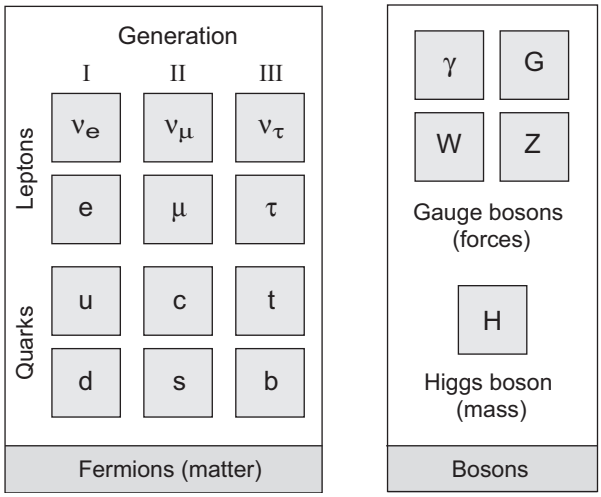


Figure 1.1: Elementary particles of the Standard Model. Photons are labeled by γ and gluons by G. Elementary particles of half-integer spin that don't undergo strong interactions are called *leptons;* electrons and electron neutrinos are examples. Particles made from quarks, antiquarks, and gluons (and 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 parallel classification of antiparticles has been omitted.

1.2.2 Particles in the Universe

Matter consists of the fermions in the Standard Model. The particles of the Standard Model are summarized in Fig. 1.1.

- In the Standard Model matter is formed from fermions,
- while interactions between elementary particles are mediated by the exchange of gauge bosons, and
- masses for bare (that is, non-interacting) particles arise from interactions with the Higgs boson.

For reasons that are not yet understood,

- the fermions (matter fields) of the Standard Model may be divided into *three generations (or families), I, II, and III*,
- with the *fermions of each generation being successively more massive than in the preceding generation*, and
- with *many properties repeating themselves in successive generations*.

> For example, the muon μ in generation II acts in many respects as if it were *a heavier version of the electron* e in generation I.

The Standard Model property of most interest to us is *electrical charge*.

- All *leptons* (weakly interacting fermions) in the Standard Model have charges that are a multiple of $|q_e|$,
- but the *hadrons* (strongly interacting fermions) are the *quarks*, with *charges that are multiples of* $\frac{1}{3}|q_e|$.

For example,

- the *charge of the up quark* u is $\frac{2}{3}|q_e|$ and
- the *charge of the down quark* d is $-\frac{1}{3}|q_e|$.

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.*

Indeed, *no free particles with 3rd-integer charges have ever been observed in experiments*.

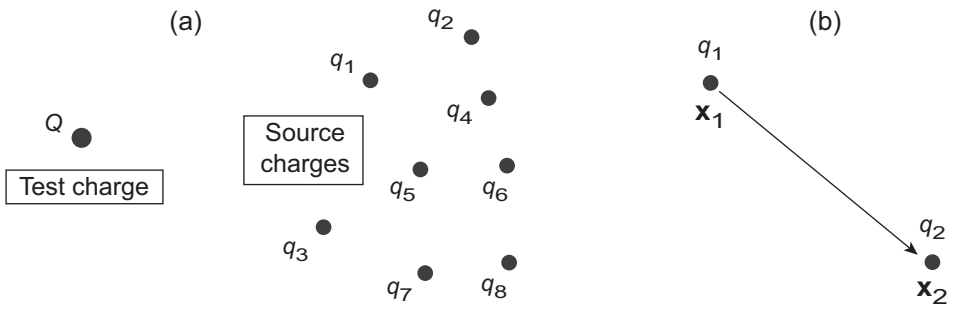
> For example, *charge is an additive quantum number* and a *proton has a udu quark structure* (two up quarks and one down quark). The *charge* Q_p *for a proton* is then
>
> $$Q_p = 2Q_u + Q_d = 2\left(\frac{2}{3}\right) + \frac{1}{3} = +1,$$
>
> in terms of the fundamental charge unit $|q_e|$, while a *neutron has a udd quark structure* and
>
> $$Q_n = Q_u + 2Q_d = \frac{2}{3} + 2\left(-\frac{1}{3}\right) = 0,$$
>
> and the neutron charge $|Q_n|$ is zero.

> Thus, some particle charges are fractions of $|q_e|$, but the *physically observable particles* all appear to have charges that are *integer multiples of the fundamental charge unit* given by $|q_e|$.

Chapter 2

Introduction to Electrostatics

The fundamental problem to be solved in electrodynamics is illustrated schematically in part (a) of the following figure



If we have a distribution of n distinct *source charges* q_1, q_2, q_3, ... q_n, what net force do they exert on a *test charge* Q?

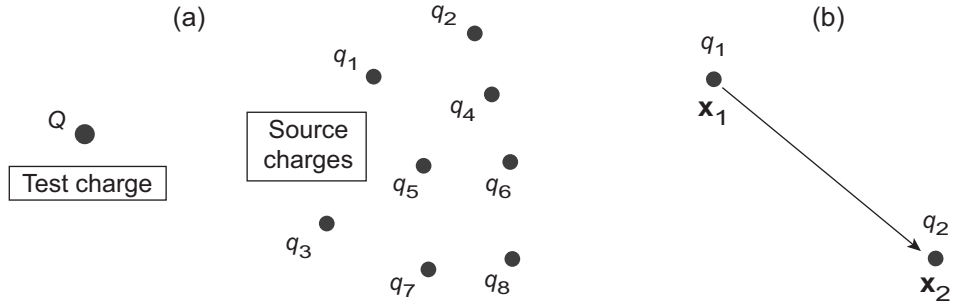


Figure 2.1: (a) The prototype electrostatics problem: interaction in vacuum of a test charge Q with a set of stationary source charges q_n. (b) Coulomb interaction between two isolated test charges.

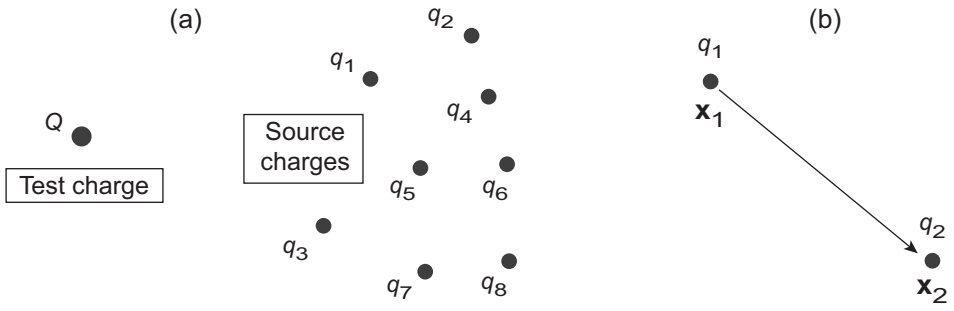
In the most general case both the source charges and test charges may be in motion but

- we shall begin with the simpler case of *electrostatics,* where the *source charges are assumed to be fixed in spatial position*.
- We shall also assume initially that the *source and test charges are embedded in vacuum*.

Let us now consider a quantitative description of this idealized problem, utilizing data, mathematics, and physical intuition as our guides.

2.1 Coulomb's Law

Consider first the force acting between two isolated charges at rest with respect to each other, as illustrated in part (b) of the following figure.



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*,

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

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 charges are opposite* and *repulsive if they are the same*.

The constant k depends on the system of units that is in use.

Two common choices are *electrostatic units* and *the SI system*.

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

2. In the *SI system* of units,

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

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

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

where the farad (F) is the derived SI unit of electrical capacitance. Thus Coulomb's law expressed in SI units is

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

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

Deviations from the Inverse-Square Law

It is common to report possible *experimental deviation from the inverse square law* in one of two ways.

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

2. Assume that the electrostatic potential has the *Yukawa form*

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

   where m_γ is the mass of the photon

   The photon should be *identically massless* for an inverse square force law.

3. Then possible deviations from the inverse square law are often reported as *an upper limit on* μ or an *upper limit on* m_γ.

1. The original *experiments of Cavendish* using concentric spheres in 1772 established an *upper limit* $|\varepsilon| \leq 0.02$.
2. *Modern determinations based on Gauss's law* have pushed this limit to $\varepsilon = (2.7 \pm 3.1) \times 10^{-16}$.
3. The best limit 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.

Experimental limits suggest that

- we can assume that *the photon is massless* and
- *deviations from Coulomb's law are negligible*

on all length scales investigated thus far.

2.2 The Electric Field

In experiments determining the interaction of charges one *typically measures a force*.

- However, much of the power of modern theoretical physics derives from the ability to *abstract broader implications from direct measurements*.
- Perhaps no abstract concept has been more powerful in the development of physics than that of a *field* (an instance of something defined at every point of spacetime).
- Let us introduce an *electric field* $\boldsymbol{E}$ acting on a test charge q by defining
$$\boldsymbol{F} = q\boldsymbol{E}.$$
- Thus the electric field is the *force per unit test charge* at a particular point in spacetime.
- Since
  - *force is a vector* and
  - *charge is a scalar*,

  the electric field generated by a charge is a *vector field* $\boldsymbol{E}(t,\boldsymbol{x})$ defined at each point of spacetime $(t,\boldsymbol{x})$.

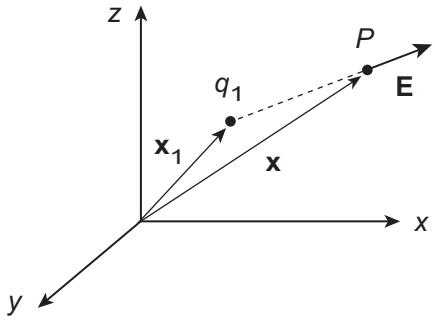


Figure 2.2: The electric field vector $\boldsymbol{E}$ at a point P, generated by a charge q_1 located at position $\boldsymbol{x}_1$.

- *Comparing with Coulomb's law*, the electric field at a point $P(\boldsymbol{x})$ produced by a point charge q_1 at $\boldsymbol{x}_1$ is

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

  as illustrated in Fig. 2.2.

- The electric field is produced by *the source charges*.

- It exists at a point whether there is a test charge there or not.

- The SI *unit of charge is the coulomb (C)* and

- the electric field $\boldsymbol{E}$ *has units of volts per meter* in the SI system.

We are presently considering electrostatics, so the time coordinate will be suppressed for now.

Significance of Fields in Electromagnetism

- As suggested by $\boldsymbol{F} = q\boldsymbol{E}$, we generally measure force and fields may be inferred from that, *suggesting a derivative role for fields*.
- But modern physics places *strong emphasis on the fields*.
- Indeed, Maxwell's equations are formulated in terms of
  - electric fields $\boldsymbol{E}$ and
  - magnetic fields $\boldsymbol{B}$,

  and *electromagnetic fields are the central concept* of the theory of electromagnetism.
- The importance of fields is *most obvious in relativistic quantum field theory*, which is not our subject here.
- However, the *field concept traces historically to the introduction of electric and magnetic fields* in the description of classical electromagnetism, which is our subject.

Originally it was believed that forces associated with charges, currents, and magnets

- constituted *action at a distance,*
- meaning that the forces acted *instantaneously over any distance*.

> *Similar statements apply to gravity*, since *Newtonian gravity acts instantaneously*. *General relativity* (in which gravity propagates at lightspeed) *eliminated action at a distance* in gravitational physics.

The modern view—shaped by experimental measurement and the development of quantum field theory is that

- *fields are every bit as fundamental as particles* (arguably even more so)
- In this picture, *forces are mediated by fields*, and
- the *lightspeed limit* set by special relativity means that

> no signal can transmit a force faster than the speed of light,

- This relegates action at a distance to the dustbin.

If, for example, a charge moves,

- the fields created by the charge change, but
- that change isn't felt immediately at every point $(\boldsymbol{x},t)$ of spacetime.
- Maxwell's equations require the change to be propagated through changes in 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 *no faster than the speed of light* c.

This abstract view was initially resisted by many, particularly because

- the fields *could exist in vacuum*, and
- *did not describe tangible matter*.

Remember that the *prevailing view until Einstein and special relativity in 1905* was that

- *waves require a medium for propagation* and that
- light propagated through a special "material" called the *aether*.

But modern physics views things quite differently:

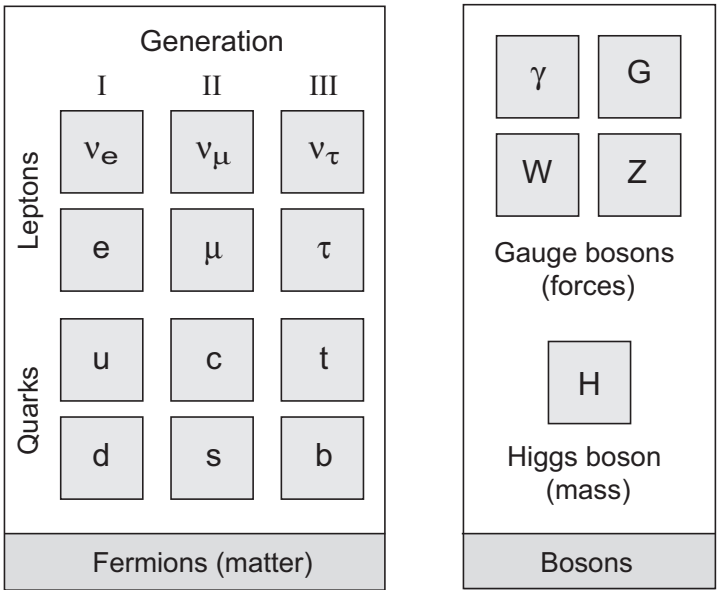
- First, the *aether is fiction* and *light can propagate in a vacuum*.
- Introduction of electric and magnetic fields allows a *simple and clean mathematical description of electromagnetic phenomena*,
- but the fields are *not just mathematical abstractions;* they are *real physical entities* ("as real as a rinoceros"—*Classical Electromagnetism in a Nutshell*, A. Garg).
- They carry concrete physical properties such as *energy*, *momentum*, and *angular momentum*.
- This is *most clear in quantum field theory*, where
  - these physical attributes become properties of *quanta of the field* (photons) and
  - one views electromagnetic interactions as being *mediated by exchange of virtual photons*.
- These photons associated with the electromagnetic field are *observable:*

  At sufficiently high energy, *collision of two photons can produce matter* in the form of an electron and positron pair; *real as a rinoceros, indeed!*

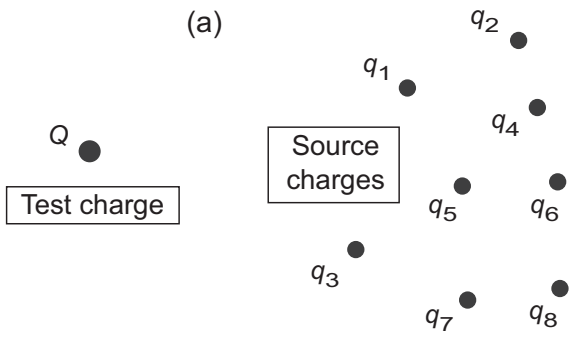
Summary: Relativistic quantum field theory implies that *all non-gravitational fundamental interactions* (electromagnetic, strong, and weak)

- are *mediated by the gauge bosons of the Standard Model* (γ, G, W, Z in figure shown below)
- that are the *quanta of gauge fields* generalizing the gauge symmetry of photons.

Thus is the *rich legacy of introducing the concept of fields in classical electromagnetism*.



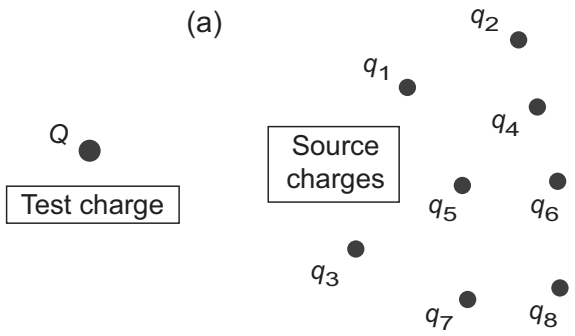
2.3 Principle of Superposition



Now let us address the more general case in the figure above of the *interaction of a test charge with n source charges*.

- In principle this could be a quite complicated problem, but
- it is an experimental fact that if the interactions between charged particles are *not too large*, and
- if we can *ignore quantum effects* at the microscopic level,
- the interaction between any two charges is unaffected by the presence of all other charges.
- Thus, the total force acting on the test charge Q can be obtained by *summing the interactions of the charges pairwise*.
- This is termed the *principle of linear superposition.*

The ultimate source of superposition is the *linearity of the Maxwell equations in* $\boldsymbol{E}$ *and* $\boldsymbol{B}$.



Since the principle of linear superposition applies to the forces it applies also to the electric fields computed from $\boldsymbol{F}$.

- The electric field acting at position $\boldsymbol{x}$ in in the above figure is given by a *vector sum of contributions from all source charges* q_i,

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

  where we work in SI units.

- In most practical problems the charges can be approximated by a *continuous charge density* $\rho(\boldsymbol{x}')$, such that

- the charge contained in a small 3D volume element $d^3x' \equiv dx'dy'dz'$ centered at $\boldsymbol{x}'$ is $\Delta q = \rho(\boldsymbol{x}')\Delta x\Delta y\Delta z$.

- Then the sum over discrete charges may be replaced by an *integral over a continuous charge distribution*,

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

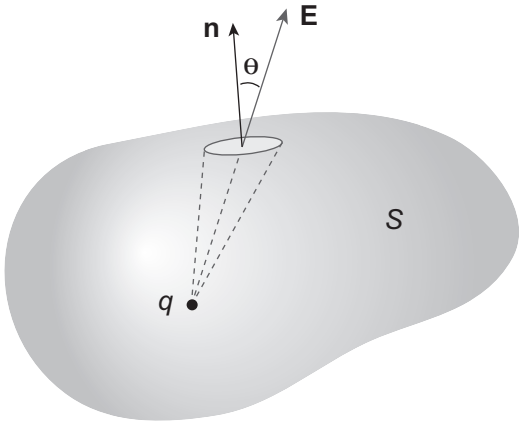


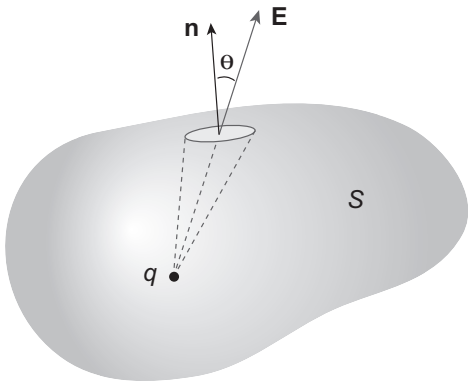
Figure 2.3: Closed surface S illustrating Gauss's law. The electric field generated by the charge q is $\boldsymbol{E}$, the normal vector to the surface is $\boldsymbol{n}$, and da is an element of surface area.

2.4 Gauss's Law

The *electric field for a continuous charge distribution* may be calculated from

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

- However, evaluating the integral in this equation isn't always the easiest solution for the electric field
- If a problem has some level of symmetry, often another integral result called *Gauss's law* can lead to an easier solution.



The figure shows a charge q enclosed by a surface S.

- The normal component of $\boldsymbol{E}$ times a surface element da is

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

- where in the last step we have used that

$$\cos\theta\,da = r^2\,d\Omega,$$

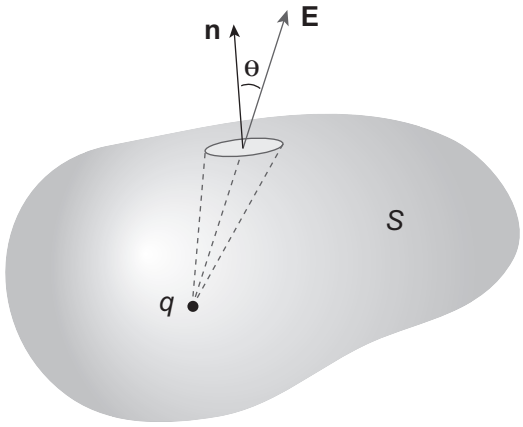
  with $d\Omega$ being the solid angle subtended by da at the position of the charge.

- If the normal component of $\boldsymbol{E}$ is now integrated over the entire surface,

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

  where $\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da$ is the *flux through the surface* S.

- This is *Gauss's law in integral form* for a single charge.



- For a *set of discrete charges* Gauss's law takes the form,

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

- where the summation over i is restricted to charges q_i that are *inside the surface* S.

- For a *continuous charge distribution* Gauss's law takes the form

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

where the right-side integration is over the volume V contained within the surface S.

Gauss's law may be viewed as *integral formulations of the law of electrostatics*. Corresponding *differential forms of Gauss's law* may be obtained using the *divergence theorem*.

> ***Divergence theorem:*** For a vector field defined within a volume V enclosed by a surface S,
>
> $$\oint_S \boldsymbol{A}\cdot\boldsymbol{n}\,da = \int_V \boldsymbol{\nabla}\cdot\boldsymbol{A}\,d^3x,$$
>
> where the *left side is the surface integral* of the outward normal component of $\boldsymbol{A}$ and the *right side is the volume integral* of the divergence of $\boldsymbol{A}$.

The *divergence theorem* allows the expression

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

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 for arbitrary V *only if the integrand vanishes*, so

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

which is the *differential form of Gauss's law* for continuous charge distributions.

> *We have obtained the first Maxwell equation.* Only three more to go!

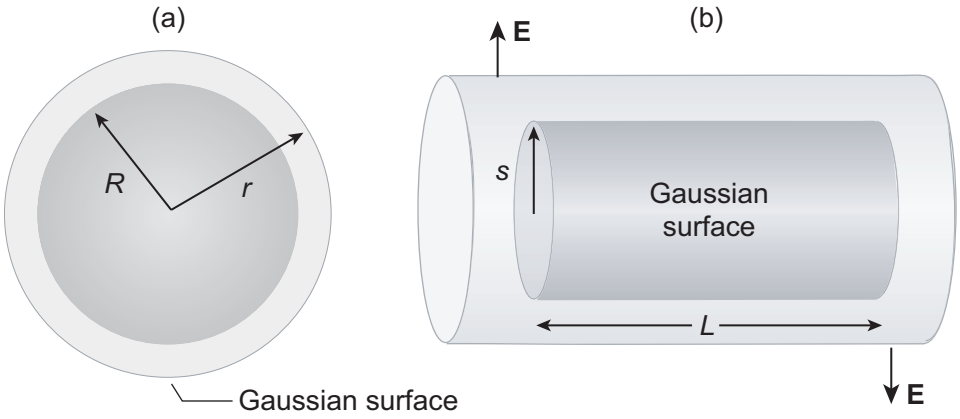


Figure 2.4: Examples of applying Gauss's law to find the electric field. (a) A ball of uniformly distributed charge with radius R. (b) A cylinder carrying a charge density proportional to the distance s from the cylindrical axis.

Example: *Figure 2.4(a) shows a charged solid 2-sphere (a ball) 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 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

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

where Q is the total charge. 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,

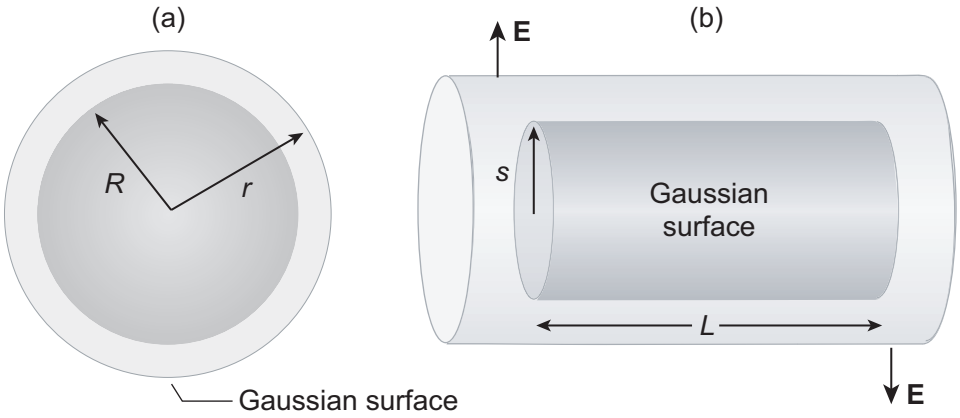
$$4\pi r^2 |\boldsymbol{E}| = \frac{1}{\varepsilon_0} Q,$$

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.



Example: *Consider Fig. (b) above, where a long cylinder carries a charge density proportional to the distance s' from the axis of the cylinder, $\rho = ks'$, where k is a constant. What is the electric field inside the cylinder?*

Let's draw a Gaussian cylinder of radius s and length L, as illustrated in Fig. (b). 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 with a cylindrical volume element $d\tau = s\, ds\, d\phi\, dz$

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

By symmetry $\boldsymbol{E}$ *must point radially outward from the cylinder's central axis*, 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,

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

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.

One sees generally from such examples that using Gauss's theorem

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

to determine the electric field

- is most useful when there is a *high degree of symmetry* that can be exploited to evaluate integrals.
- Generally our goal with Gauss's law is to *simplify the integral on the left side.*

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*.

If none of these possibilities are fulfilled, Gauss's law is still valid but it may not be the easiest way to determine the electric field.

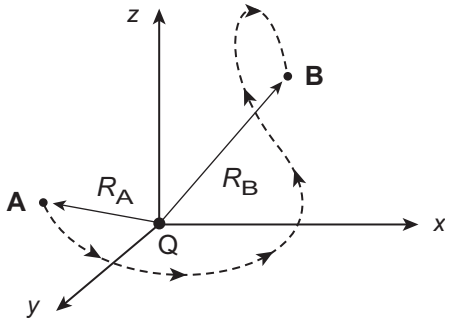


Figure 2.5: Line integral in an electric field generated by a charge Q at the origin.

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 Q at the origin, as illustrated in Fig. 2.5. In spherical coordinates the electric field $\boldsymbol{E}$ and line element $d\boldsymbol{l}$ are

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

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

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

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

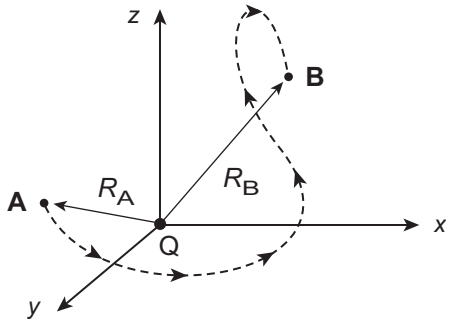
$$\oint \boldsymbol{E}\cdot d\boldsymbol{l} = 0,$$

since $R_a = R_b$ in that case. Now invoke *Stokes' theorem*.

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

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

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 $\boldsymbol{n}$.



Then Stokes' theorem

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

implies that

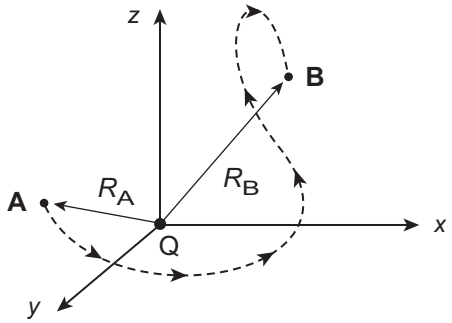
$$\int_S (\nabla \times \boldsymbol{E}) \cdot \boldsymbol{n}\, da = \oint_P \boldsymbol{E} \cdot d\boldsymbol{l} = 0,$$

and since this *must be valid for any closed path*,

- the integrand on the left side must vanish,

  $$\nabla \times \boldsymbol{E} = \boldsymbol{0},$$

  and *the curl of the electric field* $\boldsymbol{E}$ *is zero*.
- Because of $\nabla \times \boldsymbol{E}$ 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 $\boldsymbol{B}$ has *the same value for all possible paths*.



Thus we can *define a function* $\Phi(\boldsymbol{x})$ by

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

where $\mathscr{G}$ is a chosen *standard reference point*.

The function $\Phi(\boldsymbol{x})$ is called the *scalar potential* or the *electric potential*.

The *scalar potential* $\Phi(\boldsymbol{r})$ associated with a *point charge at the origin* is given by

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

where r is the separation between charge and point.

Invoking superposition, the *potential generated by a collection of n charges* is

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

For a *continuous charge distribution* the potential evaluates to

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

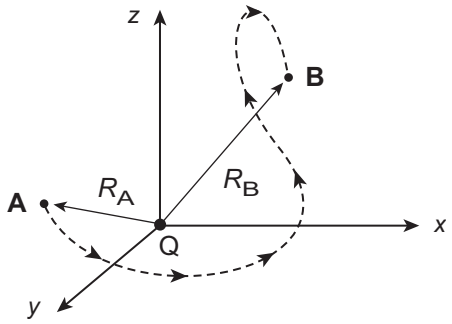
$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da' \qquad \text{(2D surface charge)},$$

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

where in these expressions

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

2.6 The Electric Field and the Scalar Potential

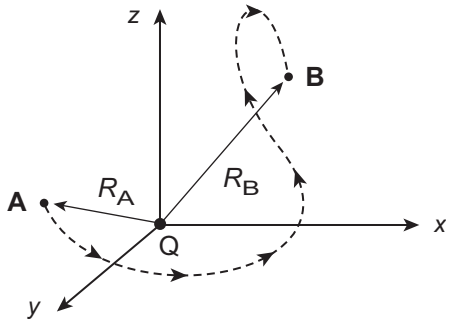


The *electric field is special* because it has *vanishing curl*.

> We shall now use this to reduce finding the electric field (a vector problem) to a simpler scalar problem.

The *potential difference* between points $\boldsymbol{A}$ and $\boldsymbol{B}$ (see figure) is given by

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



Applying the fundamental theorem for gradients,

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

to the above result

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

then gives

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

But since this must be true for any points $\boldsymbol{A}$ and $\boldsymbol{B}$, the *integrands on the two sides must be equal* and we obtain

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

which is a differential version of

$$\Phi(\boldsymbol{x}) = -\int_{\mathscr{G}}^{\boldsymbol{x}} \boldsymbol{E}(\boldsymbol{x}) \cdot d\boldsymbol{l}.$$

In SI units *force is measured in newtons* and *charge in coulombs*. Then

- electric fields $\boldsymbol{E}$ have units of *newtons per coulomb*.
- The potential Φ has units then of *newton-meters per coulomb*, or *joules per coulomb*, where

$$1\ \frac{\text{joule}}{\text{coulomb}} \equiv 1\ \text{volt}$$

A *big advantage of the potential formulation* is that

  - if you can determine Φ ,
  - then you can obtain the electric field by taking the gradient, $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$.

- 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* lies in the *restrictions following from* $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$, which implies the constraints

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

There is an *essential ambiguity in defining the potential* because of the arbitrary reference point $\mathscr{G}$ appearing in

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

since changing it *shifts the potential by a constant amount*.

- Thus *the potential itself has no physical meaning*, but
- if we *choose the same reference $\mathscr{G}$ for all potentials*,
  1. *differences between potentials* $(\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_j)$ are independent of the reference point and
  2. *gradients of the potential* $\boldsymbol{\nabla}\Phi$ are unaffected by shifting the reference point a constant amount, since the derivative of a constant is zero,

  so these have physical meaning.

The *selection of a reference point* is in principle arbitrary.

The most common choice in electrostatics is to take the *reference point for potentials* to be

- *an infinite distance away from the charges*,
- where the *potential drops to zero*.

2.7 Superposition of Scalar Potentials

The *scalar potential* Φ *obeys the superposition principle:*

- The forces acting on a test charge Q are the vector sum of contributions from each source charge q_i,

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

- and since $\boldsymbol{F} = Q\boldsymbol{E}$, dividing by Q implies linearity for the electric fields $\boldsymbol{E}$ also,

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

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

$$\Phi = \Phi_1 + \Phi_2 + \Phi_3 + \ldots$$

  meaning that the potential at a point $\boldsymbol{x}$ is the *sum of the potentials due to all source charges computed 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 $\boldsymbol{F}$ and electric fields $\boldsymbol{E}$ corresponds to *vector sums*,
  - but linear superposition of potentials entails an ordinary *arithmetic sum* over scalar quantities Φ_i,

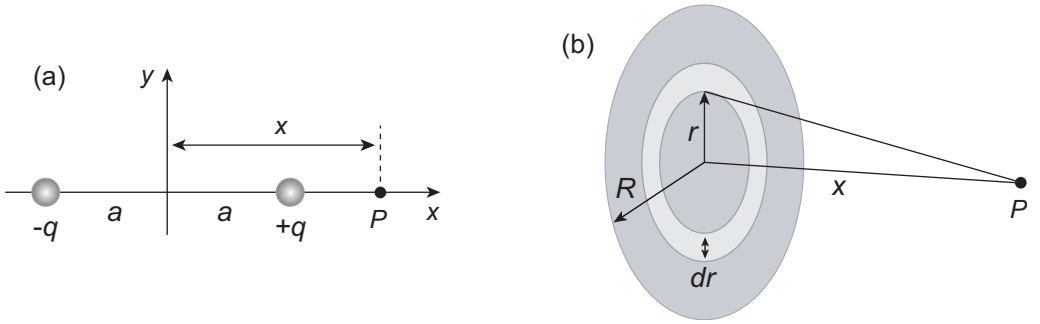


Figure 2.6: (a) An electric dipole charge configuration. (b) Uniformly charged disk of radius R and surface charge density σ.

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

The distance from $-q$ to P is $x+a$ and the distance from $+q$ to P is $x-a$. Then at the point P the potential Φ is

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

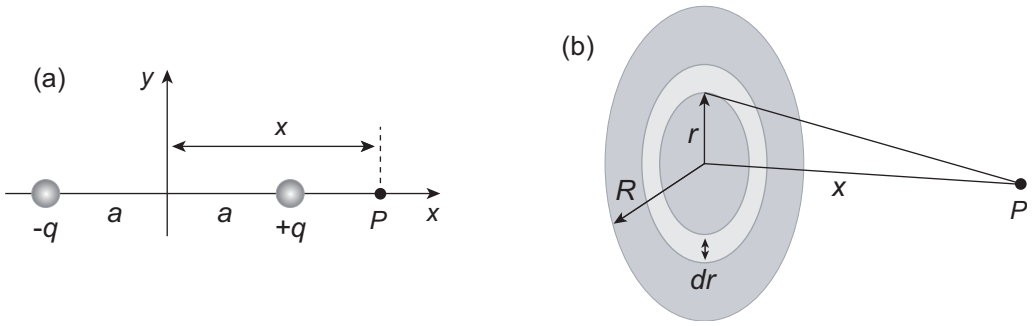
The *electric field is then oriented along the x axis* and given by *minus the gradient of the potential*

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

The scalar potential and electric field may be *approximated as*

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

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



Example: *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. (b) above.*

For a ring of radius r and width dr the charge element *at a distance* $(r^2+x^2)^{1/2}$ *from the point* P is $dq = (2\pi r dr)\sigma$. Then

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

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

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

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

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

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

- For *relatively simple charge distributions* like those in the preceding two examples, we can
  - *find the potentials* Φ easily and
  - then determine the electric field from the *gradient of the potential*, $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$.
- However, for more complex situations we may need *more powerful and systematic ways* to determine potentials.
- In the next section we show that finding the potentials can be cast in the form of *solving a second-order partial differential equation*.
- This approach may be preferred over the ones we have examined so far, particularly for *problems with complicated boundary conditions*.

What is the curl of the electric field? What is the divergence of $\boldsymbol{E}$?

2.8 The Poisson and Laplace Equations

The *curl of the electric field vanishes*, $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$. What is the *divergence of the electric field* equal to? From $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$,

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

and *comparing with Gauss's law*

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0},$$

gives *Poisson's equation,*

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

where the Laplacian operator $\boldsymbol{\nabla}^2$ *operates on a scalar to return a scalar*.

In *cartesian coordinates* (x, y, z) the Laplacian operator is

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

in *spherical coordinates* (r, θ, ϕ) the Laplacian operator is

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

and in *cylindrical coordinates* (ρ, θ, ϕ) the Laplacian operator is

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

For regions where $\rho = 0$, Poisson's equation reduces to *Laplace's equation,*

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

By construction, *solving Poisson's equation or Laplace's equation* with appropriate boundary conditions is *equivalent to solving for the potential* Φ *using*

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

from which *the electric field* $\boldsymbol{E}$ *can be calculated using* $\boldsymbol{E} = -\nabla\Phi$.

Example: *Earlier it was asserted that*

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

gives the scalar potential $\Phi(\boldsymbol{x})$ *for a 3D continuous charge distribution. If correct,* $\Phi(\boldsymbol{x})$ *should be a solution of the 3D Poisson equation*

$$\boldsymbol{\nabla}^2\Phi(\boldsymbol{x}) = -\frac{\rho(\boldsymbol{x})}{\varepsilon_0}.$$

Let's check that it is.

Applying the Laplacian operator to both sides of this equation gives

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

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

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

- *giving immediately*

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

This is *Poisson's equation*, proving the assertion that

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

is the scalar potential for a 3D continuous charge distribution.

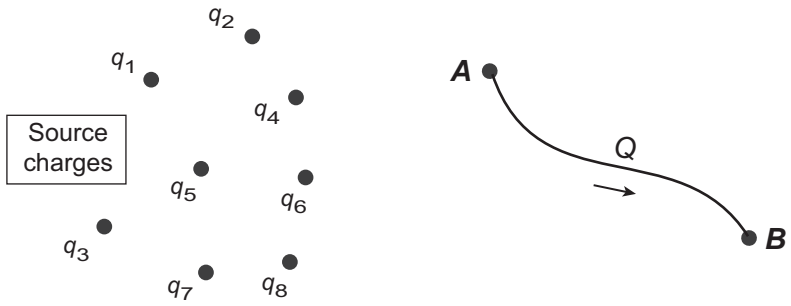


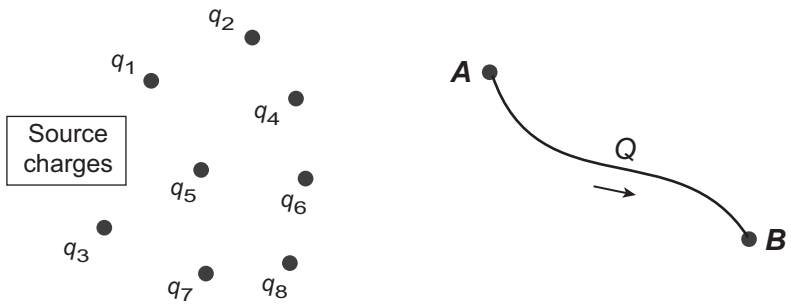
Figure 2.7: Moving an electrical charge Q on a path between endpoints ***A*** and ***B*** in an electric field generated by a set of source charges.

2.9 Work and Energy in Electric Fields

A question of fundamental importance in electrostatics is illustrated in Fig. 2.7:

> If we have a stationary configuration of source charges and a test charge is moved along some path between two points ***A*** and ***B***, *how much work will be done?*

Question: How do we calculate work done along a path?



2.9.1 Work to Move a Test Charge

The *force exerted on the test charge* Q *is* $\boldsymbol{F} = Q\boldsymbol{E}$, where the *electric field* $\boldsymbol{E}$ *is generated by the source charges*.

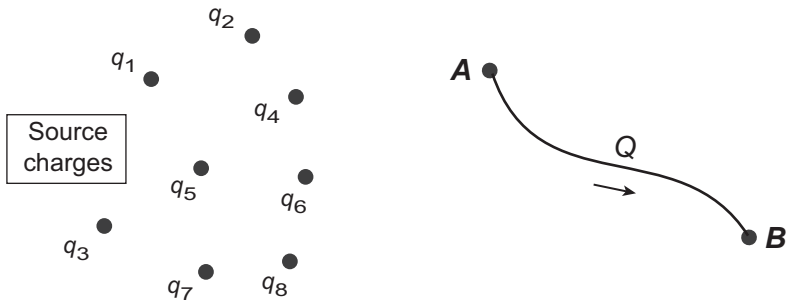
- Thus a minimal force $-Q\boldsymbol{E}$ must be exerted at each point to move the charge along the path and
- the *total work* W done is given by the line integral

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

where

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

has been used.



The work $Q[\Phi(\boldsymbol{B}) - \Phi(\boldsymbol{A})]$ is *independent of path.*

- Therefore the electrostatic force is *conservative,*
- meaning that it depends
  - only on the *difference in potentials between the end-points of the path* and
  - *not on the details* of the path followed.
- If we wish to move the charge from an infinite distance away to a point $\boldsymbol{B}$,

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

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

$$W = Q\Phi(\boldsymbol{x}) \quad \longrightarrow \quad \Phi(\boldsymbol{x}) = \frac{W}{Q}.$$

- Thus $\Phi(\boldsymbol{x}) = W/Q$ and the scalar potential $\Phi(\boldsymbol{x})$ is the *work per unit charge done against* $\boldsymbol{E}$ to move a test charge from infinity to $\boldsymbol{x}$.
- It may also be interpreted as
  - a *potential energy stored in the fields* of the assembled charge configuration
  - that could be *released by moving the charge back to infinity*.

Example: *Let's use $W = Q\Phi(\boldsymbol{x})$ 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.*

1. *The first charge costs nothing to move*, since there are no assembled charges and no electric field to fight against.
2. Adding each additional charge will require the *work summed pairwise over contributions from all charges*.
3. Therefore the *total work to assemble the charge distribution* is

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

- where r_{ij} is the distance between charges i and j,
- the initial factor of $\frac{1}{2}$ is to correct for the double counting of pairs in the double summation,
- the factor in parentheses on line 2 represents the potential $\Phi(\boldsymbol{r}_i)$ at position $\boldsymbol{r}_i$ of charge q_i generated by all the other charges.
- W represents the *work required to assemble the n charges* into some local configuration, or
- a *total (potential) energy stored in the electric fields of the final assembly* of charges

---

2.9.2 Energy of a Continuous Charge Distribution

- From the discrete-charge expression

$$W = \frac{1}{2} \cdot \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{q_i q_j}{r_{ij}} = \frac{1}{2} \sum_{i=1}^{n} q_i \Phi(\boldsymbol{r}_i),$$

we deduce that *assembling a continuous volume charge requires work*

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

where $d\tau$ is a volume element.

> This can be *rewritten to eliminate* ρ *and* Φ *in favor of* $\boldsymbol{E}$ in the following way.

- Use Gauss's law $\rho = \varepsilon_0 \boldsymbol{\nabla} \cdot \boldsymbol{E}$ to express ρ in terms of $\boldsymbol{E}$,

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

- and *integrate by parts* to give

$$W = \frac{\varepsilon_0}{2} \left(- \int_V \boldsymbol{E} \cdot (\boldsymbol{\nabla} \Phi) \, d\tau + \oint_S \Phi \boldsymbol{E} \cdot d\boldsymbol{a}, \right) \qquad d\boldsymbol{a} \equiv \boldsymbol{n} da.$$

- But $\boldsymbol{\nabla} \Phi = -\boldsymbol{E}$, so

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

where $E \equiv |\boldsymbol{E}|$, and V encloses all charge.

If we let $V \to \infty$ in

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

then the *second (surface) term tends to zero* relative to the first term and we obtain

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

where it is understood that *the integration is over all space*. The use of this equation is illustrated in the following example.

Example: *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,

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

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

Chapter 3

Electrostatic Boundary Value Problems

In the preceding chapter we introduced

- the Poisson equation, with solutions corresponding to scalar potentials $\Phi(\boldsymbol{x})$ in the presence of a charge density, and
- the Laplace equation, with solutions corresponding to scalar potentials $\Phi(\boldsymbol{x})$ in the absence of a charge density.

Those solutions result from solving the corresponding partial differential equations subject to boundary conditions, which can be a highly nontrivial matter.

In this chapter we address the nature of the solutions of the Poisson and Laplace equations, and some actual means of obtaining solutions.

3.1 Properties of Conductors

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

- *insulators,* which correspond to matter with charge carriers tightly bound to atoms or molecules that do not transport electrical charge well, and
- *conductors*, which have many delocalized charge carriers that are free to transport electrical charge. Conductors are also often called *metals*.

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

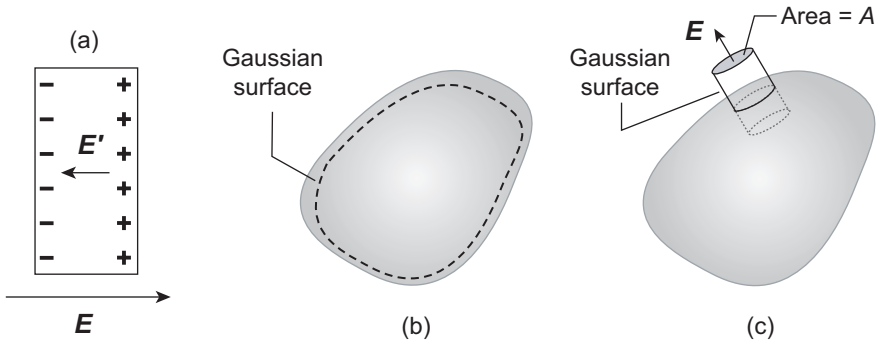


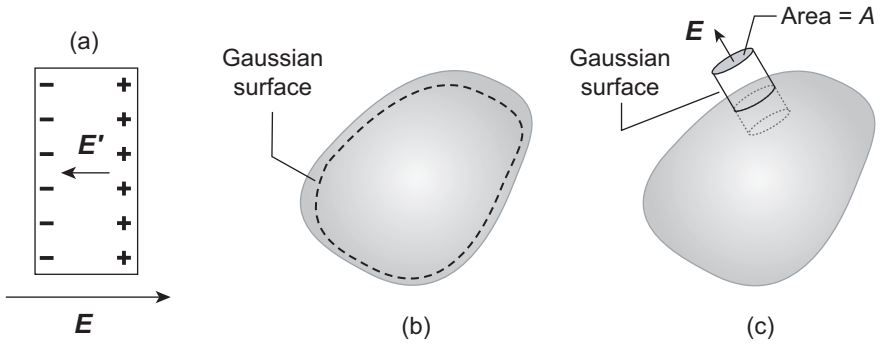
Figure 3.1: (a) Conducting slab in a uniform electric field $\boldsymbol{E}$. (b) Gaussian surface (dashed curve) inside a conductor. (c) Cylindrical gaussian surface perpendicular to the surface of a conductor.

The free mobility of charge carriers in good conductors leads to many of their basic properties:

1. $\boldsymbol{E} = 0$ *inside a conductor*. Qualitatively an electric field inside the conductor would accelerate electrons, violating electrostatic equilibrium. *More precisely,*

   - Consider Fig. 3.1(a), where a conducting slab is immersed in an external electric field $\boldsymbol{E}$.
   - Initially the external field will attract negative charges to the left side, leaving a net positive charge on the right side (an *induced charge*).
   - This polarization of charge will create an internal electric field $\boldsymbol{E}'$ that opposes the external field.
   - Charge will continue to flow until the induced internal electric field $\boldsymbol{E}'$ exactly cancels the external field, leaving a net zero electric field inside the conductor.

This is a conductor so generation of internal fields that cancel the external field typically occurs on a timescale so short that it can be assumed to be instantaneous.



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

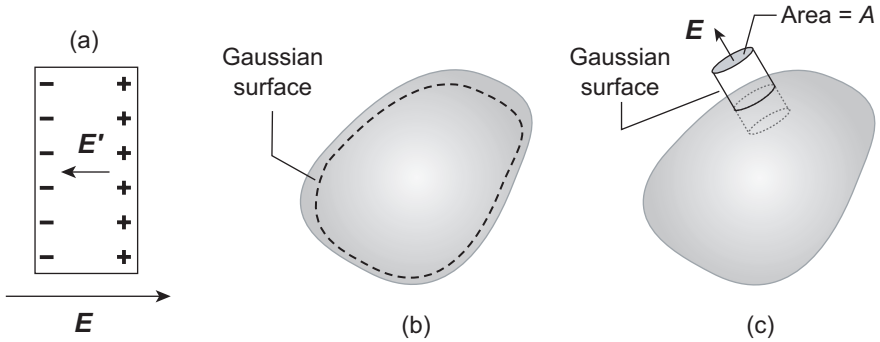
$$\oint_S \boldsymbol{E}\cdot\boldsymbol{n}\,da = \oint_S \mathbf{0}\cdot\boldsymbol{n}\,da = 0 = \frac{Q}{\varepsilon_0},$$

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

3. *Any excess charge in a conductor must reside at the surface of the conductor*. This follows immediately from point 2 above.

   – Since the gaussian surface (dashed curve) in Fig. (b) above can be place arbitrarily close to the surface,

   – any excess charge can only exist at the surface.



4. Any electric field *outside the conductor* is generated by the surface charge and is perpendicular to the surface.

   – Consider the *cylindrical gaussian surface* partially inside and partially outside the conductor in Fig. (c).

   – If $\boldsymbol{E}$ had a component tangent to the surface of the conductor, this would cause electrons to move along the surface and disturb electrostatic equilibrium.

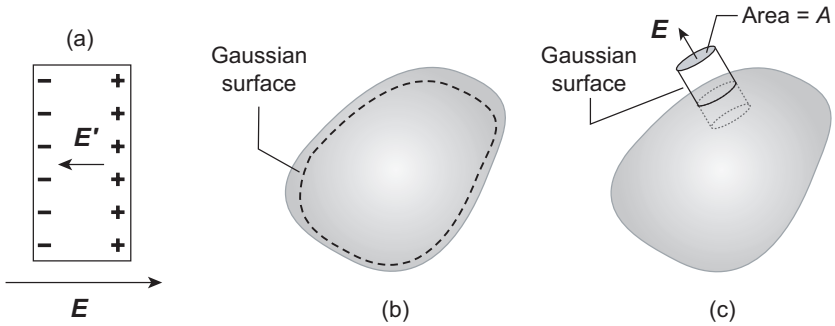
   – Thus $\boldsymbol{E}$ is perpendicular the the surface and there is no flux through the curved part of the gaussian cylinder.

   – There is also no flux through the flat face of the cylinder *inside the conductor*, because $\boldsymbol{E}$ is zero there (point 1 above).

   – Hence the net flux is only through the flat face of the gaussian cylinder outside the conductor. Applying Gauss's law to this surface,

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

   where σ is the *surface charge density* A is the *area of the cylindrical endplate*.



Thus any external electric field is proportional to the surface charge density and perpendicular to the surface.

5. The potential $\Phi(\boldsymbol{x})$ has the same value at each point in the conductor (a conductor is an *equipotential*), since for any two points in the conductor or on its surface, $\boldsymbol{E} = 0$ and

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

6. If the conductor is of irregular shape, *the surface charge density σ is greatest where the surface has the largest local curvature* (smallest radius of curvature). *Proof:*

   – Consider an irregularly shaped conductor and
   – partition the surface into small elements of area da_i subtending equal angles measured from the center.
   – Points 1 and 3 then require $\sigma_i\, da_i$ to be constant, and
   – since da_i depends on the radius of curvature, the *smaller the radius of curvature* (the larger the curvature) the *larger the local surface charge density* σ.

If conductors are present, these properties often play a large role in determining boundary conditions and solutions.

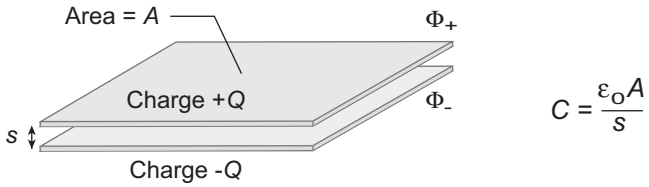


Figure 3.2: A parallel-plate capacitor.

3.2 Capacitance

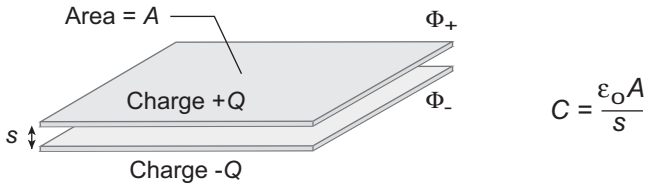
By Coulomb's law the electric field for an isolated conductor is proportional to the source charge Q and therefore the potential Φ is also, so Q and Φ are proportional to each other. The constant of proportionality is called the *capacitance* C. For an isolated conductor carrying a charge Q with potential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

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

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

> 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.

Capacitance is also a very useful concept in dealing with the change and potential associated with two or more conductors. The generic example is the *parallel-plate capacitor* illustrated in Fig. 3.2.



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 we define the difference in potential for two conductors to be V,

$$V \equiv \Phi_+ - \Phi_-,$$

the capacitance is given by

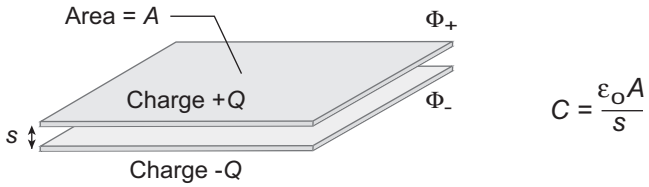
$$C = \frac{Q}{V} \qquad \text{(two conductors)}.$$

For the parallel-plate capacitor in the figure shown above the capacitance depends only on the geometry,

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

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

As will be discussed later, the capacitance of a capacitor like that in the figure above can be increased significantly by replacing the air 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 between the plates.



3.3 Energy Stored in a Capacitor

Charging a capacitor (for example, by connecting the two plates in the figure above to the poles of a battery) stores energy in the electric field created between the oppositely-charged electrodes of the device. We may determine how much energy by computing the work done to charge the capacitor to a particular level. Consider the parallel-plate capacitor shown above. If at some point in the charging process the charge on the positive plate is q, from

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

and $C = Q/V$, the work increment dW required to add the next charge increment dq is given by

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

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

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

Therefore, using $Q = CV$, the work required to charge to a potential difference V between the electrodes is

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

for a capacitor having capacitance C.

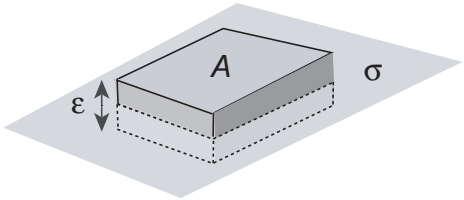


Figure 3.3: Discontinuous normal of $\boldsymbol{E}$ at surface of a conductor. The surface charge density is σ and the top area of the rectangular gaussian surface is A.

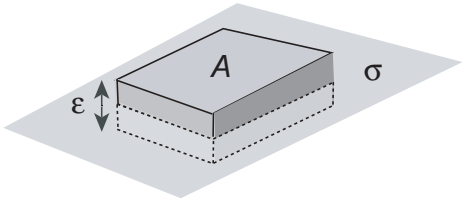
3.4 Boundary Conditions

The electric field exhibits a discontinuity if a surface charge is crossed.

- Consider Fig. 3.3, which shows a small piece of a surface having a surface charge density σ, and
- place a very thin rectangular gaussian box of height ε that extends vertically just below and just above the surface.
- From Gauss's law,

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

  where $Q = \sigma A$ is the enclosed charge and $d\boldsymbol{a} = \boldsymbol{n}a$.
- In the limit $\varepsilon \to 0$ the sides of the box contribute no flux.



- The electric fields due to the surface charge are perpendicular to the local plane and parallel to $\boldsymbol{n}$, so

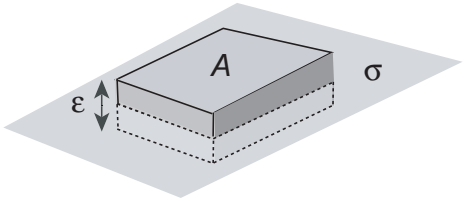
$$\oint_S \boldsymbol{E} \cdot d\boldsymbol{a} = E \int da = EA$$

and the difference between electric fields above and below the plane is

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

where $\hat{\boldsymbol{n}}$ is a unit vector perpendicular to the surface and is chosen to point upward in the figure above.

Thus, the electric field is discontinuous at the boundary, changing by an amount σ/ε_0.



The scalar potential, on the other hand, is continuous across the boundary because

- if $\boldsymbol{A}$ is a point just below the surface and $\boldsymbol{B}$ is a point just above it,

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

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

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

These boundary conditions are valid only just above and just below the surface, so the above equations are valid only in the limit that we approach the surface very closely from the top or bottom in the figure above.

3.5 Properties of Poisson and Laplace Solutions

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

3.5.1 One-Dimensional Laplace Equation

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

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

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

  $$\Phi(x) = ax + b,$$

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

> The boundary conditions in actual applications reflect the detailed physics of the system.

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 $\Phi(x+c)$,

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

for any c.

2. There can be *no local maxima or minima for the solution* as a function of the parameter x, which actually follows from the first point.

3. If there were a local maximum or minimum of Φ at some value of x it could not be the average of points on either side of it, contradicting feature 1.

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

3.5.2 2D and 3D Laplace Equations

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

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

- In general this introduces a higher level of difficulty than for the 1D case We will illustrate for 3D.
- The solutions of the Laplace and Poisson equations have two unique features that generalize those found in 1D.

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

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

     where the integral is over the surface of a sphere of radius R centered at $\boldsymbol{x}$.
  2. Because of point 1, *the solution cannot have local minima or maxima* and extrema must occur on the boundaries.

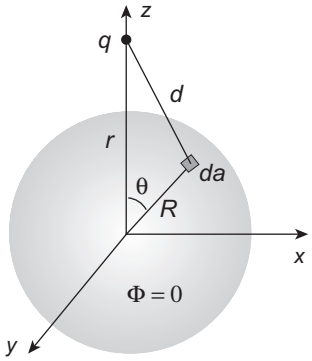


Figure 3.4: Averaging Φ over a sphere of radius R with a single point charge q external to the sphere.

Example: *A single point charge q outside a sphere.*

For Fig. 3.4 the law of cosines gives for the distance d from the charge q to the patch da,

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

At a single point on the surface of the sphere the potential is,

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

and 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 due to q at the center of the sphere.

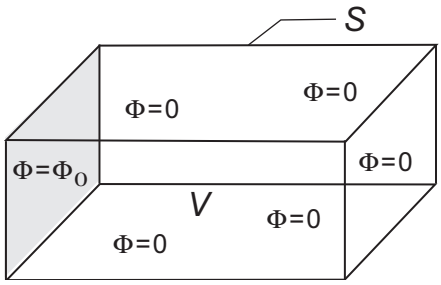


Figure 3.5: Generic example of an electrostatics problem formulated in a 3D volume V that is bounded by a 2D surface S. In this case the (finite) volume V is surrounded by a rectangular conducting box S with boundary conditions corresponding to five of the six conducting box faces held at zero potential,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

3.6 Uniqueness Theorems

Use of the Poisson or Laplace equations to determine the potential Φ requires the solution of partial differential equations with boundary conditions. A simple example of such a problem is illustrated in Fig. 3.5.

- For partial differential equations it is often not immediately obvious what constitutes appropriate boundary conditions (meaning that they allow a solution and that it is physically well-behaved).
- The proof that a given set of boundary conditions fits the bill is called a *uniqueness theorem*.
- Two categories of boundary conditions are common in electrostatics problems.

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

  This leads to two uniqueness theorems.

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

The simplest way to set boundary conditions is to specify the value of the scalar field Φ on all surfaces surrounding the region (*Dirichlet boundary conditions*).

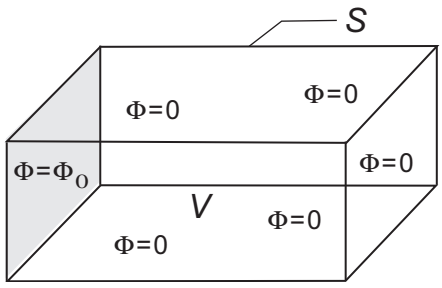
- Then the first uniqueness theorem applies.
- However, in some situations we *may not know the potential at the boundaries* but we *do know the total charge on conducting surfaces*.

This leads to a second uniqueness theorem.

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

Thus the second uniqueness theorem is appropriate for *Neumann boundary conditions*.

3.7 Uniqueness Theorems by Green's Methods



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 can be obtained using *Green's function methods*.

- These may be derived beginning with the *divergence theorem*,
$$\oint_S \boldsymbol{A}\cdot\boldsymbol{n}\,da = \int_V \boldsymbol{\nabla}\cdot\boldsymbol{A}\,d^3x,$$
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
$$\boldsymbol{\nabla}\cdot(\phi\boldsymbol{\nabla}\psi) = \phi\boldsymbol{\nabla}^2\psi + \boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\psi$$
and
$$\phi\boldsymbol{\nabla}\psi\cdot\boldsymbol{n} = \phi\frac{\partial\psi}{\partial n},$$

- where $\partial/\partial n$ is the normal derivative at the surface, directed from inside to outside the volume V.

Substitution of these equations into the divergence theorem then gives *Green's first identity*

$$\int_V (\phi \nabla^2 \psi + \nabla\phi \cdot \nabla\psi)\, d^3x = \oint_S \frac{\partial \psi}{\partial n}\, da.$$

If we then subtract from this the same expression but with ϕ and ψ interchanged, we obtain *Green's theorem* (also termed *Green's second identity*)

$$\int_V (\phi \nabla^2 \psi - \psi \nabla^2 \phi)\, d^3x = \oint_S \left[\phi \frac{\partial \psi}{\partial n} - \psi \frac{\partial \phi}{\partial n}\right] da$$

Let us now choose a particular function for the scalar 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

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

and use that

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

so that Green's theorem becomes

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

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

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

Notice *two important implications* of the result

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

1. If the surface S goes to infinity and the electric field on S falls off faster than R^{-1}, the surface term vanishes and we recover the usual result

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

2. In the second (surface) term *both* the the scalar potential Φ and the normal derivative $\partial\Phi/\partial n'$ appear.

   - Specifying Φ is associated with *Dirichlet boundary conditions*.
   - Specifying $\partial\Phi/\partial n'$ is associated with *Neumann boundary conditions*.

   Thus the system is *overdetermined* since it isn't permitted to impose both types of boundary conditions on the same closed surface.

   - Since it is overdetermined, the equation above is *not a valid solution*.
   - Later we shall address how to correct this deficiency.

Let us now use Green's first identity to show that the *use of either (but not both) Dirichlet or Neumann boundary conditions defines a unique potential problem*.

- Assume that there are two potentials, Φ_1 and Φ_2 that satisfy the Poisson equation. Let

$$U = \Phi_2 - \Phi_1$$

- Then $\boldsymbol{\nabla}^2 U = 0$ inside V and on the boundary S either
  - $U = 0$ (*Dirichlet* boundary conditions) or
  - $\partial U/\partial n = 0$ (*Neumann* boundary conditions).
- From Green's first theorem with $\phi = \psi = U$,

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

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

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

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

Thus we have shown that

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

Let us summarize some general statements about *Dirichlet* and *Neumann* boundary conditions in electrostatics problems.

1. One can specify *either Dirichlet or Neumann constraints* at each point on a boundary, *but not both*.

2. Since *either Dirichlet or Neumann conditions are adequat*e, if both are specified the system is *overdetermined*.

   Mathematically, an overdetermined system effectively has *more equations than unknowns*.

   - Such a system is *inconsistent* (no set of parameter values satisfies all equations), and
   - its equations can be manipulated to obtain *contradictory results*.

3. It is OK to specify *parts of a boundary using Dirichlet conditions* and *other parts using Neumann conditions*.

   The *uniqueness property* means that a solution obtained by any method that we wish is the *(unique) correct solution* if it

   - *satisfies the Poisson/Laplace equations* and
   - implements the *correct boundary conditions*.

Much of the uniqueness of Poisson and Laplace solutions follows from the *Helmholtz Theorem*.

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.

This is sometimes paraphrased into a uniqueness statement:

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

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

3.8 Boundary-Value Problems by Green Functions

In obtaining the result

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

(*Not valid*, because *boundary conditions are overdetermined*) we chose the function ψ to be $1/|\boldsymbol{x}-\boldsymbol{x}'|$, which satisfies

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

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

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

where we define

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

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

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

The first term on the right side of $G(\boldsymbol{x},\boldsymbol{x}')$ is the simplest Green function and is called the *Green function of free space*,

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

Recall that we derived

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

by substituting $G_0(\boldsymbol{x},\boldsymbol{x}')$ *into Green's theorem*.

- But this 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 raises the possibility that if we *substitute the generalized Green function into Green's theorem*,
- the function $F(\boldsymbol{x},\boldsymbol{x}')$ can be *chosen to eliminate from the resulting surface integral either the Dirichlet or the Neumann terms*,
- thus leaving a result with *consistent boundary conditions*.
- Indeed, if we substitute $\phi = \Phi$, and $\psi = G(\boldsymbol{x},\boldsymbol{x}')$ into Green's theorem and use

  $$\nabla^2 G(\boldsymbol{x},\boldsymbol{x}') = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}') \quad G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} + F(\boldsymbol{x},\boldsymbol{x}')$$

  we obtain

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

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

- For *Dirichlet boundary conditions*, we require that

$$G(\boldsymbol{x},\boldsymbol{x}') = 0 \qquad (\text{for } \boldsymbol{x}' \text{ on } S).$$

in the equation

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

Then the *first term in the surface integral vanishes* and the solution with *Dirichlet boundary conditions* is

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

If instead *Neumann boundary conditions* are desired, we must be a little more careful.

- The obvious choice $\partial G(\boldsymbol{x},\boldsymbol{x}')/\partial n' = 0$ for $\boldsymbol{x}'$ on S indeed banishes the second term in the surface integral of

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

- However, application of Gauss's theorem to

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

indicates that

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

implying the simplest allowable boundary condition

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

where Σ is the total area of the bounding surface S.

- Then the *solution with Neumann boundary conditions* is

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

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

Thus, we have shown formally using Green functions how to impose *consistent Dirichelet or Neumann boundary conditions* in an electrostatic boundary-value problem.

3.9 Method of Images

In the *method of images* one

- replaces the actual Poisson or Laplace problem with a different one (analog problem) that is (we can hope) easier to solve,
- but that is tailored to have boundary conditions equivalent to those of the actual problem.
- Then, since the analog problem has the same boundary conditions and is a solution of the Laplace or Poisson equation just as the actual problem,

Uniqueness Theorem I supports the validity of the solution.

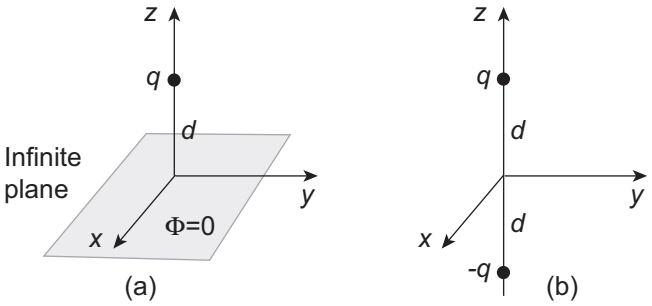


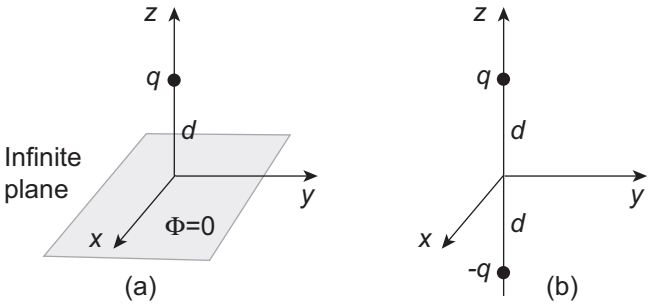
Figure 3.6: (a) Charge a distance d above an infinite conducting plane held at zero potential. (b) An analogous image-charge configuration with the charge $-q$ on the negative z axis, and no infinite plane.

In Fig. 3.6(a)

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

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

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



Now consider the completely different problem illustrated in Fig. (b) above.

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

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

- where the quantities in the denominators are distances between the charge and the point $P(x,y,z)$.

  But how does this help us? We want the solution for the problem of Fig. (a), not Fig. (b).

Well, notice that the *boundary conditions are the same* for the two problems: There is a single charge q in the region $z > 0$ (which is where we seek a solution), and from

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

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

Thus, we conclude that the solution of our original problem in Fig. (a) is given by the equation above, which is the solution of problem in Fig. (b).

This approach is called the *method of images.*

- It can be a powerful approach, but is dependent on thinking up an analog problem that has the same boundary conditions and is easy to solve.
- Notice that the image charges must in general be put in a region that is *not in the domain of the original problem*.
- In the present case we put it in the region $z < 0$ but required a solution for $z > 0$.

Now that the potential has been found for $z > 0$ we can determine the induced charge in the conducting plane caused by the positive charge at $z = d$.

- keeping the above contribution,

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

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

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

- and the derivative may be evaluated 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 the surface charge density is

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

We can then obtain the total induced charge Q by integration.

- Using polar coordinates (r, ϕ) with $r^2 = x^2 + y^2$ and $da = r\,dr\,d\phi$,

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

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

Let us also calculate the force $\boldsymbol{F}$ on q produced by the induced charge.

- From the analog problem the force is, from Coulomb's law,
$$\boldsymbol{F} = -\frac{1}{4\pi\varepsilon_0}\frac{q^2}{(2d)^2}\hat{\boldsymbol{z}},$$
where $\hat{\boldsymbol{z}}$ is a unit vector in the z direction.
- The potential, electric field, and force acting on q are expected to be the same in the analog and actual problem.
- Thus we deduce that the force for the analog problem is also the force felt in the actual problem.

3.10 Green Function for the Conducting Sphere

As discussed in Section 3.8,

- *Solution of the Laplace or Poisson equations* in a finite volume V with either
  - *Dirichlet* boundary conditions or
  - *Neumann* boundary conditions

  on the bounding surface S of V *can be obtained using Green functions*.
- For example,

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

  solves for the potential Φ in terms of a *Green function with Dirichelet boundary conditions* and
- the expression

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

  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.

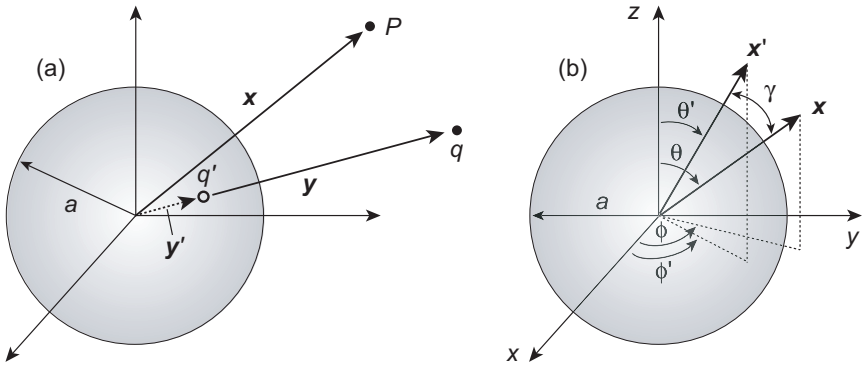


Figure 3.7: (a) Geometry of image charge method for a charge q outside a conducting sphere, with image charge q' inside the sphere. (b) Coordinates associated with the Green function $G(\boldsymbol{x},\boldsymbol{x}')$ for a conducting sphere.

As discussed by Jackson, *for image problems* like those described in Section 3.9,

- the *potential due to a unit source and its image(s)*, chosen to satisfy *homogeneous boundary conditions* is just the *Green function* of

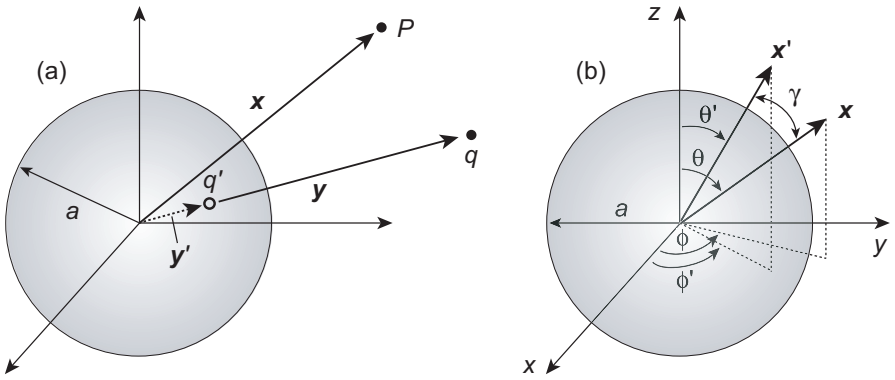
$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' - \frac{1}{4\pi}\oint_S \left[\Phi(\boldsymbol{x}')\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'}\right] da'.$$

  for *Dirichlet boundary conditions* and

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

  for *Neumann boundary conditions*.

- For example, consider a *charge outside a conducting sphere* solved by the *image method* illustrated in Fig. 3.7(a).



In the Green function $G(\boldsymbol{x},\boldsymbol{x}')$,

- $\boldsymbol{x}'$ refers to the *location of the unit source* and
- $\boldsymbol{x}$ is the point P at which *the potential is evaluated*.
- $\boldsymbol{x}$ and $\boldsymbol{x}'$, and a sphere of radius a are shown above right.
- For *Dirichlet boundary conditions* on the sphere of radius a, the *Green function* defined by

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

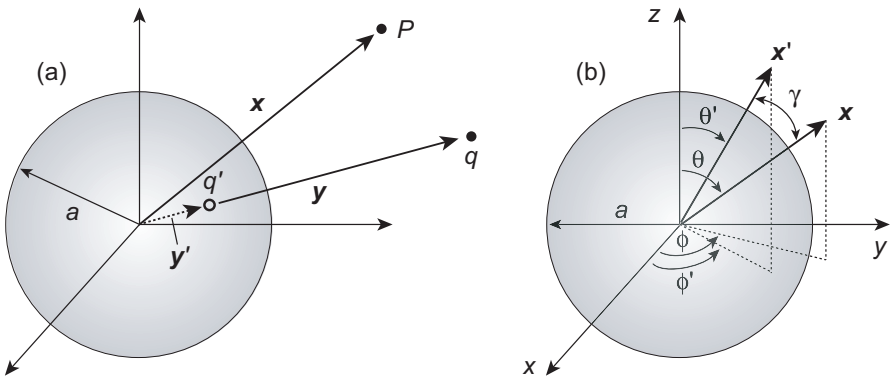
for a unit source and its image is given by the right side of

$$\Phi(\boldsymbol{x}) = \underbrace{\frac{q/4\pi\varepsilon_0}{|\boldsymbol{x}-\boldsymbol{y}|} + \frac{q'/4\pi\varepsilon_0}{|\boldsymbol{x}-\boldsymbol{y}'|}}_{\sim \text{ Conducting sphere Green function}},$$

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

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

and that q *be replaced by* $4\pi\varepsilon_0$ (unit source charge).



- *With these changes* the right side of

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

  is converted into the *Green function*

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

- which can be expressed in *spherical coordinates* as

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

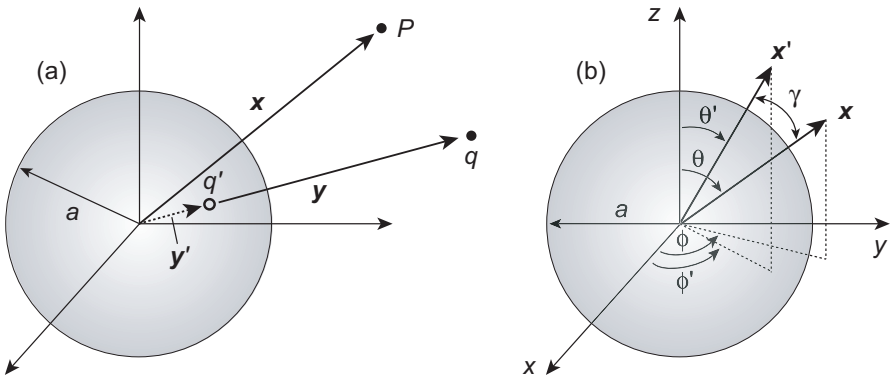
  where γ is the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$.

- For the solution

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \rho(\boldsymbol{x}')G(\boldsymbol{x},\boldsymbol{x}')d^3x' \underbrace{-\frac{1}{4\pi}\oint_S \left[\Phi(\boldsymbol{x}')\frac{\partial G(\boldsymbol{x},\boldsymbol{x}')}{\partial n'}\right] da'}_{\text{need normal derivative of Green function}},$$

  of Poisson equation with *Dirichlet boundary conditions*,

- the second term requires the *normal derivative* $\partial G/\partial n'$.



$\boldsymbol{n}'$ is the unit normal vector *outward from the volume of interest*.

- Thus, for the solution *outside the sphere* (top left figure), it is inward along $\boldsymbol{x}'$, toward the origin and

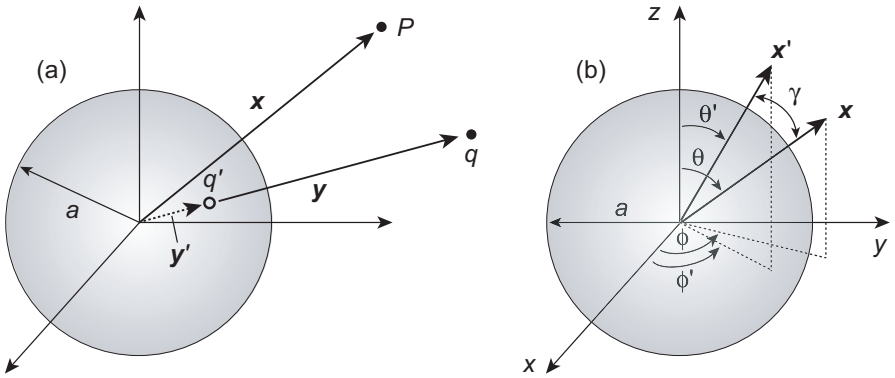
$$\left.\frac{\partial G}{\partial n'}\right|_{x'=a} = -\frac{x^2-a^2}{a(x^2+a^2-2ax\cos\gamma)^{3/2}}.$$

- Therefore, 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

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

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

$$\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi-\phi').$$



- If there *is a charge distribution* $\rho(\boldsymbol{x}')$, add the first term of

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

- with the associated Green function

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

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 changes.
- Thus the interior solution is

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

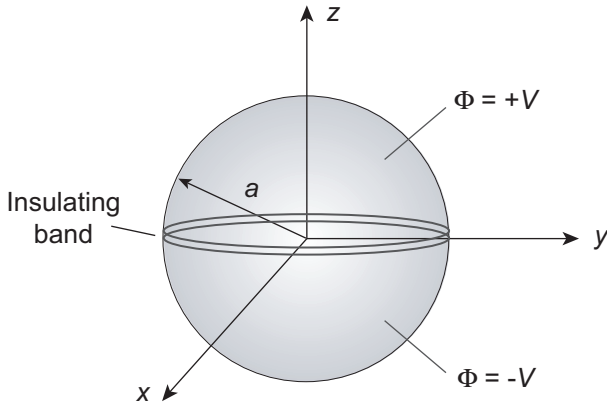


Figure 3.8: (Example 3.2) Two hemispheres of radius a separated by an insulating band lying in the $z = 0$ plane, with the upper hemisphere kept at potential $+V$ and the lower hemisphere kept at potential $-V$.

Example 3.2

Let's illustrate use of the *exterior solution* by considering a *conducting sphere of radius* a, divided into *two hemispherical shells separated by an insulating ring*, as illustrated in Fig. 3.8.

- From

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

  the solution for $\Phi(x,\theta,\phi)$ is

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

- 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

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

- Because of the complicated dependence among the angles in

$$\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi-\phi').$$

this *cannot be integrated easily in closed form*.

- If one *restricts to the positive z axis* the integrals can be done, with the result

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

which correctly reduces to $\Phi = V$ at $z = a$.

- Of potentially more use is to *expand the denominator in a power series* and integrate term by term, which yields

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

- This expansion has been shown to *converge rapidly for large* x/a, and agrees with the special solution on the positive z axis for $\cos\theta = 1$.

3.11 Solving the Poisson and Laplace Equations Directly

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

3.11.1 Separation of Variables in Cartesian Coordinates

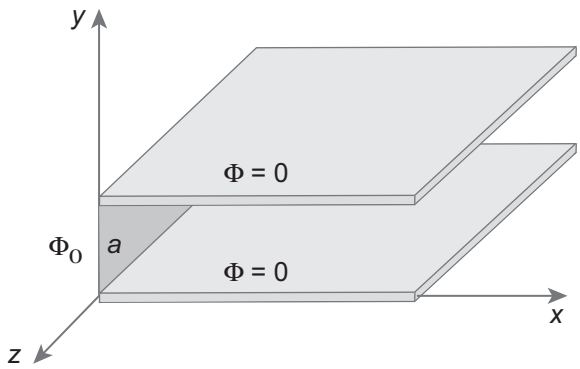
A typical case to be solved:

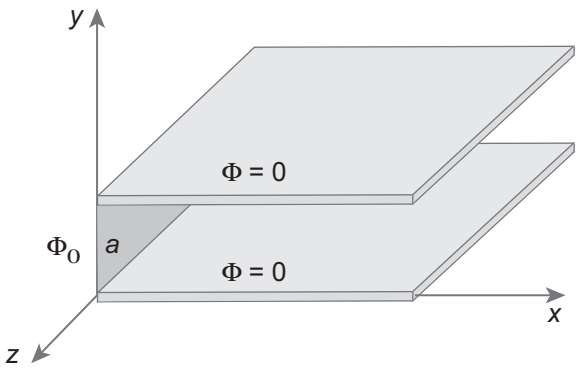
- the *potential* or the *charge density* is *specified on the boundaries* of a region, and
- we wish to *find the potential in the interior*.
- A *standard approach* to such a problem is to solve Laplace's equation by the *separation of variables method*, which we consider initially in *cartesian coordinates*.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

$$\Phi(x,y,z) = X(x)Y(y)Z(z),$$

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 the following figure.





- Electrodes assumed to extend infinitely in the x and z directions.
- 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*,

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

subject to the *boundary conditions*

$$\begin{aligned} \Phi &= 0 \qquad && (y = 0, y = a), \\ \Phi &= \Phi_0 \qquad && (x = 0), \\ \Phi &\to 0 \qquad && (x \to \infty). \end{aligned}$$

- Inserting the 2D product function

$$\Phi(x,y) = X(x)Y(y)$$

and dividing through by $\Phi = X(x)Y(y)$ gives

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

- where we write the derivatives as ordinary derivatives since $X(x)$ and $Y(y)$ are functions of a single variable.
- Now the second term 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,

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

- This 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*

$$\frac{d^2X}{dx^2} - k^2X = 0,$$
$$\frac{d^2Y}{dy^2} + k^2Y = 0.$$

As verified by substitution, *the second equation has a solution*

$$Y = A\sin(ky) + B\cos(ky),$$

where A and B are arbitrary constants.

- *Determine constants* by imposing *boundary conditions*.
- $\Phi = 0$ at $y = 0$ requires that $B = 0$, and
- $\Phi = 0$ at $y = a$ requires that

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

- Therefore,

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

- The equation for X now takes the form

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

  which has a solution

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

- However, the boundary condition $\Phi \to 0$ as $x \to \infty$ can be satisfied only if $G = 0$, so we discard the first term and insert previous results to give

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

  with C another arbitrary constant.

The solution still does not satisfy the boundary condition $\Phi = \Phi_0$ at $x = 0$.

- 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 in the form
$$\Phi(x,y) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right) e^{-n\pi x/a},$$
- The boundary condition $\Phi = \Phi_0$ at $x = 0$ implies that
$$\Phi(0,y) = \Phi_0 = \sum_{n=0}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right),$$
which is a *Fourier sine series*,
- so we can *use the corresponding orthogonality properties* to *evaluate the coefficients* c_n.
- Specifically, *multiply both sides by* $\sin[(pxy)/a]$, where p is an integer, and *integrate from* $y = 0$ *to* $y = a$,
$$\int_0^a \Phi_0 \sin\left(\frac{p\pi y}{a}\right) dy = \int_0^a \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi y}{a}\right) \sin\left(\frac{p\pi y}{a}\right) dy.$$

For the integral on the left side of this equation

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

while for the integral on the right side,

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

Comparing, we conclude that the expansion coefficients are

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

and the potential as a function of x and y is

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

- This series should converge fairly rapidly because of the rapid decrease of the factor $e^{-n\pi x/a}/n$ with n.
- A convergent power series is adequate to define a solution, but it happens that *the series can be summed exactly*, giving

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

  in a more convenient closed form.

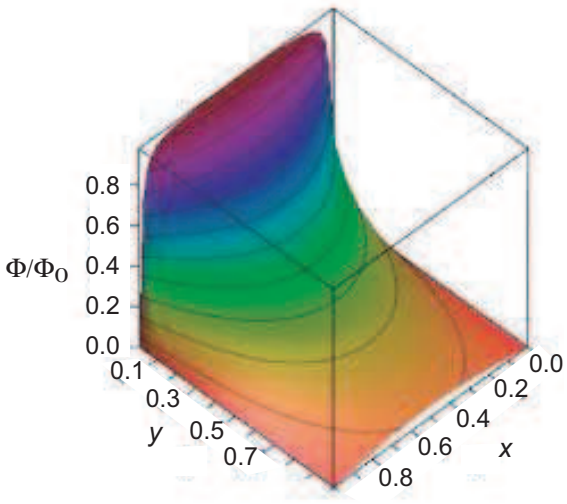
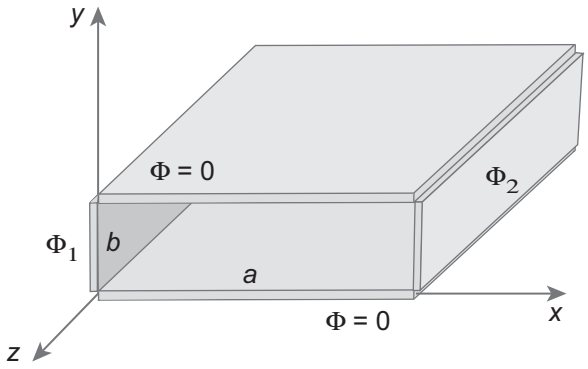


Figure 3.9: Solution for 2D Laplace equation by separation of variables. (Excuse quality; temporary placeholder).

The solution is plotted in Fig. 3.9.

Example: *Field between conducting plates*

Consider the following figure,



where we wish to calculate the electrostatic potential in the region between the parallel-plane electrodes (which extend infinitely in the z direction).

- There is *no z dependence* so again let's solve the 2D Laplace equation by the method of separation of variables.
- This problem has much in common with the example just worked out, but the *boundary conditions are different*.

$$\begin{aligned} \Phi &= 0 \qquad (y = 0, y = b), \\ \Phi &= \Phi_1 \qquad (x = 0), \\ \Phi &= \Phi_2 \qquad (x = a), \end{aligned}$$

- The most general solution is of the form

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

where the constants A_n and B_n may be *determined using the boundary conditions*.

- From the *boundary condition* $\Phi = \Phi_1$ *at* $x = 0$ we have

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

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

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

implying that

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

- The boundary conditions $\Phi = \Phi_2$ at $x = a$ yields a second relationhip between A_n and B_n:

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

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

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

- Then,

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

where $n = 1, 3, 5, \cdots$. This gives $\Phi(x, y)$ when inserted in

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

This solution is plotted in Fig. 3.10.

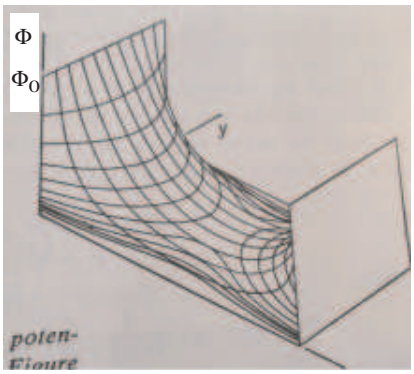


Figure 3.10: Solution for 2D Laplace equation by separation of variables. (Excuse quality; temporary placeholder).

3.11.2 Separation of Variables in Spherical Coordinates

In the preceding examples of solving the Laplace equation by separation of variables the geometry was rectangular and cartesian coordinates were appropriate.

- However, for some problems other coordinate systems such as *spherical* or *cylindrical* may be more natural.
- In this section we consider *solution of Laplace's equation in spherical coordinates* (r, θ, ϕ)

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

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

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

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

$$\Phi(r,\theta)=R(r)\,\Theta(\theta),$$

where $R(r)$ is a function only of r and $\Theta(\theta)$ is a function only of θ.

Substituting $\Phi(r,\theta) = R(r)\,\Theta(\theta)$ and dividing through by $R\Theta$,

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

where now we use total rather than partial derivatives.

- The second term is independent of r so the first term must also be independent of r.
- Thus we write the separated equations as two ODEs

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = k \qquad (R \text{ equation})$$

$$\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) = -k \qquad (\Theta \text{ equation})$$

  where the constants are $k = -k$, since the *sum of the two equations must be zero*.
- Consider first the R equation. Multiplying both sides by R and carrying out the leftmost d/dr operation gives

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

  which has a solution

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

  that requires n and k to be related by,

$$n(n+1) = k.$$

  when inserted into the differential equation.

Now consider the Θ *equation*. Multiplying by $\Theta \sin\theta$, it may be written

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

- This is a *famous equation* but it isn't in its standard form.
- To recognize it, *let's change variables* by letting $\mu = \cos\theta$.
- By the *chain rule*, for any function $f(\mu)$ of μ,

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

  where we have used $d\mu/d\theta = -\sin\theta$ and $1-\mu^2 = \sin^2\theta$.

- Then,

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

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

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

- A *general solution of Laplace's equation* in *spherical coordinates assuming axial symmetry* is then given by

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

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

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

- which is important in evaluating the coefficients in the equation above for $\Phi(r,\theta)$.

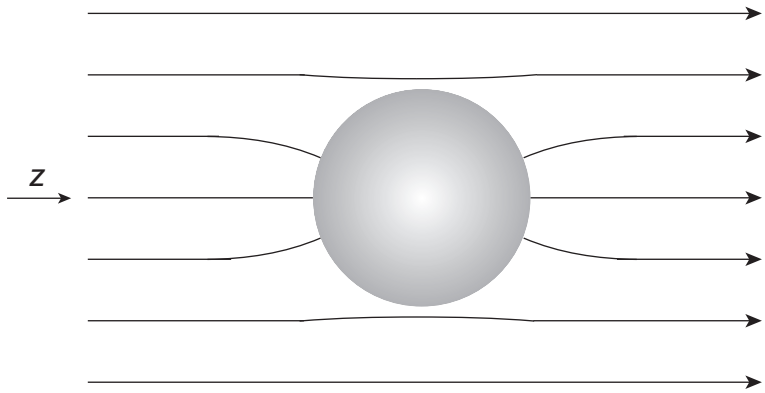


Figure 3.11: Conducting ball of radius a in an external electric field $\boldsymbol{E}_0$ directed along the $\boldsymbol{z}$ axis. Assume axial symmetry around the left-right axis.

Example: *Consider the conducting ball of radius a in an electric field illustrated in Fig. 3.11.*

- Inside the ball the field will be zero.
- Far outside it will be the undisturbed field $\boldsymbol{E}_0$.
- Near the ball the field will be *distorted by polarization*.
- Let's *solve the Laplace equation for the outside field* in 2D by separation of variables in spherical coordinates, assuming axial symmetry.
- We take as *boundary conditions*

$$\Phi = 0 \qquad (r = a)$$
$$\Phi = -E_0 r\cos\theta \qquad (r = \infty)$$

- At $r = a$ (radius of ball),

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

- The coefficients A_n and B_n may be evaluated using the *orthogonality of the Legendre polynomials*.
- Multiply

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

by $P_m(\cos\theta)$ and integrate over $d(\cos\theta)$ using

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

- *The only non-vanishing terms are those for which* $n = m$. Thus,

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

- from which the coefficients are related by

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

- Therefore,

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

Now as $r \to \infty$ the second term in parentheses in the above equation becomes negligible compared with the first, and the boundary condition

$$\Phi = -E_0 r \cos\theta \qquad (\text{as } r \to \infty)$$

requires that

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

- 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 all the $B_n = 0$ except for

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

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

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

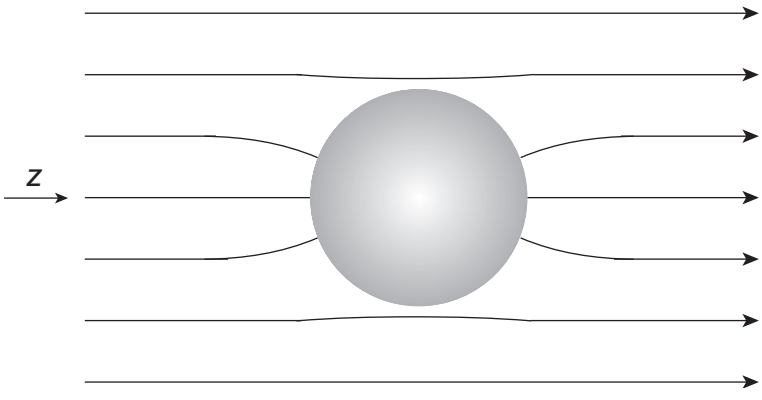
- where the *first term comes from the applied field* E_0 and
- the second term is the *induced polarization potential*.
- The *electric field components* follow by taking the gradient,

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

- At the surface of the conductor, we know that $E_r|_{r=a} = \frac{\sigma}{\varepsilon_0}$, so solving for σ gives

$$\sigma = 3\varepsilon_0 E_0\cos\theta,$$

for the *induced charge density*.



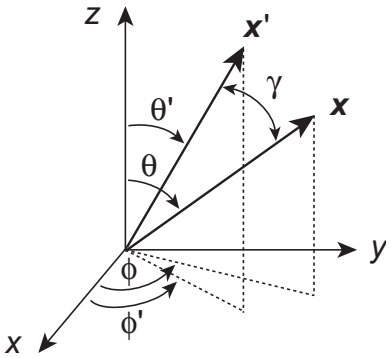


Figure 3.12: Geometry for the Legendre polynomials and the spherical harmonics.

3.12 Multipole Expansions

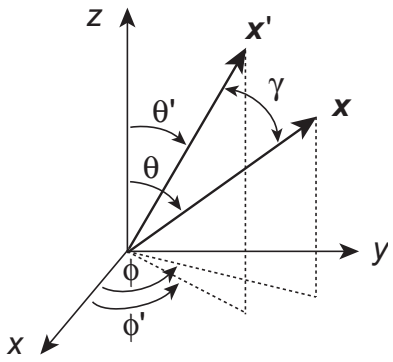
The *3D volume charge density* $\rho(\boldsymbol{x})$,

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

describes a charge localized in some finite volume V.

- In many situations we are interested in the potential $\Phi(\boldsymbol{x})$ at large distances such that $|\boldsymbol{x}| \gg |\boldsymbol{x}'|$.
- Then a series expansion of $1/|\boldsymbol{x}-\boldsymbol{x}'|$ is suggested.
- Two types of expansions are common in the literature:
  1. A Taylor series in the cartesian coordinates $(x, y.z)$.
  2. An expansion in terms of spherical harmonics depending on spherical coordinates (r, θ, ϕ).

> We will primarily discuss the multipole expansion in spherical harmonics or Legendre polynomials.



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

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

where $P_l(\cos\theta)$ is a Legendre polynomial, $r_<$ is the smaller and $r_>$ the larger of $|\boldsymbol{x}|$ and $|\boldsymbol{x}'|$, and θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$ (see figure above).

This 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 the following.

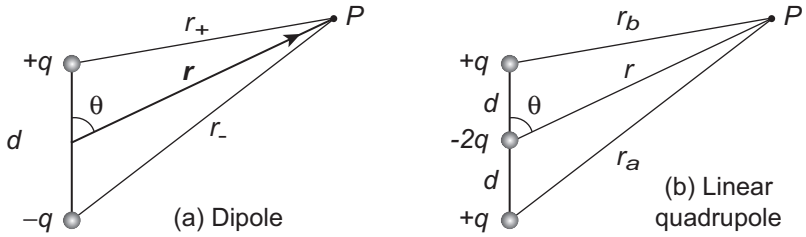


Figure 3.13: Point charge distributions for a (a) dipole, and (b) linear quadrupole.

What is the potential generated at P in Fig. 3.13?

- *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 the distances are

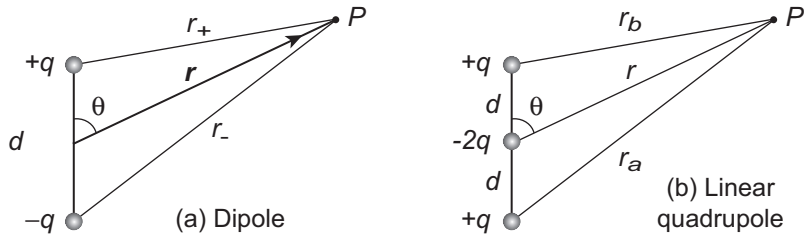
$$\begin{aligned} r_\pm^2 &= r^2 + \left(\frac{d}{2}\right)^2 \mp rd\cos\theta \\ &= r^2\left(1 \mp \frac{d}{r}\cos\theta + \frac{d^2}{4r^2}\right) \simeq r^2\left(1 \mp \frac{d}{r}\cos\theta\right), \end{aligned}$$

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

- Thus,

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

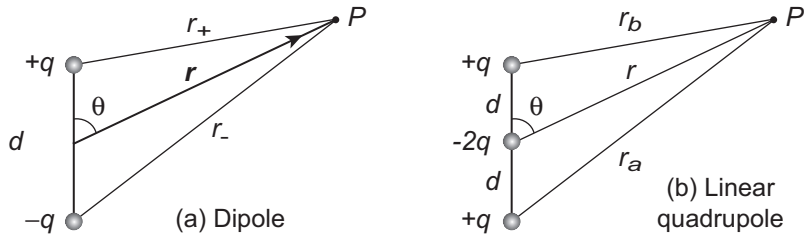
  where a *4*binomial expansion was used in the last step.



- Therefore,

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

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



- *Quadrupole Potential:* Carrying out a similar analysis for the linear quadrupole in Fig. (b) above, the linear quadrupole potential is given by

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

  which is proportional to $P_2(\cos\theta)$ and varies as $1/r^3$.

- Comparison of these expressions with the terms in

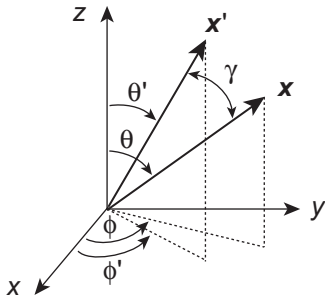
$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \sum_{l=0}^{\infty}\frac{r_<^l}{r_>^{l+1}}P_l(\cos\theta),$$

  using that

$$r_> = r \qquad r_< = d \qquad (\text{since } r \gg d)$$

  suggests why this equation 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*.)



The multipole expansion

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

can be expressed in terms of spherical harmonics using the *spherical harmonic addition theorem*

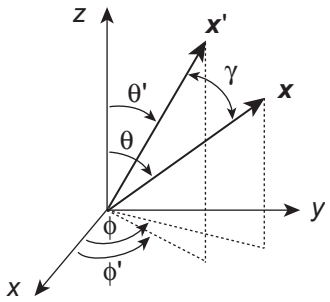
$$P_l(\cos\gamma) = \frac{4\pi}{2l+1}\sum_{m=-l}^{l} Y_{lm}^*(\theta',\phi')\,Y_{lm}(\theta,\phi),$$

- 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'),

where

$$\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi-\phi').$$

is the angle γ between the vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$.



The spherical harmonic addition theorem

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

may then be used to express the expansion

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

of the potential at $\boldsymbol{x}$ due to a unit charge at $\boldsymbol{x}'$ as

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

where angles are defined in the figure above.

This expression gives the potential *completely factorized in the coordinates* $\boldsymbol{x}$ *and* $\boldsymbol{x}'$.

If the localized distribution of charge $\rho(\boldsymbol{x}')$ 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

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

- The expansion coefficients q_{lm} are given by

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

  and are called *multipole moments*.

- If we define

  1. a total charge Q (monopole moment),
  2. an electric dipole moment vector by

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

  3. and a *quadrupole moment tensor* Q_{ij} by

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

the multipole moments q_{lm} in

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

may be written explicitly in cartesian coordinates as

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

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

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

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

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

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

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 $1/|\boldsymbol{x}-\boldsymbol{x}'|$. We quote the result without proof,

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

The utility of a multipole expansion like

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

or like

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

is that at large distance from the source-charge distribution,

- the potential may be well approximated by retaining only the first few terms in the multipole expansion, since
- the terms fall off with distance r as $r^{-(l+1)}$.

Example 3.5

Consider a localized charge density

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

Let's make a *multipole expansion* of the potential associated with this charge density and determine all the *non-vanishing multipole moments*.

The charge distribution is *axially symmetric*, so only Y_{lm} with $m = 0$ are non-zero. The *moments may be written*

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

where we have *used in the third line*,

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

and *used in the fifth line*,

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

and *used in the last step*

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

and the *radial integral was evaluated* using the tabulated definite integral

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

Thus the *multipole moments are*

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

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

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

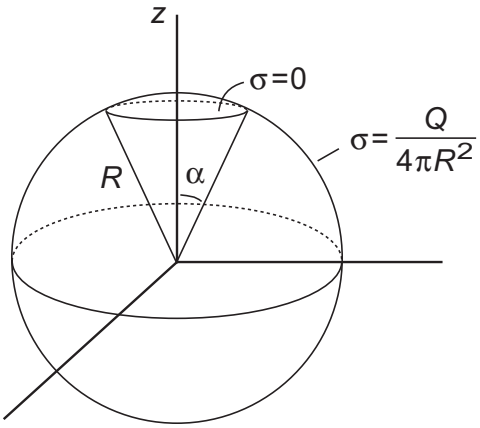


Figure 3.14: Example 3.6: A spherical conducting surface has a uniform surface charge of density $\sigma = Q/4\pi R^2$, except for a spherical cap at the north pole defined by a cone with opening $\theta = \alpha$ where $\sigma = 0$.

Example 3.6

A *spherical conducting 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. 3.14. *Use the jump condition at a charge layer* 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 3.4),

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

which allows us to solve for the potential $\Phi(r,\theta)$.

- Because of the *azimuthal symmetry about the z axis* (no dependence on ϕ) we may *expand the potential in Legendre polynomials*,

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

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

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

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 of*

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

  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

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

  and *solving for* A_l,

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

- 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

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

- The expresssion

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

can be integrated by inserting *(primes indicate derivatives)*,

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

- which gives

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

where in the last step $P_l(-1) = (-1)^l$ *was used*.

- Substituting in the original expansions, for the *inside solution*,

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

- and for the *outside solution*,

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

3.12.1 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

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

with definite (l,m) the *spherical electric field components* are

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

with q_{lm} *the multipole moments*.

Example 3.7

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

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

for the electric-field components.

3.12.2 Energy of a Charge Distribution in an External Field

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

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

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

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

where *in line 2 the definition of the electric field* $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ was used.

Inserting the expansion in the first equation, the energy takes the form

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

where q *is the total charge*, the *dipole moment* $\boldsymbol{p}$ is

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

and the *quadrupole moment* Q_{ij} is

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

Example 3.8

Many atomic nuclei have *charge distributions exhibiting a quadrupole deformation*.

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

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

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

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.

Chapter 4

Electrostatics in Matter

To this point we have considered electrostatics in vacuum,

- but many applications of electromagnetic theory involve interactions in matter.
- Therefore, in this chapter we begin to address how the properties of matter influence the equations of electrostatics.
- The influence of matter on electrostatics will in most cases require some amount of *averaging over the detailed and complex microscopic interactions in the matter*.
- One important concept will be to compute the *average electric field* in some volume of matter.
- We will find that *electric fields in matter are largely dipole fields*.

The net dipole moment induced in matter has *two basic sources:*

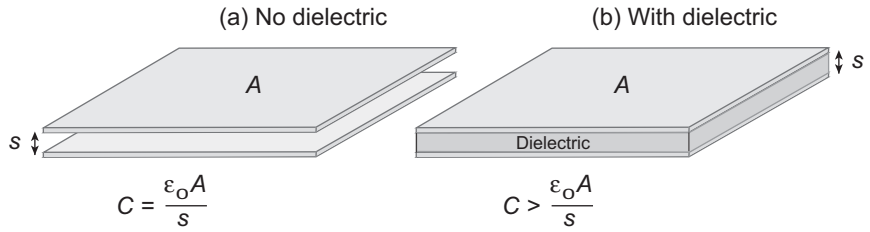
1. *An electric field polarizes matter*, even if the atoms or molecules have no significant dipole moment in the absence of the applied field. This polarization effect is characterized by a quantity called the *atomic polarizability*.
2. *Some molecules have an intrinsic dipole moment*, and the external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

In either case the polarization of the material my be quantified in terms of a

- *polarization density* $\boldsymbol{P}$ and
- an *electric susceptibility*, which is proportional to the ratio of $\boldsymbol{P}$ to the electric field.
- The *net effect of the polarization density* 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*.
- A capacitor with dielectric material between the metal plates has *increased capacitance because of this induced surface charge*.

Capacitors and Dielectrics

The *simplest 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.
- This is because the *induced surface charge on the dielectric* produced by the electric field between the plates partially *cancels the opposite charge on the adjacent plate*.

Understanding this mechanism leads to fundamental insight into the role that an electric field plays in a dielectric.

4.1 Dielectrics

Let us investigate in more depth the effect of a dielectric on a parallel plate capacitor. A parallel-plate capacitor with *no material between the plates* has a capacitance C defined by

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

- Q is the *magnitude of the charge* on the plates,
- Φ_{12} is the *difference in potential* between the two plates,
- A is the *surface area of a plate*, and
- s is the *separation of the plates*.

4.1.1 Capacitors and Dielectrics

Now suppose that a layer of dielectric material is placed between the capacitor plates.

- One will generally find that the capacitance can still be defined by the ratio of charge to potential difference between the plates, $C = Q/\Phi_{12}$, but
- the actual value of C will be *increased* over that found with no dielectric between the plates.

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

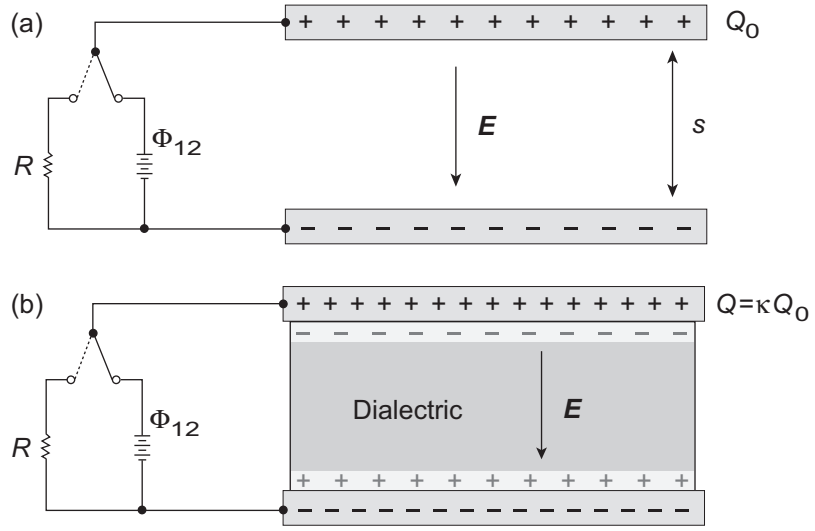
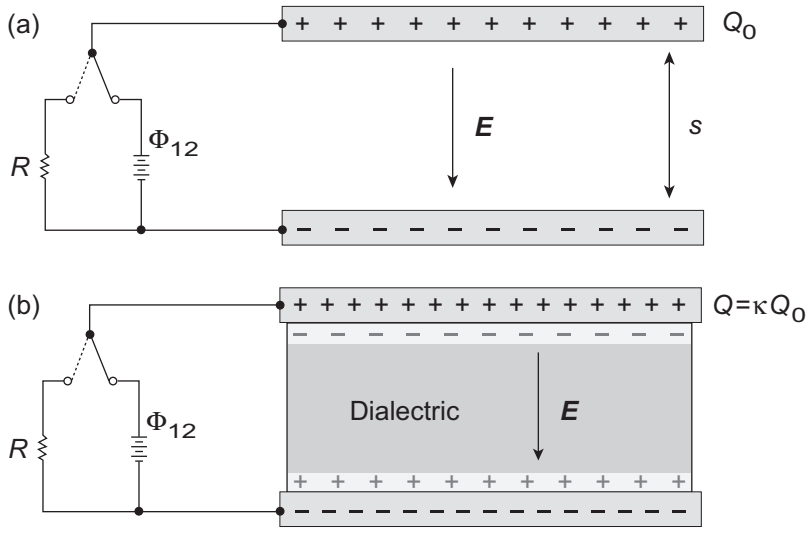


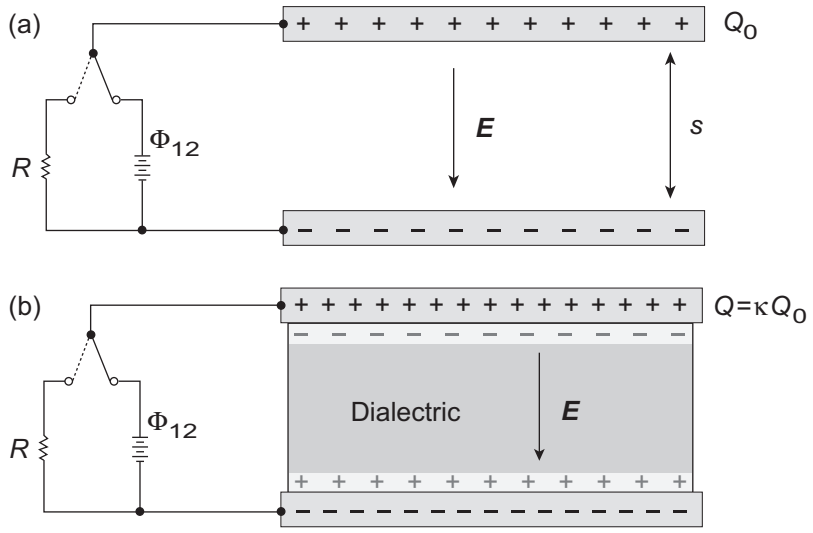
Figure 4.1: Increase of capacitance by insertion of a dielectric between the plates. (a) No dielectric. (b) With dielectric.

Qualitatively, this influence of the dielectric on the capacitance is easily understood.

- The material of the dielectric consists of atoms or molecules with negatively charged electrons and positively charged nuclei.
- The electrical field between the plates *polarizes the charge distribution of the dielectric*.
- If we assume the upper plate to have positive charge and the lower plate negative charge, *negative charges will be pulled upward* and *positive charges pushed down*.
- This is illustrated in Fig. 4.1,



- This has the effect of *exposing a layer of uncompensated negative charge near the top* of the dielectric and a *layer of uncompensated positive charge near the bottom*.
- The charge Q on the upper plate will increase because of the induced top layer of negative charge below it.
- We shall show later that Q must increase until the algebraic sum of Q and the induced charge layer is equal to the total charge on the top plate Q_0 before the dielectric layer was inserted.
- Thus, the total charge Q in the top layer is *larger than the charge* Q_0 of the top plate in Fig. (a) before the dielectric was inserted.
- Thus, the charge is the Q appearing in the dielectric equation, and is, in the above circuit the *charge supplied by the battery* while the switch is in the right (charging) position.



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

Table 4.1: Dielectric constants κ of some substances

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

4.1.2 Dielectric Constants

Different dielectric materials would be expected to have different efficiencies for increasing the charge capacity of a capacitor, according to the ease with which electrons can be displaced with respect to the atomic nuclei by the applied electric field.

- The factor Q/Q_0 by which the charge and capacitance is increased is called the *dielectric constant* κ .

$$Q = \kappa Q_0 \quad \leftrightarrow \quad C = \kappa C_0.$$

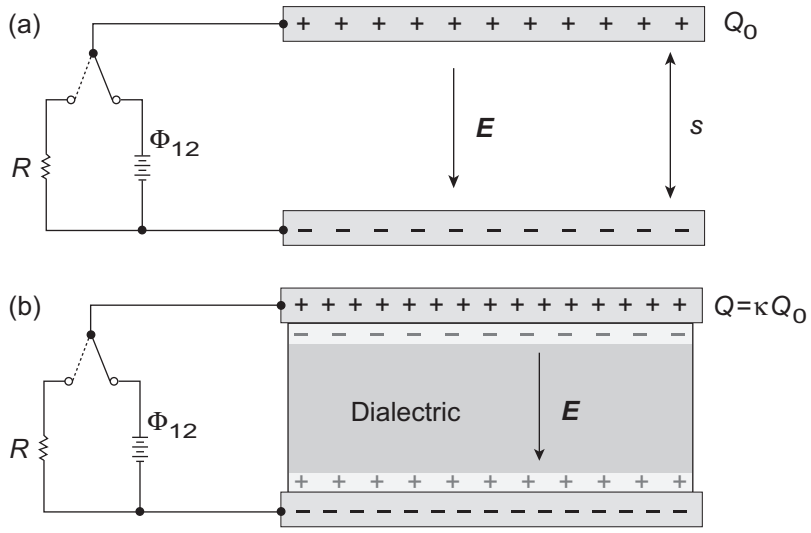
- Dielectric constants are dimensionless; a few are given in Table 4.1 for some representative substances.
- Note that the *dielectric constant of the vacuum is* $\kappa = 1.0000$ (by definition),
- typical gases have κ slightly larger than one,
- and liquids and solids can have dielectric constants varying widely in the range $\kappa \sim 1 - 100$.

The reason for the remarkably large value $\kappa = 80.4$ for water merits an explanation that will be given later.

4.1.3 Bound Charge and Free Charge

In considering the effect of a dielectric on a capacitor, it is useful to introduce some *terminology distinguishing between the charge associated with the dielectric itself and the mobile charges that can charge the plates*.

- Those charges associated with the dielectric are termed *bound charges*.
- They are *not mobile* because they are *attached to the atoms and molecules making up the dielectric*.
- They can be polarized through *electric fields causing tiny displacements*, but
- *if the capacitor is discharged the bound charges remain with the dielectric*, which becomes unpolarized as the electric field vanishes.
- The bound charges are *not part of the charge Q of the capacitor plates*.



On the other hand, those charges that are *not bound in dielectrics* are termed *free charges*.

- They are the charges that we have some agency over in an experiment.
- (for example, through manipulation of the simple electrical circuit using the switch to charge or discharge the capacitor).

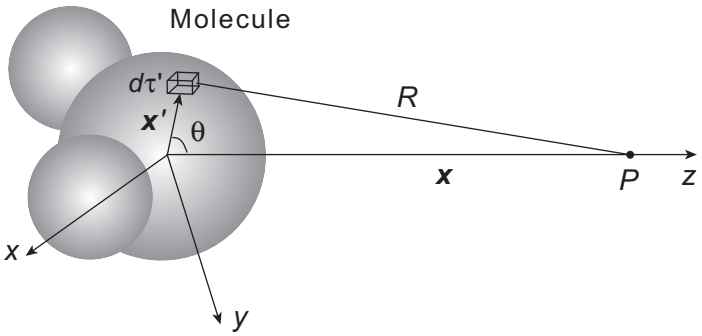


Figure 4.2: Diagram for calculation of the potential at point P of a molecular charge distribution.

4.2 Moments of a Molecular Charge Distribution

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

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

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

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

where we define $r = |\boldsymbol{x}|$ and $r' = |\boldsymbol{x}'|$, and the *equation for the potential becomes*

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

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

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

Then from the preceding equations *the potential can be written as*

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

where $r = |\boldsymbol{x}|$ is a constant and has been brought outside the integrals (the integration variable is $\boldsymbol{x}'$).

- Now this is a *power series in* $1/r$ *with coefficients that are constants* depending only on integrals over the charge distribution and independent of the distance to P.
- Thus we can write the potential as a *power series (multipole expansion) with constant coefficients*

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

where the *coefficients* C_i *are the integrals over the internal charge distribution* in the preceding equation.

The *electric field then follows from* $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi_P$. This result is valid only along the z axis, but the power series

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

illustrates the central point:

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

Let us look at the coefficients in

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

more closely.

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

As we shall see, for our main task here of *understanding the behavior of dielectrics* only

- the *monopole strength* (the total charge) and
- the *dipole strength* of the molecular building blocks of the dielectric

are *important in determining its electric-field properties*.

- Thus, for the *properties of dielectrics* we will find that *all multipole moments of the charge distribution of order greater than the dipole can usually be ignored*.
- A net dipole moment can come about because of *induced polarization by an electric field*, or
- because of *molecules that have a permanent dipole moment*.

We address induced moments and permanent dipole moments in the following sections.

4.3 Induced Dipole Moments

The *simplest atom is hydrogen*, consisting of one nucleus and one electron.

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

The electron cloud of the undisturbed hydrogen atom is *spherically symmetric* but

- if it is placed in an electric field pointing upward along a z-axis, *the charge cloud of the atom will be distorted*,
- with the *negative electron charge cloud pulled down and the nucleus pushed up*.
- This distorted atom will have a *net 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* $\sim e\Delta z$, where e is the total electron charge.

We may make a very *rough estimate of how much distortion will be caused by a field of strength* E by noting that

- *a strong electric field already exists holding the unperturbed H atom together*, which may be estimated as

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

  where a is a characteristic atomic length scale (for example, some multiple of the Bohr radius).

- If we assume that a field of the same order of magnitude would be required to produce significant distortion, we may estimate the distortion at

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

- Inserting typical numbers suggests that $4\pi\varepsilon_0 a^2 \sim 10^{11}$ volts/m, which is enormous (thousands of times larger than any field produced in a laboratory).

- Obviously the *distortion of the atom will be tiny indeed*.

- The corresponding *induced dipole moment vector* $\boldsymbol{p}$ will point in the direction of $\boldsymbol{E}$.

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

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

It is common to report atomic polarizabilities as $\alpha/4\pi\varepsilon_0$.

Table 4.2: 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

Some *atomic polarizabilities are given in Table 4.2*. Elements are listed there in order of increasing electron number. The *large variations in polarizability* may be attributed to

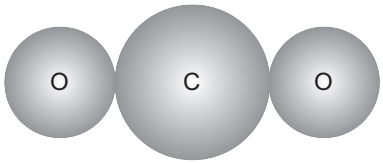
- *differences in electron number* and to
- detailed *valence electronic structure*.

Let's estimate the charge displacement induced by a typical electric field.

From the polarizability of atomic hydrogen in Table 4.2, the magnitude of the dipole moment induced by an *electric field of one megavolt/meter* is $p < 10^{-34}$ coulomb-meters.

Asymmetric Polarization

Molecules may be very asymmetric and can 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 $\boldsymbol{p} = \alpha \boldsymbol{E}$ 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 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.

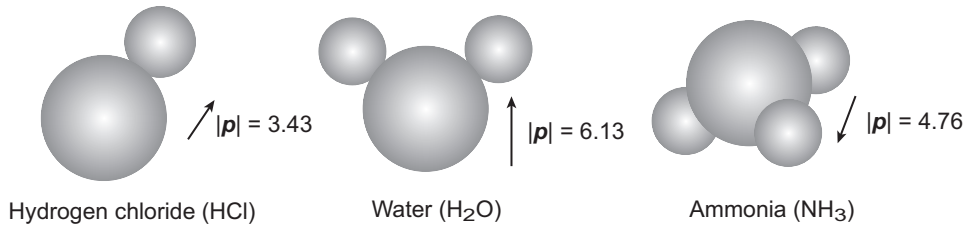
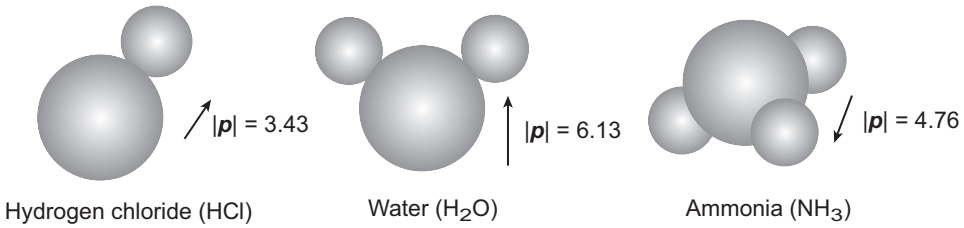


Figure 4.3: Approximate geometry of the electron cloud and the observed dipole moment vector $\boldsymbol{p}$ for some polar molecules. Magnitudes of the dipole moment vector are in units of 10^{-30} coulomb-meters

4.4 Permanent Dipole Moments

Some molecules have asymmetric shapes and permanent dipole moments in their normal ground state, even in the absence of an external field; a few examples are shown in Fig. 4.3.

- Molecules with a permanent dipole moment are termed *polar molecules*.
- As a rule, *intrinsic dipole moments (when they exist) are much larger than induced dipole moments*.



As shown in the above figure,

- the magnitude of the (permanent) dipole moment for the molecule HCl is

$$p = 3.43 \times 10^{-30} \text{ coulomb-meters}$$

- which is equivalent to shifting one electron by a distance of 0.2 angstrom (0.2×10^{-8} cm).
- From an earlier example, 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 representative and when they exist *permanent dipole moments are typically orders of magnitude larger than those induced by laboratory electric fields.*

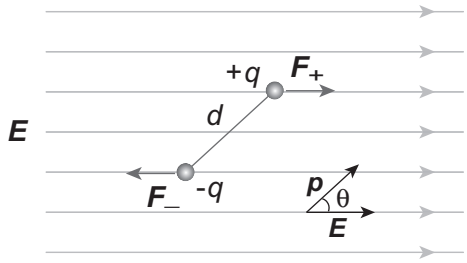


Figure 4.4: Torque applied to a molecule with an intrinsic dipole moment in a uniform electric field.

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. 4.4),
- but the *torque* $\boldsymbol{N}$ *acting on the dipole* is

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

  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*.

4.5 Polarization

We have discussed two basic mechanism that lead to polarization of a dielectric:

1. *distortion of the charge distribution by an electric field*, which induces many tiny dipoles pointing in the same direction as the field if the dielectric substance consists of atoms or non-polar molecules, and
2. *alignment of the existing intrinsic dipole moments by an electric field* if the substance consists of polar molecules.

The net effect of either is to produce a set of dipoles aligned or partially aligned with the electric field.

- At this point our primary interest is in the effect of this polarization, without regard to the way it was formed.
- Therefore, a measure of how polarized the matter is can be formed by asking how many dipoles N of average dipole moment $\boldsymbol{p}$ there are in a unit volume, without regard to the source of the dipoles.
- The total dipole strength of a volume element is then $\boldsymbol{p}Nd\tau$ and the *polarization density* $\boldsymbol{P}$ can be defined by

$$\boldsymbol{P} \equiv \boldsymbol{p}N = \frac{\text{dipole moments}}{\text{unit volume}},$$

  which has units of $\text{C-m/m}^3 = \text{C/m}^2$.

Let's now estimate the the potential produced by the polarized dielectric material.

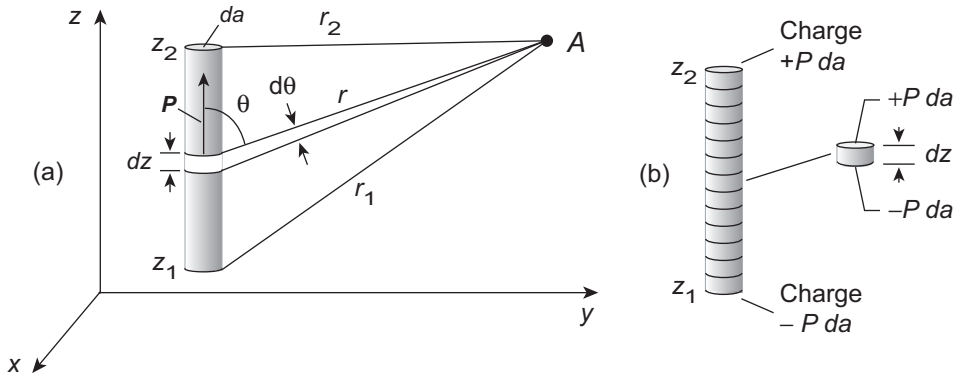


Figure 4.5: (a) A column of polarized material with cross section da observed at a distant point A produces the same field as two charges, one at each end of (b).

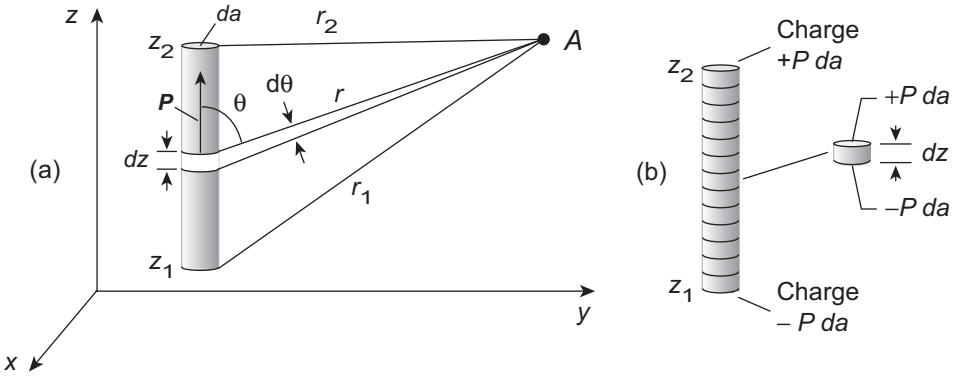
4.6 Field Outside Polarized Dielectric Matter

If the dielectric was assembled from neutral matter

- there is no net charge so there is *no monopole term* in the multipole expansion

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

- Therefore to very good approximation *only the dipole moments need be considered as sources* of a field.
- Consider a thin vertical cylinder of dielectric matter in an electric field, as illustrated in Fig. 4.5.
- The polarization density $\boldsymbol{P}$ is uniform and points in the positive z direction, and we wish to calculate the electric potential at the point A.

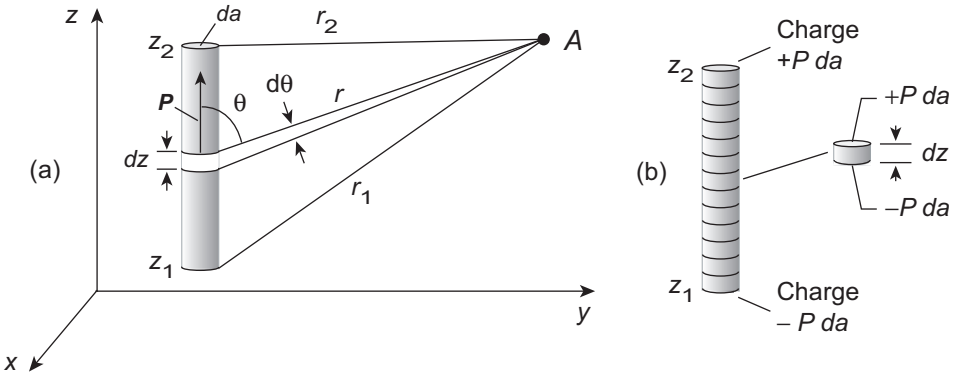


An element of the cylinder of height dz in Fig. (a) has a dipole moment

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

and its contribution to the potential at point A is

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

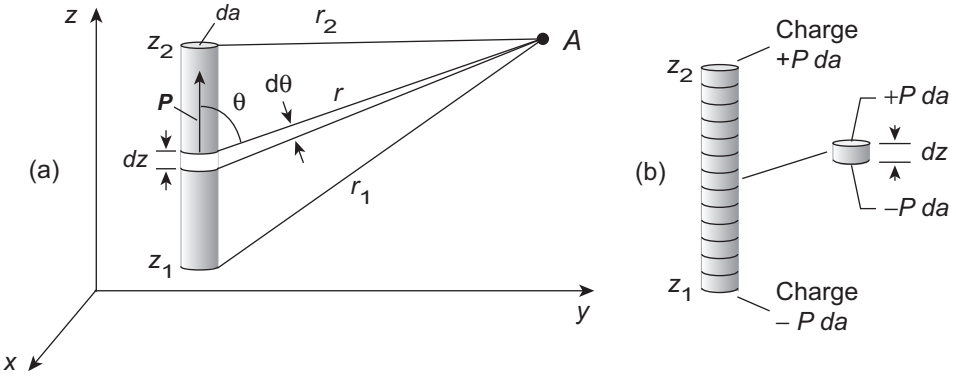


Then the potential at A produced by the entire column of polarized matter is obtained by integration,

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

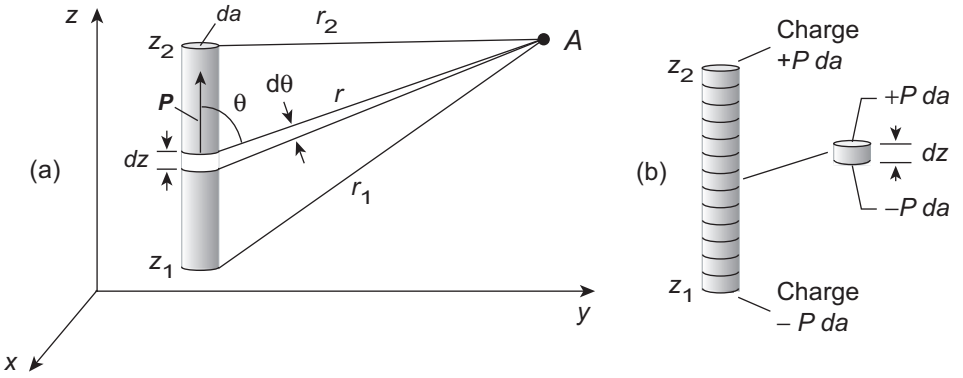
But this is exactly the same result for the potential at A that would be produced by

- a positive point charge of magnitude Pda at the top of the column at a distance r_2 from A and
- a negative point charge of the same magnitude placed at the bottom of the column a distance r_1 from A.



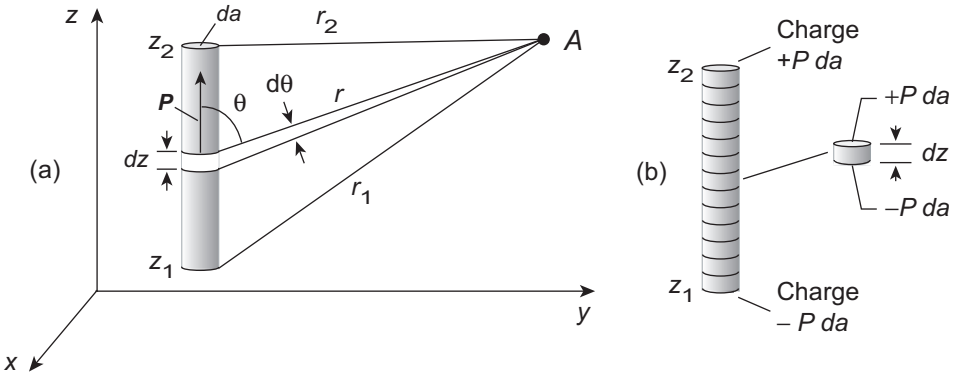
A column of uniformly polarized matter

- 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.

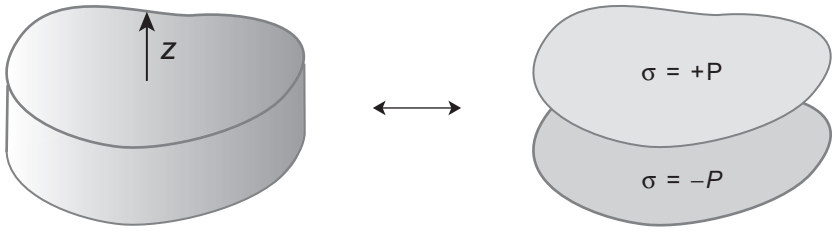


- Consider making the cylinder shown in Fig. 4.5(a) by stacking on top of each other small cylinder segments of height dz, with a charge of $+P\,da$ on its top face and $-P\,da$ on its bottom face.
- But now within the column the $+$ charge on the top of a segment will be cancelled by the $-$ charge on the bottom of the next segment up,
- *except for the top-most segment*, which will have an uncancelled charge on its top face of $+P\,da$, and
- *except for the bottom-most segment*, which will have an uncancelled charge on its bottom face of $-P\,da$.

Thus the column of tiny dipoles will seem to be a *single large dipole* with end charges $+P\,da$ and $-P\,da$, separated by a distance $z_2 - z_1$.



- Note that nowhere in this derivation have we assumed that *A* is particularly distant;
- we have assumed only that the distance to *A* is much larger than the lengths of the individual microscopic dipoles and much larger than the width of the column in Fig. 4.5(a), both of which are very small.



(a) Block of polarized dielectric (b) Two layers of surface charge

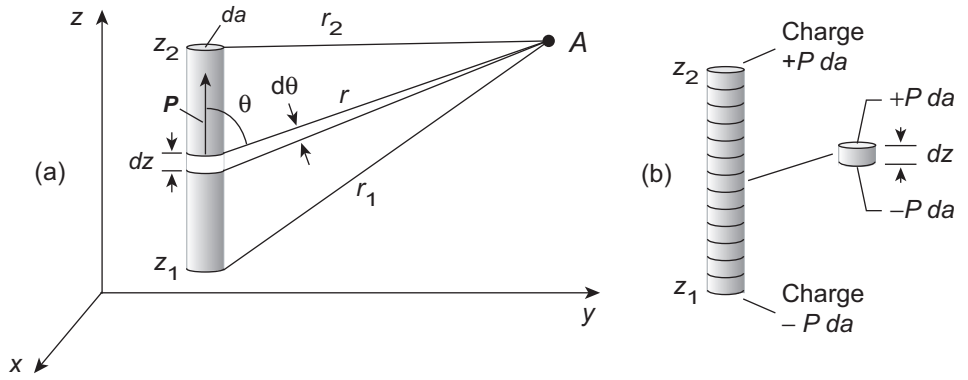
Figure 4.6: The external field due to (a) a block of dielectric polarized by an electric field in the z direction is the same as would be produced by (b) two sheets of charge $\pm P$ at the positions of the top and bottom surfaces of the block.

We can take a slab of dielectric and

- divide it up infinitesimally into such columns to
- conclude that the electric field outside the slab is
- the same as if two sheets of surface charge located where the top and bottom of the slab are located
- each carrying constant surface charge density $\sigma = +P$ and $\sigma = -P$, respectively.

See Fig. 4.6.

4.7 Field Inside Polarized Dielectric Matter



Outside the dielectric we are justified in assuming that the point *A* in the above figure is at a large distance.

- Inside the dielectric we can hardly assume that a point is at a much larger distance than the size of all the dipoles.
- Let us average over a region that is macroscopically small but microscopically large.

That is, a region large enough to suppress statistical sampling fluctuations, but small enough that ***P*** doesn't vary substantially over the volume.

The field inside polarized dielectric matter can be quite complicated.

- The spatial average of $\boldsymbol{E}$ over a volume V inside the polarized matter is given by

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

  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 the above equation.

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

1. The fundamental relation $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$ that is obeyed by the microscopic field *remains valid for the macroscopic field*.
2. This implies that the macroscopic electric field is still derivable from a scalar potential in electrostatics through $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$.

If an electric field is applied to a medium consisting of a large number of atoms or molecules.

- The dominant multipole mode response to the applied field is the dipole, with an electric polarization

$$\boldsymbol{P}(\boldsymbol{x}) = \sum_i N_i \langle \boldsymbol{p}_i \rangle,$$

- where $\boldsymbol{p}_i$ is the dipole moment of the ith species (atoms or molecules), and the average $\langle\ \rangle$ is taken over a small volume centered on $\boldsymbol{x}$ and N_i is the average number per unit volume of the ith species at $\boldsymbol{x}$.
- We can build up the macroscopic potential or field by linear superposition of contributions from each macroscopically small volume element ΔV at the variable point $\boldsymbol{x}'$.
- If there are no macroscopic multipole moments higher than dipole, then from

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

the *macroscopically averaged potential* is

$$\Delta\Phi(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{4\pi\varepsilon_0}\left[\frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\Delta V + \frac{\boldsymbol{P}(\boldsymbol{x}')\cdot(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\Delta V\right],$$

assuming $\boldsymbol{x}$ to lie outside ΔV.

Setting $\Delta V \to d^3x'$ and integrating over all space gives

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int d^3x' \left[\frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} + \boldsymbol{P}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)\right],$$

where the prime on $\boldsymbol{\nabla}'$ indicates that the divergence acts on $\boldsymbol{x}'$ instead of $\boldsymbol{x}$. Integration by parts of the second term leads to

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int d^3x' \frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\left[\rho(\boldsymbol{x}') - \boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')\right],$$

which is the usual expression for the potential generated by an effective charge distribution $\tilde{\rho}(\boldsymbol{x}') = \rho(\boldsymbol{x}') - \boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')$.

Then the first Maxwell equation (*Gauss's law*) reads

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = \frac{1}{\varepsilon_0}\left[\rho - \boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')\right],$$

which reduces to

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} = \frac{\rho}{\varepsilon_0},$$

in the absence of polarization.

The second (divergence) term appears in the effective charge density

$$\tilde{\rho} = \frac{1}{\varepsilon_0}[\rho - \underbrace{\boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{x}')}]$$

because for *non-uniform polarization* there can be a net change in charge within any small volume.

4.8 Displacement and Constituitive Relations

If we define the *electric displacement* $\boldsymbol{D}$ by

$$\boldsymbol{D} \equiv \varepsilon_0 \boldsymbol{E} + \boldsymbol{P},$$

then

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \frac{1}{\varepsilon_0} \left[\rho - \boldsymbol{\nabla}' \cdot \boldsymbol{P}(\boldsymbol{x}') \right],$$

can be written as

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho.$$

Solution for potentials or fields requires *constituitive relations that relate* $\boldsymbol{D}$ *and* $\boldsymbol{E}$.

1. A common assumption is that the *response to the applied field is linear*, $\boldsymbol{P} \propto \boldsymbol{E}$.

2. A second common assumption is that the *medium is isotropic*, so that $\boldsymbol{P}$ is parallel to $\boldsymbol{E}$ and the coefficient of proportionality has no angular dependence,

   $$\boldsymbol{P} = \varepsilon_0 \chi_e \boldsymbol{E},$$

   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

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} \qquad \varepsilon \equiv \varepsilon_0(1+\chi_e) \qquad \kappa \equiv \frac{\varepsilon}{\varepsilon_0} = 1+\chi_e,$$

where ε is *electric permittivity* and κ is the *dielectric constant*.

If the dielectric medium is *uniform in addition to isotropic*, then ε is independent of position and the divergence equation $\nabla \cdot \boldsymbol{D} = \rho$ can be written

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon}.$$

For a medium fulfilling all conditions given above,

- electrostatics problems in that medium are reduced to vacuum solutions found previously,
- except that the *electric fields must be reduced by a factor* $\varepsilon_0/\varepsilon$.

Physically this results from *polarized atoms producing fields that oppose the applied field*.

One consequence is that

- *capacitance* of a capacitor is *increased by a factor* $\varepsilon/\varepsilon_0$
- if the empty space between the electrodes is *filled with a material having dielectric constant* $\kappa = \varepsilon/\varepsilon_0$

if fringing fields can be neglected.

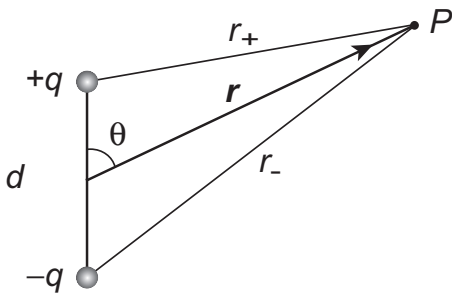
If different media are juxtaposed, *boundary conditions* must be considered *for* $\boldsymbol{D}$ *and* $\boldsymbol{E}$ *at interfaces between media*.

- This will be addressed in Section 4.11.
- The results for electrostatics are that the *normal components of* $\boldsymbol{D}$ and the *tangential components of* $\boldsymbol{E}$ on either side of an interface between medium 1 and medium 2
- must satisfy the *boundary conditions*

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n} = \sigma,$$
$$(\boldsymbol{E}_2 - \boldsymbol{E}_1) \times \boldsymbol{n} = 0,$$

  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).

4.9 Surface and Volume Bound Charges



- The potential created by a *single dipole* can be written

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\frac{\boldsymbol{p}\cdot\hat{\boldsymbol{r}}}{r^2},$$

where $r = |\boldsymbol{r}|$ and $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$ and $\boldsymbol{p}$ is the dipole moment vector.

- The *total potential contributed by dipoles* then follows by integration,

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \frac{\boldsymbol{P}(\boldsymbol{x}')\cdot\hat{\boldsymbol{r}}}{r^2}\,d\tau',$$

where $d\tau' \equiv d^3x'$ and $\boldsymbol{P}$ is the *dipole moment density*. This can be rewritten using

$$\boldsymbol{\nabla}'\left(\frac{1}{r}\right) = \frac{\hat{\boldsymbol{r}}}{r^2}$$

in the form

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \boldsymbol{P}\cdot\boldsymbol{\nabla}'\left(\frac{1}{r}\right)d\tau'.$$

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \boldsymbol{P}\cdot\boldsymbol{\nabla}'\left(\frac{1}{r}\right)d\tau'.$$

can be *integrated by parts using the product rule*

$$\boldsymbol{\nabla}'\cdot(f\boldsymbol{A}) = f(\boldsymbol{\nabla}'\cdot\boldsymbol{A}) + \boldsymbol{A}\cdot(\boldsymbol{\nabla}' f).$$

Setting $f = 1/r$ and $\boldsymbol{A} = \boldsymbol{P}$, and *integrating both sides over the volume* V, the product rule becomes

$$\int_V \boldsymbol{\nabla}'\cdot\left(\frac{\boldsymbol{P}}{r}\right)d\tau' = \int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau' + \int_V (\boldsymbol{P}\cdot\boldsymbol{\nabla}')\frac{1}{r}d\tau'$$

Now apply the *divergence theorem*

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

to the term on the left side to give

$$\oint_S \frac{1}{r}\boldsymbol{P}\cdot d\boldsymbol{a}' = \int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau' + \int_V (\boldsymbol{P}\cdot\boldsymbol{\nabla}')\frac{1}{r}d\tau'.$$

Multiply both sides by $1/4\pi\varepsilon_0$ and rearrange to give

$$\underbrace{\frac{1}{4\pi\varepsilon_0}\int_V (\boldsymbol{P}\cdot\boldsymbol{\nabla}')\frac{1}{r}d\tau'}_{\Phi(\boldsymbol{x})} = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{r}\boldsymbol{P}\cdot d\boldsymbol{a}' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau',$$

and finally, comparing with the original expression,

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int_V \boldsymbol{P}\cdot\boldsymbol{\nabla}'\left(\frac{1}{r}\right)d\tau',$$

this becomes

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{r}\boldsymbol{P}\cdot d\boldsymbol{a}' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau'.$$

Notice that in

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{r}\boldsymbol{P}\cdot d\boldsymbol{a}' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{r}(\boldsymbol{\nabla}'\cdot\boldsymbol{P})d\tau'.$$

1. The first term of looks like the potential generated by a *surface charge*

$$\sigma_b \equiv \boldsymbol{P}\cdot\hat{\boldsymbol{n}},$$

with $\hat{\boldsymbol{n}}$ the *unit vector normal to the surface* and the subscript *"b" indicates that it originates in the bound charges*.

2. The second term looks like the potential generated by a *volume charge*

$$\rho_b = -\boldsymbol{\nabla}\cdot\boldsymbol{P}.$$

This equation can thus be written

$$\Phi = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{\sigma_b}{r}da' + \frac{1}{4\pi\varepsilon_0}\int_V \frac{\rho_b}{r}d\tau'.$$

- The *potential* Φ and the *field* $\boldsymbol{E}$ of a polarized object is *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 density $\boldsymbol{P}$. Thus it contributes *only if the polarization is spatially non-uniform*].

Let's use the result

$$\Phi = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{\sigma_b}{r}da' + \frac{1}{4\pi\varepsilon_0}\int_V \frac{\rho_b}{r}d\tau'.$$

to determine the potential for a *uniformly polarized sphere of radius* R. Since we assume *uniform polarization*,

- the second (volume-charge) term makes no contribution (it is non-zero only if the derivative of the polarization density is non-zero) and the *potential is generated entirely by the surface charge*,

$$\sigma_b = \boldsymbol{P}\cdot\hat{\boldsymbol{n}} = P\cos\theta,$$

  where the z-axis is chosen as the *direction of polarization*.

Thus, we need to evaluate the *potential for a sphere with the surface charge* σ_b *painted on it*.

- This problem was *already solved* (homework),

$$\Phi(r,\theta) = \begin{cases} \dfrac{Pr}{3\varepsilon_0}\cos\theta & (r \le R), \\ \dfrac{PR^3}{3\varepsilon_0 r^2}\cos\theta & (r \ge R). \end{cases}$$

- The *electric field* is then given by $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$.
- Since $r\cos\theta = z$, *the field is uniform inside the sphere:*

$$\boldsymbol{E} = -\boldsymbol{\nabla}\Phi = -\frac{P}{3\varepsilon_0}\hat{\boldsymbol{z}} = -\frac{1}{3\varepsilon_0}\boldsymbol{P} \qquad (r < R).$$

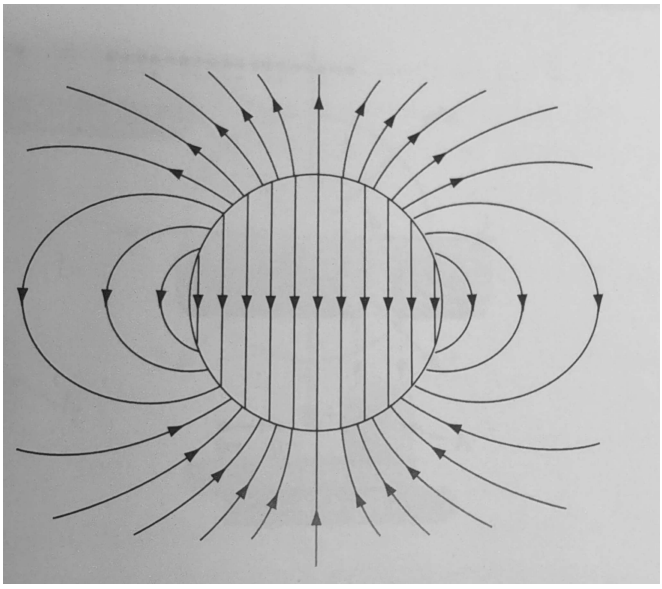


Figure 4.7: Electric field for uniformly polarized sphere.

Outside the sphere the field approximates a *dipole form*

$$\Phi = \frac{1}{4\pi\varepsilon_0}\frac{\boldsymbol{p}\cdot\hat{\boldsymbol{r}}}{r^2} \qquad (r \geq R),$$

where the dipole moment is equal to the *total dipole moment of the sphere*

$$\boldsymbol{p} = \frac{4}{3}\pi R^3 \boldsymbol{P}.$$

The *field lines for* $\boldsymbol{E}$ of the *uniformly polarized sphere* are plotted in Fig. 4.7.

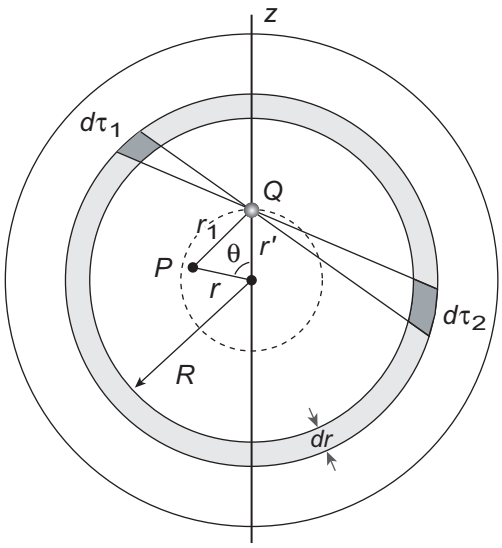


Figure 4.8: Charge Q located a distance r' from the center of a sphere of radius R.

4.10 Average Electric Fields in Matter

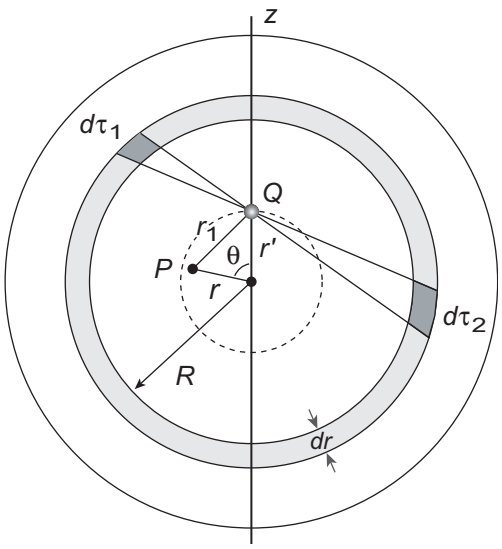
Let's begin by considering a sphere containing a *single point charge* Q located a distance r' from the origin along the z-axis, as illustrated in Fig. 4.8.

- By symmetry the average field over the entire volume must be along the z-axis The average field is then

$$\langle E_z \rangle = \frac{\int_\tau E_z \, d\tau}{\int_\tau d\tau} = \frac{1}{\tau} \int_\tau E_z \, d\tau,$$

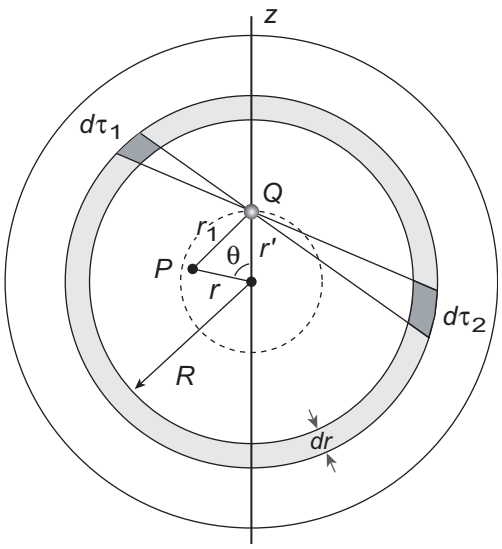
  where τ is the volume of the sphere.

- 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 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 $d\tau_1$ and 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 the above figure.

- The potential at P is

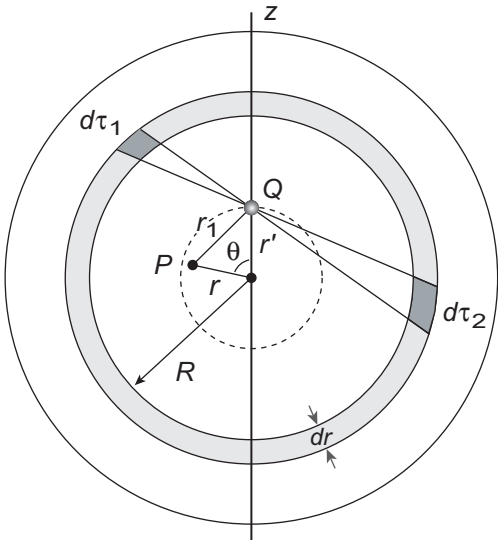
$$\Phi = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r_1} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{|\boldsymbol{x}-\boldsymbol{x}'|},$$

  where from the figure $r_1 = |\boldsymbol{x}-\boldsymbol{x}'|$.

- Thus, we may expand in the multipole expansion,

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

  where we've used from the diagram that $r < r'$.



Writing this expansion out and substituting explicit values for the Legendre polynomials,

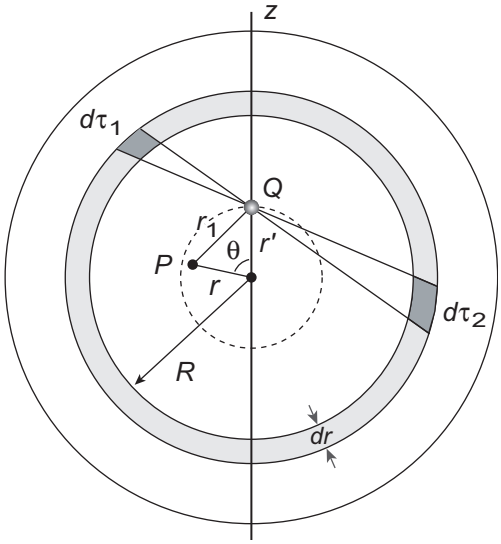
$$\Phi = \frac{Q}{4\pi\varepsilon_0}\frac{1}{r'}\left[1+\frac{r}{r'}\cos\theta+\frac{1}{2}\frac{r^2}{r'^2}(3\cos^2\theta-1)\right.$$
$$\left.+\frac{1}{2}\frac{r^3}{r'^3}(5\cos^3\theta-3\cos\theta+\cdots\right].$$

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

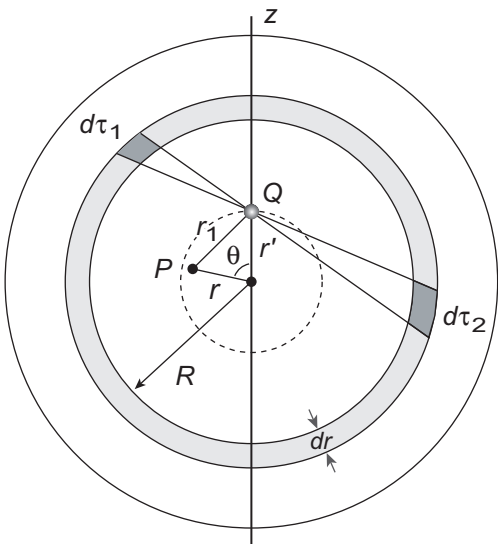
$$\langle E_z \rangle_1 = -\frac{Q}{4\pi\varepsilon_0 r'^2} \int_0^{\pi} \int_0^{r'} 2\pi r^2 \sin\theta \, dr \, d\theta = -\frac{Qr'}{3\varepsilon_0 \tau}.$$

- 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 (because of superposition), except that

$$\langle E_z \rangle = -\frac{p_{\text{total}}}{4\pi\varepsilon_0 R^3},$$

- where p_{total} is the *total dipole moment of the arbitrary charge distribution* within the sphere of radius R.

4.11 Boundary Conditions at Interfaces

We must often consider problems in which

- *media with different properties* are adjacent and
- *boundary conditions at interfaces* must be evaluated.

In a medium perhaps *polarized by electric or magnetic fields*,

- the *vacuum Maxwell equations* are modified to the *Maxwell equations in medium*,

$$\begin{aligned} \boldsymbol{\nabla}\cdot\boldsymbol{D} &= \rho && \text{(Gauss's law)},\\ \boldsymbol{\nabla}\times\boldsymbol{E}+\frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\ \boldsymbol{\nabla}\cdot\boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\ \boldsymbol{\nabla}\times\boldsymbol{H}-\frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)}, \end{aligned}$$

- where $\boldsymbol{D}$ and $\boldsymbol{H}$ are defined by

$$\boldsymbol{D} \equiv \varepsilon_0 \boldsymbol{E} + \boldsymbol{P} \qquad \boldsymbol{H} = \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M},$$

  with $\boldsymbol{P}$ the *density of electric dipole moments* and $\boldsymbol{M}$ the *density of magnetic dipole moments* (see Ch. 6).

- *Note:* The *vacuum Maxwell equations can be recovered* by substituting $\boldsymbol{D} = \varepsilon_0 \boldsymbol{E}$ and $\boldsymbol{H} = \boldsymbol{B}/\mu_0$ into the above in-medium equations, remembering that $\mu_0\varepsilon_0 = 1/c^2$.

The *in-medium Maxwell equations* in *differential form*,

$$\begin{aligned} \boldsymbol{\nabla}\cdot\boldsymbol{D} &= \rho && \text{(Gauss's law)},\\ \boldsymbol{\nabla}\times\boldsymbol{E}+\frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\ \boldsymbol{\nabla}\cdot\boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\ \boldsymbol{\nabla}\times\boldsymbol{H}-\frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)}, \end{aligned}$$

can be cast in *integral form* using

- the *divergence theorem* and
- *Stokes' theorem*,

as we now demonstrate.

Consider the *differential form* of *Maxwell's equations in medium*,

$$\begin{aligned} \boldsymbol{\nabla}\cdot\boldsymbol{D} &= \rho && \text{(Gauss's law)},\\ \boldsymbol{\nabla}\times\boldsymbol{E}+\frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\ \boldsymbol{\nabla}\cdot\boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\ \boldsymbol{\nabla}\times\boldsymbol{H}-\frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)}. \end{aligned}$$

- Let V be a *finite volume* bounded by a *closed surface* 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*

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

applied to Gauss's law $\boldsymbol{\nabla}\cdot\boldsymbol{D}=\rho$ then yields

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\,d^3x,$$

This *integral form of Gauss's law* requires that

- the *total flux* $\boldsymbol{D}$ out through the surface is
- *equal to the charge* contained in the volume.

Likewise, applying the *divergence theorem*

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

to the *third Maxwell equation*,

$$\begin{aligned}
\boldsymbol{\nabla}\cdot\boldsymbol{D} &= \rho && \text{(Gauss's law)},\\
\boldsymbol{\nabla}\times\boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\
\boldsymbol{\nabla}\cdot\boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\
\boldsymbol{\nabla}\times\boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)}.
\end{aligned}$$

yields the *third Maxwell equation in integral form*,

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0,$$

which is the *magnetic analog* of

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\,d^3x,$$

requiring *no net flux of* $\boldsymbol{B}$ through the closed surface because *magnetic charges do not 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.

Applying *Stokes' theorem*,

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

to the *second Maxwell equation*

$$\begin{aligned}
\boldsymbol{\nabla} \cdot \boldsymbol{D} &= \rho && \text{(Gauss's law)},\\
\boldsymbol{\nabla} \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\
\boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\
\boldsymbol{\nabla} \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)}.
\end{aligned}$$

gives

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}'\, da,$$

which is an *integral form of Faraday's law* of magnetic induction.

Likewise, applying *Stokes' theorem*

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

to the *4th Maxell equation*

$$\begin{aligned}
\nabla \cdot \boldsymbol{D} &= \rho && \text{(Gauss's law)},\\
\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\
\nabla \cdot \boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\
\nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)}.
\end{aligned}$$

gives

$$\oint_C \boldsymbol{H} \cdot d\boldsymbol{l} = \oint_{S'} \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \right) \cdot \boldsymbol{n}'\, da,$$

which is an *integral form of the Ampère–Maxwell law* of magnetic fields.

Summarizing:

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\, d^3x \qquad \text{(Gauss's law)}$$

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}'\,da, \qquad \text{(Faraday's law)}$$

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0, \qquad \text{(No magnetic charges)}$$

$$\oint_C \boldsymbol{H}\cdot d\boldsymbol{l} = \oint_{S'} \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}\right)\cdot\boldsymbol{n}'\,da \qquad \text{(Ampère–Maxwell law)}$$

define *Maxwell's equations, in medium, in integral form*.

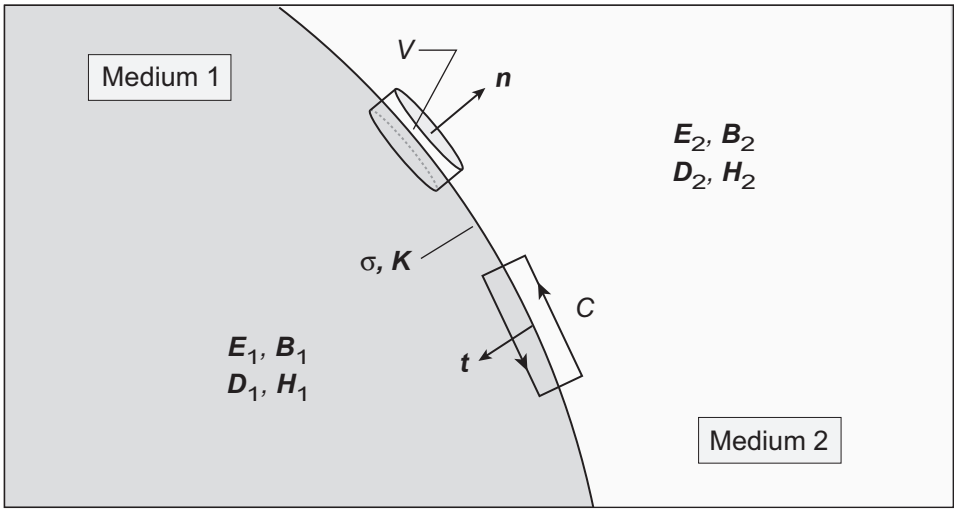
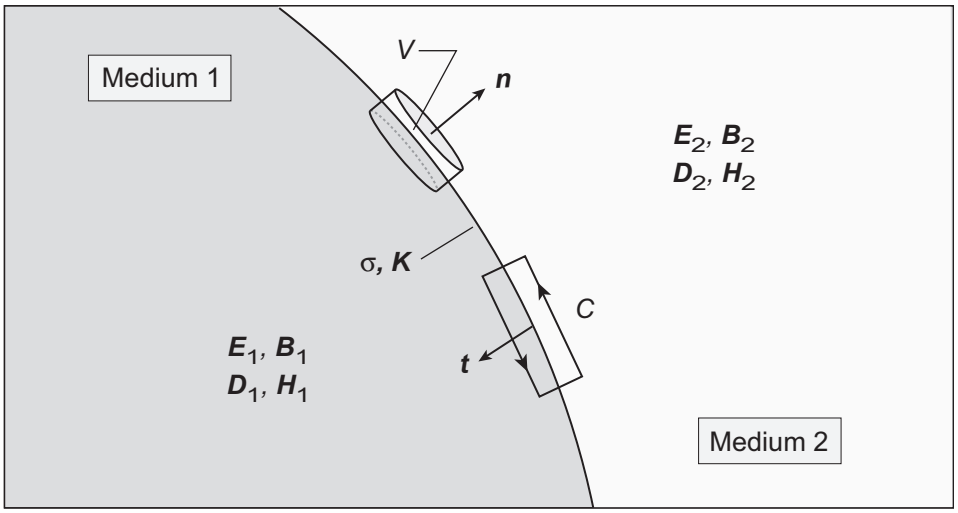


Figure 4.9: Schematic illustration of a boundary surface between different media that is assumed to carry surface charge σ and surface current density $\boldsymbol{K}$. The infinitesimal cylinder of volume V is half in one medium and half in the other, with the normal $\boldsymbol{n}$ to its surface pointing from medium 1 into medium 2. The infinitesimal rectangular contour C is partly in one medium and partly in the other, with its plane perpendicular to the surface so that its normal $\boldsymbol{t}$ (pointing out of the page) is tangent to the surface.

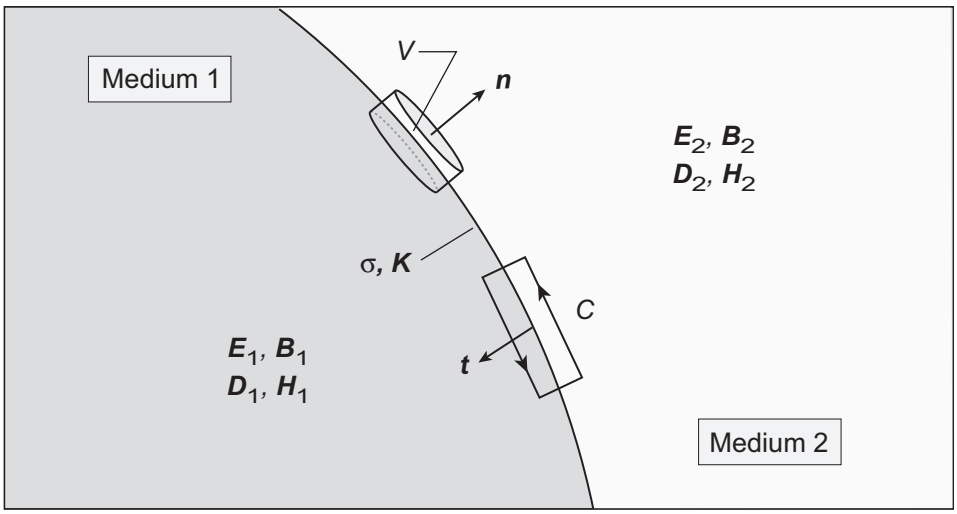
Maxwell's integral equations in medium may be used to determine the relationship of

- *normal and tangential components* of the fields on *either side of an interface* between media having *different electromagnetic properties*.
- Consider Fig. 4.9, where there is a *boundary* between medium 1 and medium 2, and
- we allow the possibility that there is a *surface charge* σ and a *current density* $\boldsymbol{K}$ at the interface.



To facilitate our analysis,

- an *infinitesimal cylindrical Gaussian pillbox* of volume V straddles the surface between the two media.
- In addition, an *infinitesimal rectangular contour* C has long sides on either side of the boundary and
- is oriented so that the *normal* $\mathbf{t}$ *to the rectangular surface* points *out of the page* and is *tangent to the interface*.



Let us first apply

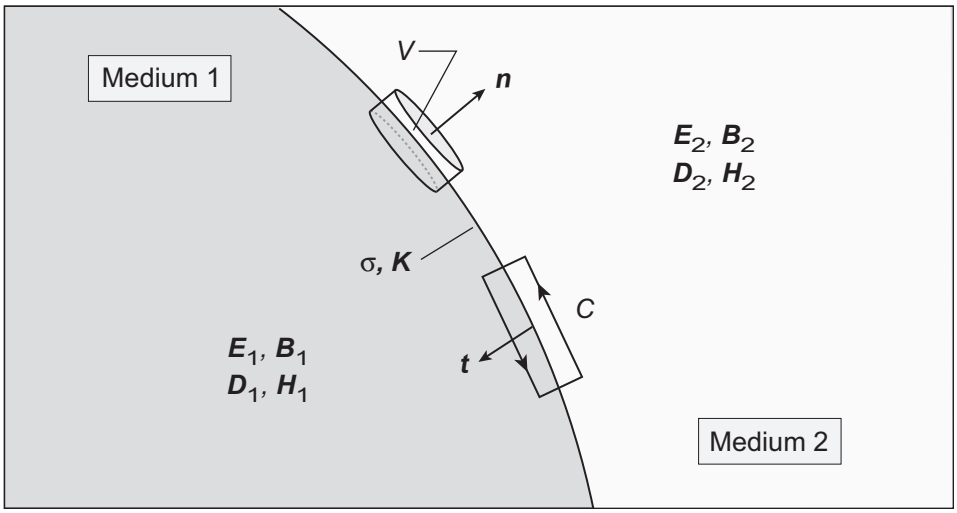
$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = \int_V \rho\, d^3x \qquad \text{(Gauss's law)}$$

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0, \qquad \text{(No magnetic charges)}$$

to the cylindrical pillbox in the above figure.

- In the limit that the pillbox is very shallow the *side of the cylinder does not contribute* to the integrals on the left side of these equations.
- If the top and bottom of the cylinder are *parallel to the interface* and have area Δa, then
- the *integral on the left side of Gauss's law* is

$$\oint_S \boldsymbol{D}\cdot\boldsymbol{n}\,da = (\boldsymbol{D}_2 - \boldsymbol{D}_1)\cdot\boldsymbol{n}\,\Delta a.$$



Applying a similar argument to the left side of

$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = 0 \qquad \text{(No magnetic charges)},$$

the integral is

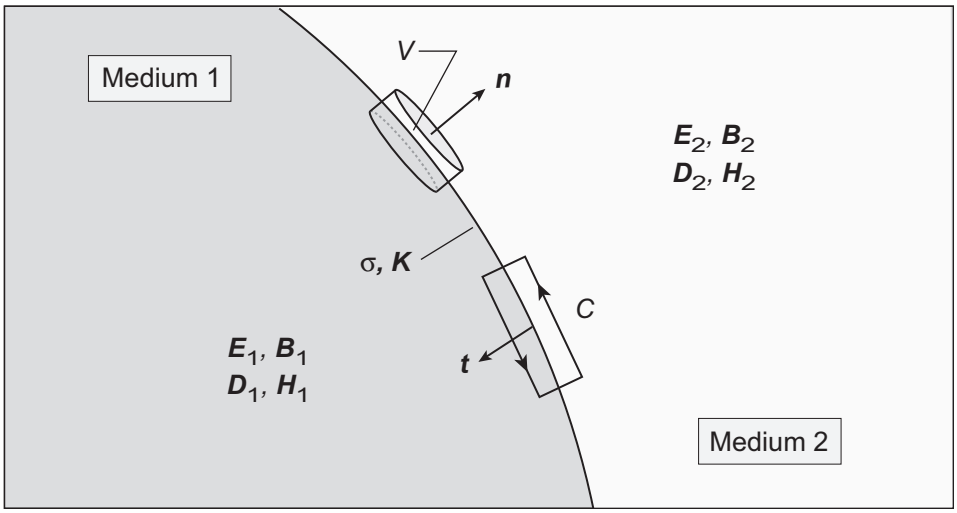
$$\oint_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = (\boldsymbol{B}_2 - \boldsymbol{B}_1)\cdot\boldsymbol{n}\,\Delta a.$$

- If the charge density ρ produces an *idealized surface charge density* σ, the integral on the right side of Gauss's law evaluates to

$$\int_V \rho\, d^3x = \sigma\Delta a,$$

and the *normal components of* $\boldsymbol{D}$ *and* $\boldsymbol{B}$ on the *two sides of the interface* are related by

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1)\cdot\boldsymbol{n} = \sigma,$$
$$(\boldsymbol{B}_2 - \boldsymbol{B}_1)\cdot\boldsymbol{n} = 0.$$

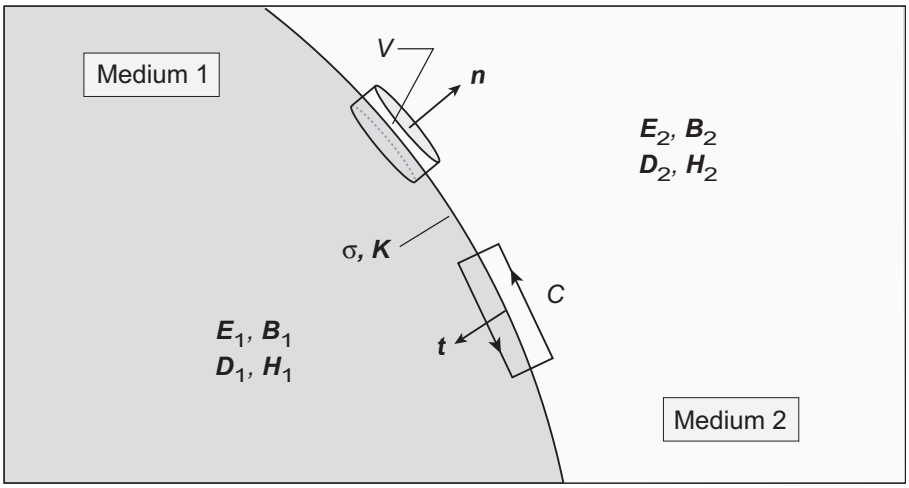


Thus, in words

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n} = \sigma,$$
$$(\boldsymbol{B}_2 - \boldsymbol{B}_1) \cdot \boldsymbol{n} = 0.$$

means that

1. the *normal component of* $\boldsymbol{B}$ *is continuous across the interface* but
2. the *discontinuity of the normal component of* $\boldsymbol{D}$ at any point is equal to the *surface charge density* σ at the point.



In a similar manner, the *infinitesimal rectangular contour* C in the figure above can be used with *Stokes' theorem*

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

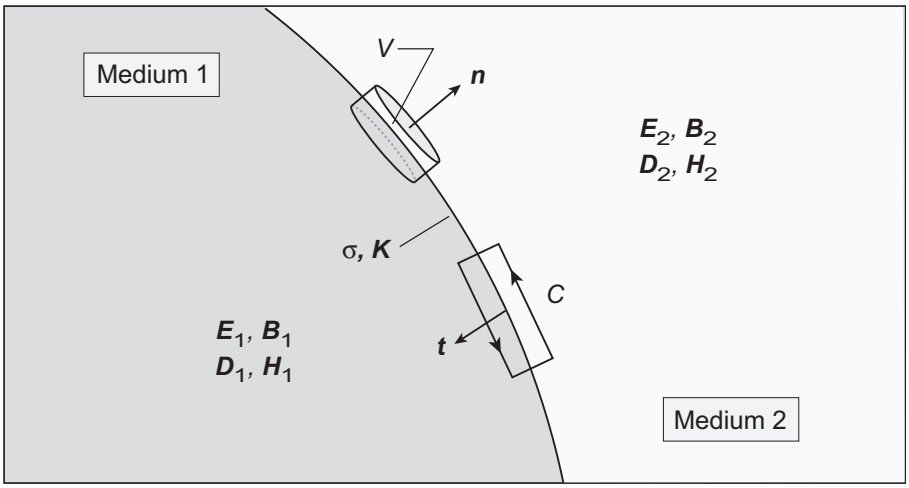
to determine the *discontinuities in the tangential components* of $\boldsymbol{E}$ and $\boldsymbol{H}$.

- 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

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}'\, da, \qquad \text{(Faraday's law)}$$

is given by

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{l} = (\boldsymbol{t} \times \boldsymbol{n}) \cdot (\boldsymbol{E}_2 - \boldsymbol{E}_1) \Delta l.$$



Likewise, the *integral on the left side* of

$$\oint_C \boldsymbol{H}\cdot d\boldsymbol{l} = \oint_{S'} \left(\boldsymbol{J}+\frac{\partial \boldsymbol{D}}{\partial t}\right)\cdot \boldsymbol{n}' da \quad \text{(Ampère–Maxwell law)}$$

is

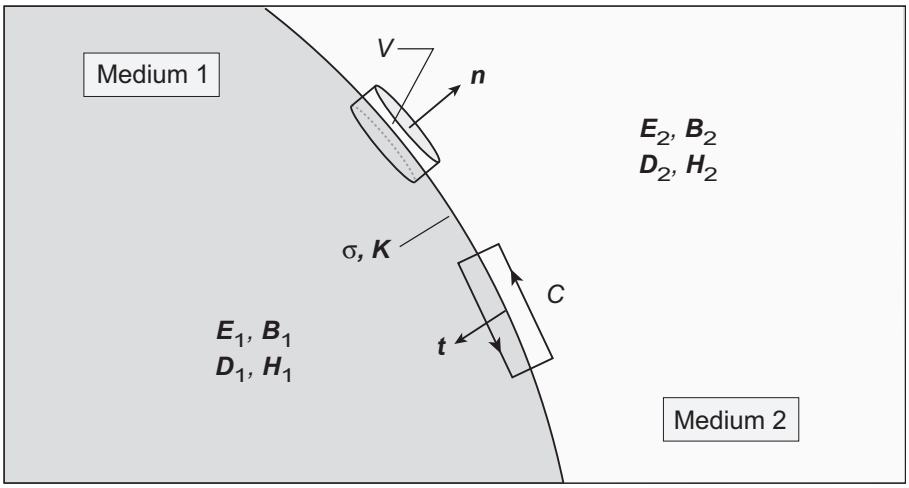
$$\oint_C \boldsymbol{H}\cdot d\boldsymbol{l} = (\boldsymbol{t}\times\boldsymbol{n})\cdot(\boldsymbol{H}_2-\boldsymbol{H}_1)\Delta l.$$

The *right side* of

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot \boldsymbol{n}' da, \qquad \text{(Faraday's law)}$$

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

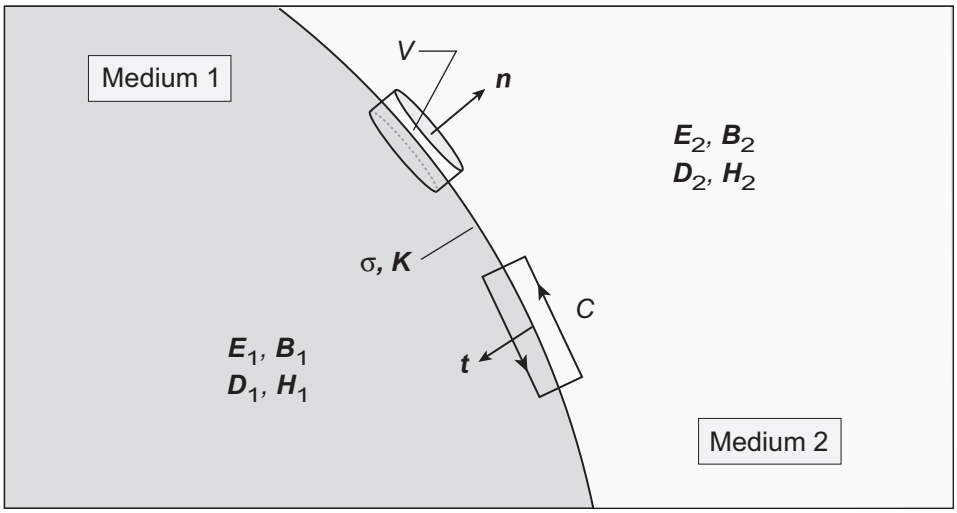
$$\oint_C \boldsymbol{H}\cdot d\boldsymbol{l} = \oint_{S'} \left(\boldsymbol{J}+\frac{\partial \boldsymbol{D}}{\partial t}\right)\cdot \boldsymbol{n}'\,da \quad \text{(Ampère–Maxwell law)}$$

is equal to

$$\oint_{S'} \left(\boldsymbol{J}+\frac{\partial \boldsymbol{D}}{\partial t}\right)\cdot \boldsymbol{t}\,da = \boldsymbol{K}\cdot\boldsymbol{t}\,\Delta l,$$

where the *second term vanishes* by the same argument as for the right side of

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot \boldsymbol{n}'\,da, \qquad \text{(Faraday's law)}$$



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

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

this implies that

$$\begin{aligned}\boldsymbol{n}\times(\boldsymbol{E}_2-\boldsymbol{E}_1)&=0,\\ \boldsymbol{n}\times(\boldsymbol{H}_2-\boldsymbol{H}_1)&=\boldsymbol{K},\end{aligned}$$

where it is understood that the surface current $\boldsymbol{K}$ has only components parallel to the interface at every point.

Thus, the *matching conditions* are

$$\boldsymbol{n} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1) = 0,$$
$$\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{K},$$

which implies that

1. the *tangential component of* $\boldsymbol{E}$ *is continuous* across an interface, but
2. the *tangential component of* $\boldsymbol{H}$ *is discontinuous* across the interface by an amount with magnitude equal to $|\boldsymbol{K}|$ and direction given by the direction of $\boldsymbol{K} \times \boldsymbol{n}$.

SUMMARY: The *discontinuity equations* are

$$
\begin{aligned}
(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n} &= \sigma, \\
(\boldsymbol{B}_2 - \boldsymbol{B}_1) \cdot \boldsymbol{n} &= 0. \\
\boldsymbol{n} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1) &= 0, \\
\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) &= \boldsymbol{K},
\end{aligned}
$$

for the fields $\boldsymbol{B}$, $\boldsymbol{E}$, $\boldsymbol{D}$, and $\boldsymbol{H}$.

- These allow *solving the Maxwell equations in different regions* having potentially different electromagnetic properties, and
- then *connecting the solutions* to obtain the fields over all of the space.

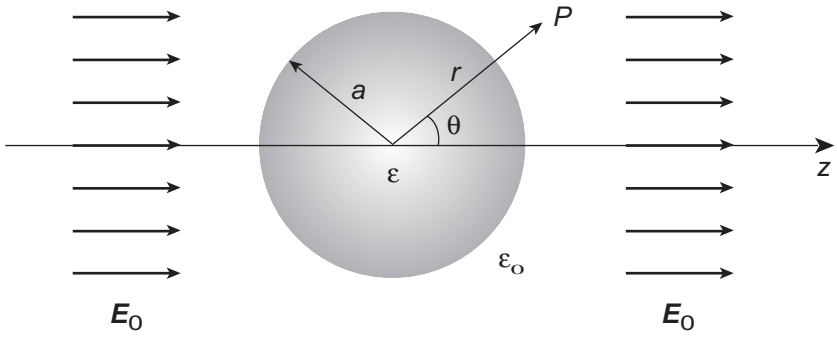


Figure 4.10: Delectric ball of radius a and dielectric constant $\kappa = \varepsilon/\varepsilon_0$ in an initially uniform external electric field $\boldsymbol{E}_0$ directed along the z axis.

4.12 Dielectric Boundary Value Problems

The methods developed in previous chapters may be *adapted to handle the presence of dielectrics*.

Example: Dielectric Ball in Uniform Electric Field

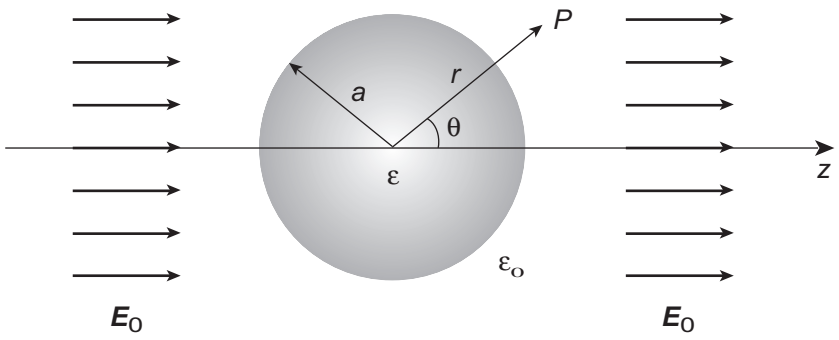
A *dielectric ball* with dielectric constant $\kappa = \varepsilon/\varepsilon_0$ is placed in an initially *uniform electric field directed along the z axis*, as illustrated in Fig. 4.10.

Let's determine the

- *potential* Φ and
- the *electric field* $\boldsymbol{E}$

inside and *outside the ball*, assuming *no free charges* inside or outside of the ball.

There are no free charges so the solution will involve *solving the Laplace equation* with *appropriate boundary conditions*.



- We assume *axial symmetry* about the z axis.
- Solving the *Laplace equation in spherical coordinates* by *separation of variables*, general solutions are of the form,

$$\Phi(r,\theta) = \sum_{l=0}^{\infty} (A_l r^l + B_l r^{-(l+1)}) P_l(\cos\theta).$$

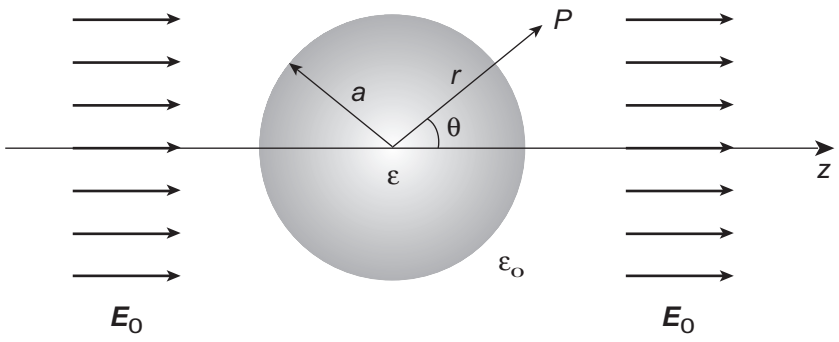
- However, there are *no charges at the origin* and
- the *requirement that* $\Phi(r,\theta)$ *be finite there* demands that for the interior solution, $B_l = 0$, so
- the *interior solution is of the form*

$$\Phi_{\text{in}}(r,\theta) = \sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta),$$

- 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,

$$\Phi(r \to \infty) = -E_0 z = -E_0 r\cos\theta = -E_0 r P_1(\cos\theta).$$

- This requires that

$$-E_0 r P_1(\cos\theta) = [B_l r^l + C_l r^{-(l+1)}]P_l(\cos\theta) \simeq B_l r^l P_l(\cos\theta)$$

since the C_l term vanishes as $r \to \infty$.

- Multiply both sides by $P_1(\cos\theta)$ and integrate,

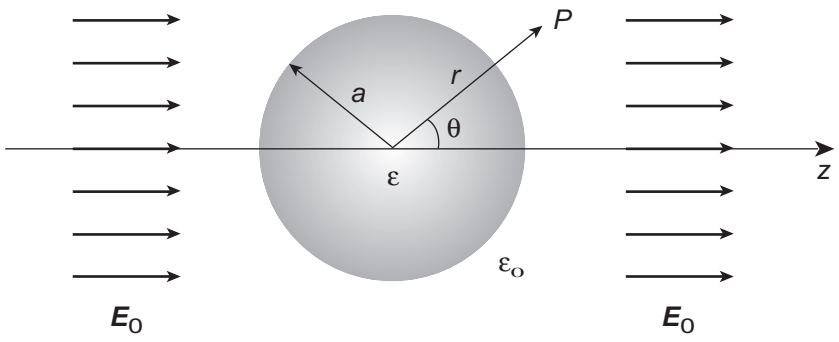
$$-E_0 r \int P_1(\cos\theta)P_1(\cos\theta)d(\cos\theta) = B_l r^l \int P_1(\cos\theta)P_l(\cos\theta)d(\cos\theta).$$

- Using the orthogonality relation

$$\int_{-1}^{+1} P_k(\cos\theta)P_l(\cos\theta)d(\cos\theta) = \frac{2}{2l+1}\delta_{kl},$$

all terms vanish except for $l = 1$ and

- we obtain $B_1 = -E_0$, with all other B_l equal to zero.



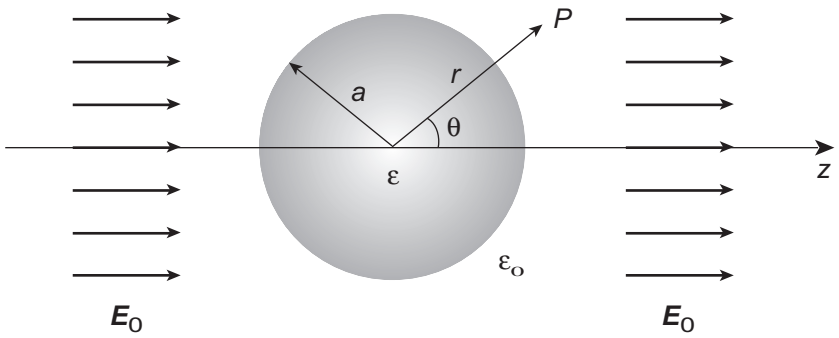
- Thus, the *exterior solution* becomes,

$$\Phi_{\text{out}}(r,\theta) = -E_0 r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)} P_l(\cos\theta),$$

- Now let's use the *boundary conditions at the edge of the sphere* to fix the other constants,
- by requiring the matching at the surface $r = a$ for *tangential and normal components* of Φ,

$$-\frac{1}{a}\frac{\partial \Phi_{\text{in}}}{\partial \theta}\bigg|_{r=a} = -\frac{1}{a}\frac{\partial \Phi_{\text{out}}}{\partial \theta}\bigg|_{r=a} \qquad \text{(tangential)},$$

$$-\varepsilon\frac{\partial \Phi_{\text{in}}}{\partial r}\bigg|_{r=a} = -\varepsilon_0\frac{\partial \Phi_{\text{out}}}{\partial r}\bigg|_{r=a} \qquad \text{(normal)}.$$



- Now *substitute the expansions*

$$\Phi_{\text{in}}(r,\theta) = \sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta),$$

$$\Phi_{\text{out}}(r,\theta) = -E_0 r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)} P_l(\cos\theta),$$

into these matching equations and solve to determine the constants.

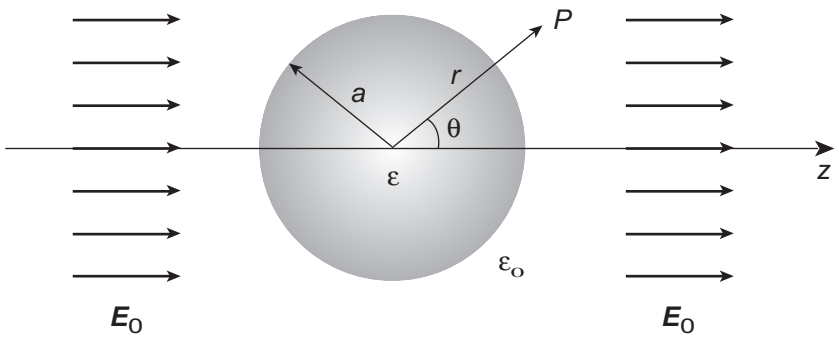
- First consider the *tangential matching*. Substituting the above equations into

$$-\frac{1}{a}\frac{\partial \Phi_{\text{in}}}{\partial \theta}\bigg|_{r=a} = -\frac{1}{a}\frac{\partial \Phi_{\text{out}}}{\partial \theta}\bigg|_{r=a}$$

leads to

$$-\frac{1}{a}\frac{\partial}{\partial\theta}\sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta)\bigg|_{r=a} =$$

$$-\frac{1}{a}\frac{\partial}{\partial\theta}\left[-E_0 r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)}\right]_{r=a} P_l(\cos\theta),$$



- which simplifies to

$$\sum_{l=0}^{\infty} a^l A_l \frac{\partial}{\partial\theta} P_l(\cos\theta) =$$
$$-aE_0 \frac{\partial}{\partial\theta} P_1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \frac{\partial}{\partial\theta} P_l(\cos\theta).$$

- Let's convert the derivatives of Legendre polynomials $P_n(x)$ to associated Legendre polynomials $P_n^m(x)$ using the general relationship,

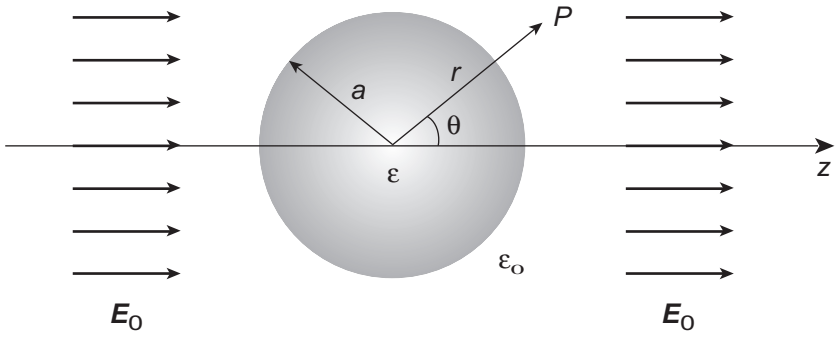
$$P_n^m(x) = (1-x^2)^{m/2} \frac{d^m}{dx^m} P_n(x),$$

- which upon specializing to $m = 1$ yields

$$\frac{dP_n(x)}{dx} = (1-x^2)^{-1/2} P_n^1(x),$$

or letting $x = \cos\theta$,

$$\frac{dP_n(\cos\theta)}{d\theta} = -P_n^1(\cos\theta),$$



- Then

$$\sum_{l=0}^{\infty} a^l A_l \frac{\partial}{\partial\theta} P_l(\cos\theta) = -aE_0 \frac{\partial}{\partial\theta} P_1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \frac{\partial}{\partial\theta} P_l(\cos\theta).$$

  becomes

$$-\sum_{l=0}^{\infty} a^l A_l P_l^1(\cos\theta) = -aE_0 P_1^1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)} C_l P_l^1(\cos\theta)$$

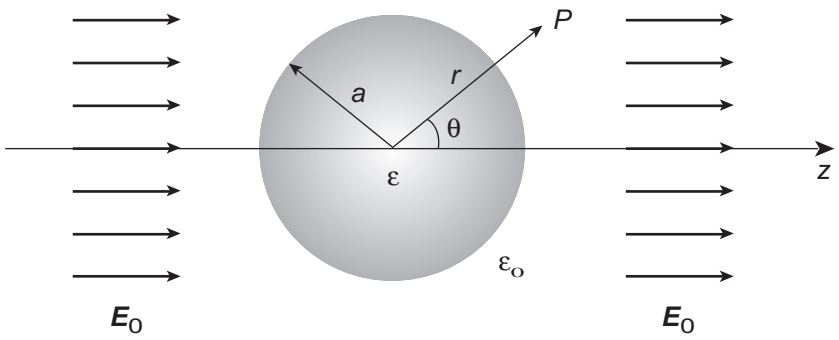
- We will now exploit the *orthogonality properties of the associated Legendre polynomials*, which obey

$$\int_{-1}^{+1} P_p^m(x) P_q^m(x) dx = K_{qm}\delta_{pq} \quad K_{qm} \equiv \frac{2}{2q+1}\frac{(q+m)!}{(q-m)!}$$

  or in *spherical coordinates*

$$\int_0^{\pi} P_p^m(\cos\theta) P_q^m(\cos\theta) \sin\theta d\theta = K_{qm}\,\delta_{pq}$$

There are two solutions, a *special solution* for $l = 1$ and a *general solution* for all $l \neq 1$.



To obtain the *special solution* let $x = \cos\theta$, multiply both sides by $P_1^1(x)$, and integrate.

$$-\sum_{l=0}^{\infty} a^l A_l \int P_1^1(x) P_l^1(x)\,dx =$$

$$-aE_0 \int P_1^1(x) P_1^1(x)\,dx + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \int P_l^1(x) P_1^1(x)\,dx$$

Utilizing

$$\int_{-1}^{+1} P_p^m(x) P_q^m(x)\,dx = K_{qm}\delta_{pq} \qquad K_{qm} \equiv \frac{2}{2q+1}\frac{(q+m)!}{(q-m)!}$$

this gives

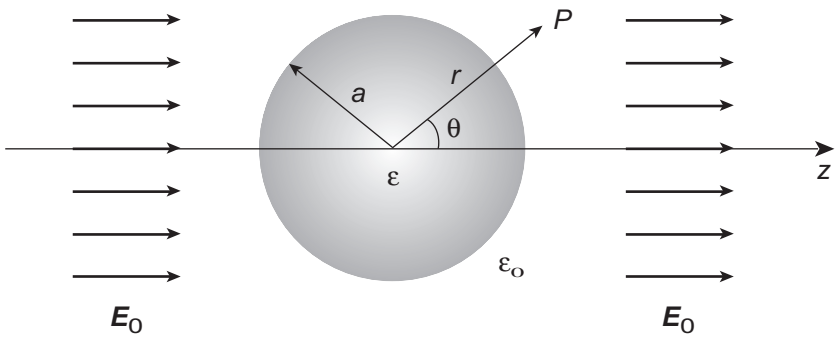
$$-\sum_{l=0}^{\infty} a^l A_l \delta_{1l} = -aE_0 + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \delta_{1l}$$

and finally

$$aA_1 = -aE_0 + a^{-2}C_1$$

or equivalently,

$$A_1 = -E_0 + \frac{C_1}{a^3}.$$



Now let's find the *general solution* for $l \neq 1$.

- Let $x = \cos\theta$ and multiply both sides of

$$-\sum_{l=0}^{\infty} a^l A_l P_l^1(\cos\theta)$$

$$= -aE_0 P_1^1(\cos\theta) + \sum_{l=0}^{\infty} a^{-(l+1)} C_l P_l^1(\cos\theta)$$

by $P_k^1(x)$ and integrate,

$$-\sum_{l=0}^{\infty} a^l A_l \int P_k^1(x) P_l^1(x)dx =$$

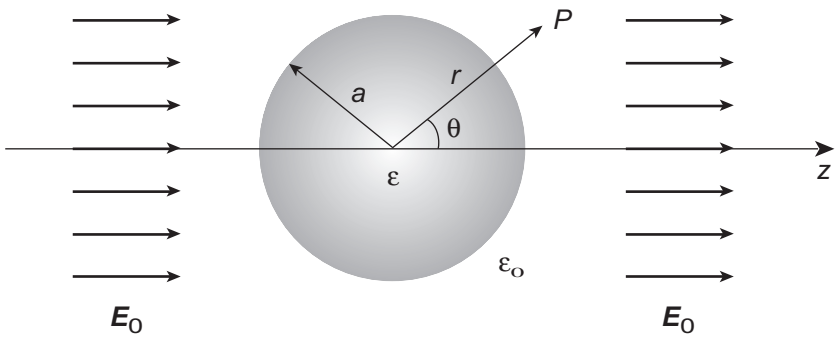
$$-aE_0 \int P_k^1(x) P_1^1(x)dx + \sum_{l=0}^{\infty} a^{-(l+1)} C_l \int P_k^1(x) P_l^1(x)dx$$

- Utilizing the orthogonality relation

$$\int_{-1}^{+1} P_p^m(x) P_q^m(x)dx = K_{qm}\delta_{pq} \qquad K_{qm} \equiv \frac{2}{2q+1}\frac{(q+m)!}{(q-m)!}$$

this gives

$$a^l A_l = a^{-(l+1)} C_l \quad \rightarrow \quad A_l = \frac{C_l}{a^{2l+1}}$$



We now have *two relations among coefficients*

$$A_1 = -E_0 + \frac{C_1}{a^3} \qquad A_l = \frac{C_l}{a^{2l+1}}.$$

obtained from the *tangential matching condition*

$$-\frac{1}{a}\frac{\partial \Phi_{\text{in}}}{\partial \theta}\bigg|_{r=a} = -\frac{1}{a}\frac{\partial \Phi_{\text{out}}}{\partial \theta}\bigg|_{r=a} \qquad \text{(tangential)}.$$

Now let's use the *normal matching condition*

$$-\varepsilon\frac{\partial \Phi_{\text{in}}}{\partial r}\bigg|_{r=a} = -\varepsilon_0\frac{\partial \Phi_{\text{out}}}{\partial r}\bigg|_{r=a} \qquad \text{(normal)}.$$

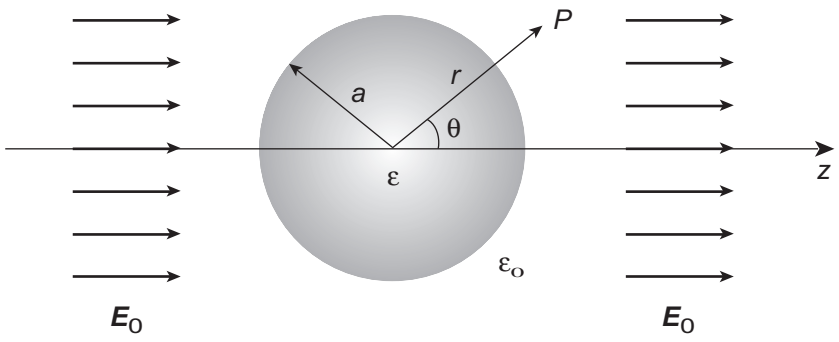
to obtain *additional constraints*. Upon substituting

$$\Phi_{\text{in}}(r,\theta) = \sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta),$$

$$\Phi_{\text{out}}(r,\theta) = -E_0 r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)} P_l(\cos\theta),$$

we obtain

$$\varepsilon \sum_{l=0}^{\infty} l a^{l-1} A_l P_l(x) = \varepsilon_0 B_1 P_1(x) + \sum_{l=0}^{\infty} -(l+1) C_l a^{-(l+2)} P_l(x).$$



As before, there are *two solutions*,

- a *special solution* when $l = 1$, and
- a *general solution* when $l \neq 1$.

Lets find the *special solution* by multiplying

$$\varepsilon \sum_{l=0}^{\infty} l a^{l-1} A_l P_l(x) = \varepsilon_0 B_1 P_1(x) + \sum_{l=0}^{\infty} -(l+1) C_l a^{-(l+2)} P_l(x).$$

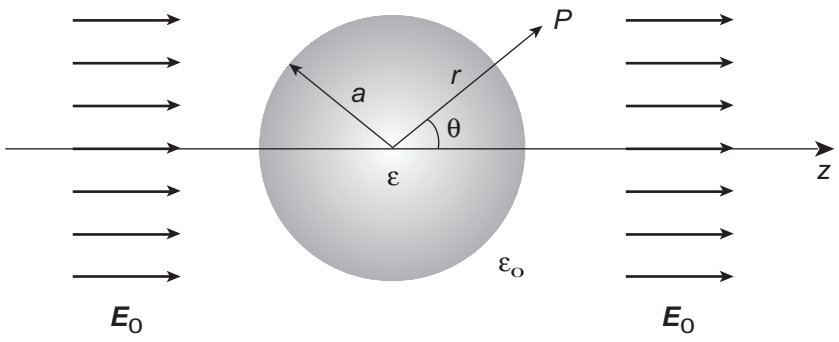
by $P_1(x)$ and integrating.

- Upon exploiting the *orthogonality properties*

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

the result is another relationship among coefficients:

$$\frac{\varepsilon}{\varepsilon_0} A_1 = -E_0 - 2\frac{C_1}{a^3} \qquad (l = 1).$$



The *general solution* may be obtained by multiplying both sides of

$$\varepsilon \sum_{l=0}^{\infty} l a^{l-1} A_l P_l(x) = \varepsilon_0 B_1 P_1(x) + \sum_{l=0}^{\infty} -(l+1) C_l a^{-(l+2)} P_l(x).$$

by $P_k(x)$, integrating, and exploiting the orthogonality constraints

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

The result is

$$\frac{\varepsilon}{\varepsilon_0} l A_l = -(l+1) \frac{C_l}{a^{2l+1}} \qquad (l \neq 1).$$

- We now have *four equations*,

$$A_1 = -E_0 + \frac{C_1}{a^3}$$
$$A_l = \frac{C_l}{a^{2l+1}}$$
$$\frac{\varepsilon}{\varepsilon_0} A_1 = -E_0 - 2\frac{C_1}{a^3}$$
$$\frac{\varepsilon}{\varepsilon_0} l A_l = -(l+1)\frac{C_l}{a^{2l+1}}$$

  to solve simultaneously for the unknown coefficients.

- Eqs. 2 and 4 are *satisfied only if* $A_l = C_l = 0$ for all $l \neq 1$,

- Solving Eqs. 1 and 3 simultaneously for $l = 1$ gives

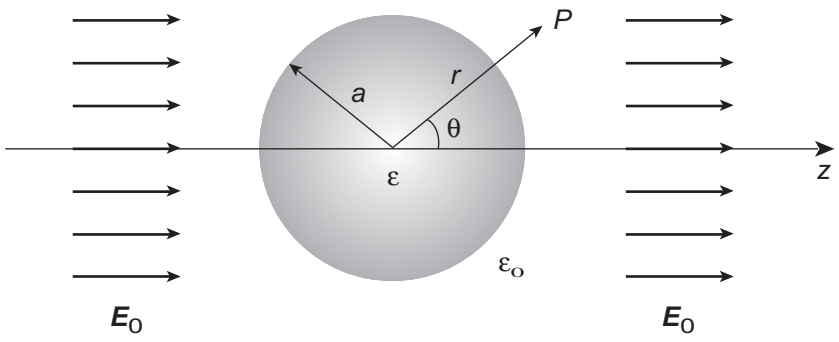
$$A_1 = -\left(\frac{3}{\varepsilon/\varepsilon_0 + 2}\right) E_0 \qquad C_1 = \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) a^3 E_0.$$

- Inserting these results into

$$\Phi_{\text{in}}(r,\theta) = \sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta),$$
$$\Phi_{\text{out}}(r,\theta) = -E_0 r^l P_1(\cos\theta) + \sum_{l=0}^{\infty} C_l r^{-(l+1)} P_l(\cos\theta),$$

  gives for the *interior and exterior potentials*

$$\Phi_{\text{in}} = -\left(\frac{3}{2+\varepsilon/\varepsilon_0}\right) E_0 r\cos\theta = -\left(\frac{3}{2+\varepsilon/\varepsilon_0}\right) E_0 z,$$
$$\Phi_{\text{out}} = -E_0 z + \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) E_0 a^3 \frac{\cos\theta}{r^2}.$$



The *electric fields* then follow from $\boldsymbol{E} = -\nabla\Phi$. *For the interior*,

$$E_{\text{in}} = -\frac{\partial}{\partial z}\left[\left(\frac{-3}{\varepsilon/\varepsilon_0+2}\right)E_0 z\right] = \left(\frac{3}{\varepsilon/\varepsilon_0+2}\right)E_0,$$

from which we conclude that

1. the *interior field* E_{in} *is constant*, proportional to the constant exterior field,
2. if $\varepsilon = \varepsilon_0$, then the *interior field is equal to the exterior field*, $E_{\text{in}} = E_0$, and
3. if $\varepsilon > \varepsilon_0$, then the *interior field is reduced* relative to the exterior field, $E_{\text{in}} < E_0$.

In the equation for the exterior potential,

$$\Phi_{\text{out}} = -E_0 z + \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) E_0 a^3 \frac{\cos\theta}{r^2}.$$

- it is clear that the *first term is due to the unperturbed exterior field* $\boldsymbol{E}_0$ and
- the *second term is a correction* associated with the *polarized dielectric sphere*.
- If the *induced electric dipole moment* $\boldsymbol{p}$ of the polarized sphere is taken to have magnitude

  $$p = 4\pi\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) a^3 E_0,$$

  then the *exterior potential* can be written

  $$\Phi_{\text{out}} = -E_0 z + \frac{p}{4\pi\varepsilon_0}\frac{\cos\theta}{r^2},$$

- making clear that the external potential is
  - the *unperturbed external potential* (1st term)
  - *modified by a potential from an induced electric dipole* associated with the polarized dielectric sphere (2nd term).

- From
$$\boldsymbol{P} = \varepsilon_0 \chi_e \boldsymbol{E} \qquad \kappa \equiv \frac{\varepsilon}{\varepsilon_0} = 1 + \chi_e$$
the *polarization* $\boldsymbol{P}$ is given by
$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E},$$
and from
$$\boldsymbol{P} \equiv \boldsymbol{p}N = \frac{\text{dipole moments}}{\text{unit volume}},$$
the polarization may also be defined as the *density of dipole moments*.

- Then if the *induced electric dipole moment* $\boldsymbol{p}$ *of the polarized sphere* is taken to have magnitude
$$p = 4\pi\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) a^3 E_0,$$
the *polarization of the dielectric sphere* is given by
$$\boldsymbol{P} = (\varepsilon - \varepsilon_0)\boldsymbol{E} = \left(\frac{\text{dipole moments}}{\text{unit volume}}\right)$$
$$= \frac{p}{\frac{4}{3}\pi a^3} = 3\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2}\right) \boldsymbol{E}_0.$$

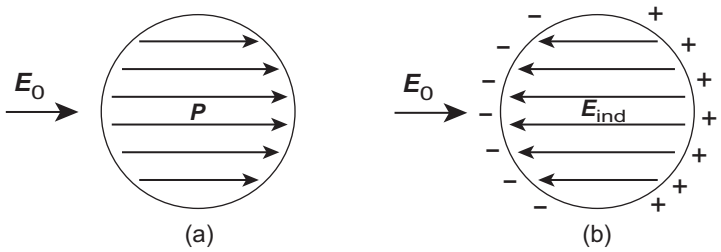
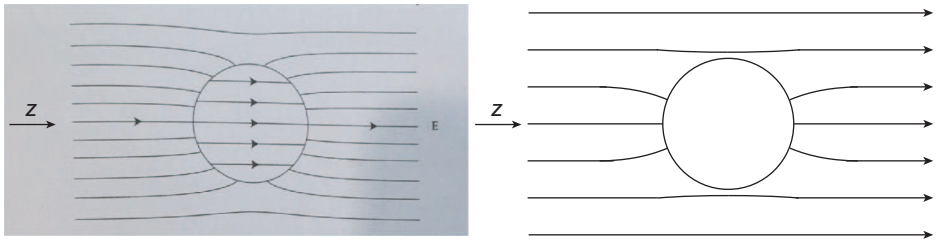


Figure 4.11: (a) Polarization of a dielectric ball by an external electric field $\boldsymbol{E}_0$. (b) The polarization of charge induces a field $\boldsymbol{E}_{\text{ind}}$ that opposes the applied field $\boldsymbol{E}_0$.

The results of this example are displayed in Fig. 4.11.

- As indicated in Fig. 4.11(a), the *external field* $\boldsymbol{E}_0$ aligned in the z direction *polarizes the dielectric sphere*,
- with the *polarization vector* $\boldsymbol{P}$ oriented in the *direction of the external field*.
- The polarization is the *density of induced dipole moments*.
- This corresponds to the *polarization of charge* illustrated in Fig. 4.11(b) with positive charge accumulating on the right side of the sphere and negative charge on the left side.
- The *polarization of charge* indicated in Fig. 4.11(b) *induces an electric field* $\boldsymbol{E}_{ind}$ *that opposes the applied field* $\boldsymbol{E}_0$ but *does not completely cancel it* as would happen for a conducting filled sphere.
- The electric field $\boldsymbol{E}_{\text{in}}$ inside the sphere is is *constant*.



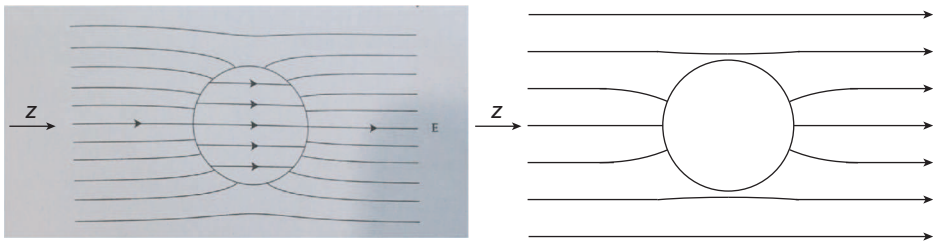
(a) Dielectric filled sphere (a) Metal filled sphere

Figure 4.12: (a) Electric field lines for dielectric filled sphere in an external electric field $\boldsymbol{E}_0$ directed along the $\boldsymbol{z}$ axis. (b) Electric field lines for a conducting ball in an external electric field $\boldsymbol{E}_0$ directed along the $\boldsymbol{z}$ axis. Original calculation in Fig. 3.11.

- If $\varepsilon > \varepsilon_0$ the *electric field inside acts in the opposite direction* as the applied field and
- is *reduced in strength by a factor* $3/(\varepsilon/\varepsilon_0 + 2)$ relative to the applied field.

This is indicated schematically in Fig. 4.12(a); notice that

1. The *field inside* $\boldsymbol{E}_{in}$ is in the *same direction as the applied field* $\boldsymbol{E}_0$, but
2. is *reduced in strength by the induced field* $\boldsymbol{E}_{\text{ind}}$ acting against the applied field.
3. This *reduction in strength for the interior field* is indicated by the *decreased density of field lines inside the sphere* relative to the outside.
4. There is a *discontinuity of the electric field lines* at the boundary of the sphere
5. because of the *surface charge* that has accumulated there as a consequence of *polarization*.



(a) Dielectric filled sphere (a) Metal filled sphere

As indicated by

$$p = 4\pi\varepsilon_0 \left(\frac{\varepsilon/\varepsilon_0 - 1}{\varepsilon/\varepsilon_0 + 2} \right) a^3 E_0,$$

$$\Phi_{\text{out}} = -E_0 z + \underbrace{\frac{p}{4\pi\varepsilon_0} \frac{\cos\theta}{r^2}}_{\text{polarization}},$$

- the external potential consists of *the original potential contributed by the external field* plus
- a correction term that may be viewed as the *potential generated by a set of electric dipoles* centered on the sphere because of the *charge polarization*.
- *Far from the sphere* the external potential is that of the original applied field $\boldsymbol{E}_0$, but
- *near the sphere* the field lines in Fig. (a) above are distorted by the increasing importance of the second term the equation above for Φ_{out},
  - which *scales as* r^{-2} and therefore
  - *grows rapidly* as the sphere is approached.

Chapter 5

Magnetostatics

- 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*, and
- we desire to *calculate the electric field* at arbitrary locations in space produced by those charges.
- This can be done using the *principle of superposition:*
  - *Calculate the contribution of each source charge* q_i or infinitesimal piece of a continuous charge $d\rho(\boldsymbol{x})$ to the electric field at some point, and
  - *sum them* to get the total electric field at that point.
- We have developed a number of *sophisticated ways to do this beyond brute force summation or integration*, but that is the essential idea.

In this chapter we wish to expand upon this idea by *allowing the source charges to move*.

5.1 Introduction

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.

As a consequence, it took *centuries after electrostatic and magnetic phenomena were first identified* 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*, 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

$$\nabla \cdot \boldsymbol{B}(\boldsymbol{x}) = 0,$$
$$\nabla \times \boldsymbol{B}(\boldsymbol{x}) = \mu_0 \boldsymbol{j}(\boldsymbol{x}).$$

  These equations *specify the divergence and the curl* of $\boldsymbol{B}(\boldsymbol{x})$. This should be sufficient to specify $\boldsymbol{B}(\boldsymbol{x})$ uniquely by the *Helmholtz theorem*.

- *Applying the divergence operation* to both sides of

$$\nabla \times \boldsymbol{B}(\boldsymbol{x}) = \mu_0 \boldsymbol{j}(\boldsymbol{x})$$

  using the *vector identity* $\nabla \cdot (\nabla \times \boldsymbol{A}) = 0$ gives

$$\nabla \cdot (\nabla \times \boldsymbol{B}(\boldsymbol{x})) = 0 = \mu_0 \nabla \cdot \boldsymbol{j}(\boldsymbol{x}).$$

- Thus, *magnetostatic current densities* satisfy $\nabla \cdot \boldsymbol{j}(\boldsymbol{x}) = 0$.
- In *electrostatics* stationary charges produce *electric fields constant in time*.
- In *magnetostatics* stationary (unchanging) currents produce *magnetic fields that are constant in time*.

The *more complex nature of magnetostatics* relative to electrostatics arises in part because

- magnetostatics involves a *vector current density* and
- the force law (below) involves a *cross product* of vectors,
- 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*.

  – *Quantum orbital angular momentum* can be viewed semiclassically as charge in motion on a classical orbit, but

  – *quantum spin angular momentum* has no corresponding classical analog.

- It follows that *magnetizable matter* exhibits much greater variety than *polarizable matter*.

For example, permanent magnetism (*ferromagnetism*) is much more common than than the dielectric counterpart of *ferroelectricity*.

From the *static Maxwell equations* (all time derivatives set to zero),

- the *basic equations governing electrostatics* are

$$\nabla \times \boldsymbol{E}(\boldsymbol{x}) = 0$$
$$\nabla \cdot \boldsymbol{E}(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x})}{\varepsilon_0}.$$

- Comparing with the *basic equations governing magnetostatics*,

$$\nabla \cdot \boldsymbol{B}(\boldsymbol{x}) = 0,$$
$$\nabla \times \boldsymbol{B}(\boldsymbol{x}) = \mu_0 \boldsymbol{j}(\boldsymbol{x}).$$

  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 is only a *choice of observer reference frame*:
  - what *looks like an electric field* from one inertial frame
  - *looks like a magnetic field* from a different inertial frame,

  and vice versa
- We shall take that up in *later chapters* when we discuss
  - the *special theory of relativity*,
  - *Lorentz transformations*, and
  - formulation of the *Maxwell equations in a manifestly Lorentz-covariant manner*.
- For now we note that in our low-energy world
- there is some *pragmatic utility* in using our present formalism that *distinguishes clearly between electric and magnetic phenomena*
- through the use of equations that are (secretly) *actually Lorentz covariant*, but are *not manifestly so*.

5.2 Magnetic Forces

A magnetic field $\boldsymbol{B}$ generated by source charges in motion and the electric field $\boldsymbol{E}$ associated with those charges

- lead to a force $\boldsymbol{F}$ that is found to act on a test charge q having relative velocity v according to the *Lorentz force law*,

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}).$$

- As we will find, the very different phenomenology of magnetic effects relative to electrostatic effects
- is partially due to the *nature of this force law*, which mixes the effect of the electric field and magnetic field in a non-trivial way.

Characteristic *cyclotron motion* of a charged particle in a magnetic field is circular with the centripetal acceleration provided by the magnetic force.

How much *work* can be *done by the Lorentz force*?

- Previously we found 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 with $\boldsymbol{B} = 0$ the same result is obtained (reassuring!).
- On the other hand, let's set $\boldsymbol{E} = 0$ 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 *work done by the magnetic field* is

$$dW = \boldsymbol{F} \cdot d\boldsymbol{L} = Q(v \times \boldsymbol{B}) \cdot \boldsymbol{v}\, dt = 0.$$

- The *magnetic field does no work.*
- This obviously is because *the cross product* $\boldsymbol{v} \times \boldsymbol{B}$ is *perpendicular to the velocity* $\boldsymbol{v}$,
- so the *scalar product* $(v \times \boldsymbol{B}) \cdot \boldsymbol{v}$ *vanishes*.

As seen in the *cyclotron motion* discussed below,

- *magnetic forces* can *alter the direction of charged-particle motion*, but
- they *cannot change the speed* (magnitude of the velocity).

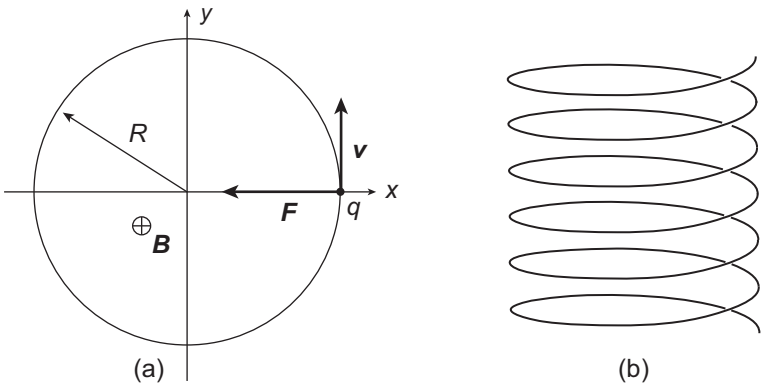


Figure 5.1: (a) Cyclotron motion caused by magnetic field $\boldsymbol{B}$ pointing into the page along the z axis (not shown) acting on a charge q with a velocity $\boldsymbol{v}$. (b) An electric field $\boldsymbol{E}$ perpendicular to the magnetic field converts circular cyclotron motion into spiral motion because of the Lorentz force law.

Cyclotron Motion

As shown in Fig. 5.1(a), the field $\boldsymbol{B}$, *oriented into the page*,

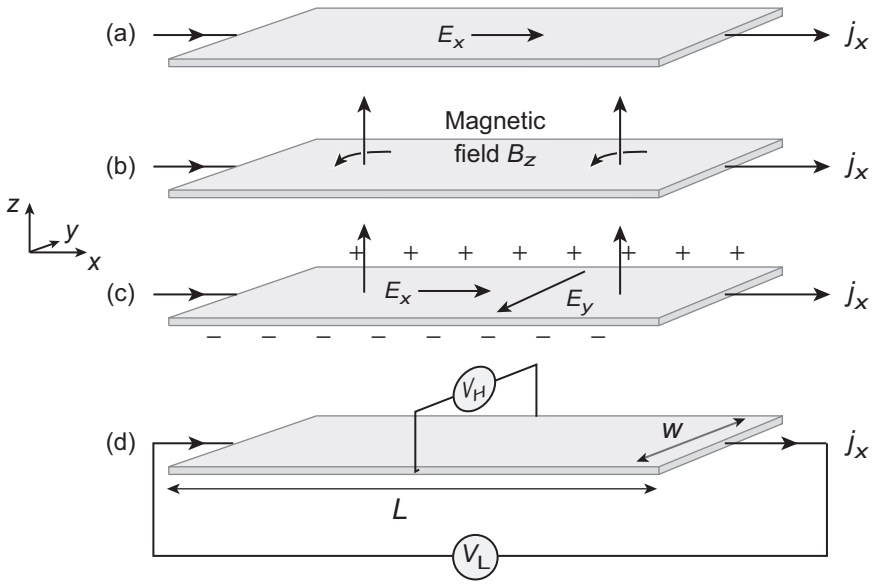
- *produces through the* $q\boldsymbol{v} \times \boldsymbol{B}$ *component of the Lorentz force* a centripetal force
- that is *directed toward the center of the circle* for a particle of charge q and velocity $\boldsymbol{v}$,
- causing the particle to be *deflected in a circular path*.

The motion becomes more complex if an $\boldsymbol{E}$ field is present also.

- For an $\boldsymbol{E}$ perpendicular to $\boldsymbol{B}$,
- the circular cyclotron motion due to the magnetic field
- is converted into a the *spiral motion* depicted in Fig. 5.1(b) by the $q\boldsymbol{E}$ *component of the Lorentz force*.

The Hall Effect

A scientifically important consequence of the Lorentz force is exemplified by the *classical Hall effect*.



(a) For a 2D sample an *electric field* E_x causes a *current density* j_x *in the* x *direction*.

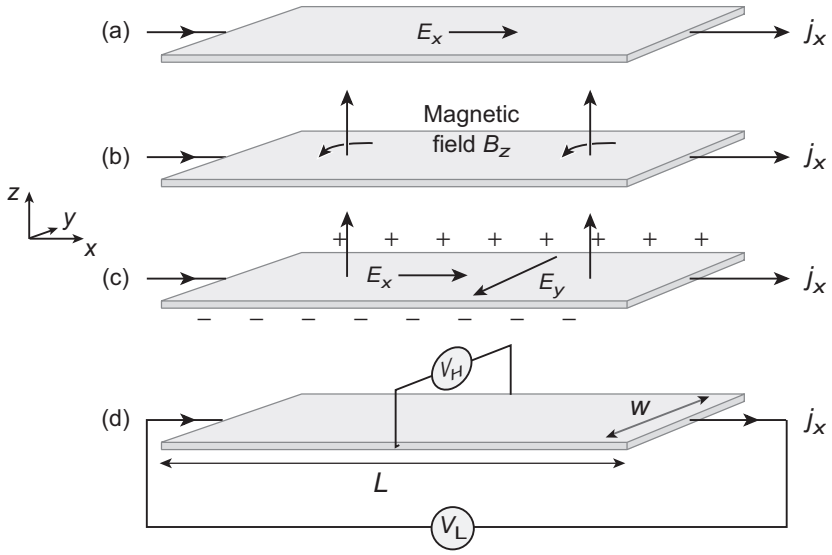
(b) A *uniform magnetic field* $\boldsymbol{B}$ placed on the sample *in the* $+z$ *direction* deflects electrons *in the* $-y$ *direction*.

(c) Negative charge accumulates on one edge and a positive charge excess on the other edge,

- producing a *transverse electric field* E_y (*Hall field*)
- that just *cancels the magnetic field 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

- which is inferred from $R_H = V_H/wj_x$,
- depends on the *sign and density of charge carriers*.
- Thus the classical Hall effect is important as a *diagnostic for charge carriers* in materials samples.

At very high magnetic fields, *quantum effects become significant*.

- These *quantum Hall effects* are of even greater importance, since they
- first *revealed topological effects* that were the harbingers of the modern *topological matter revolution* in condensed matter and materials science.

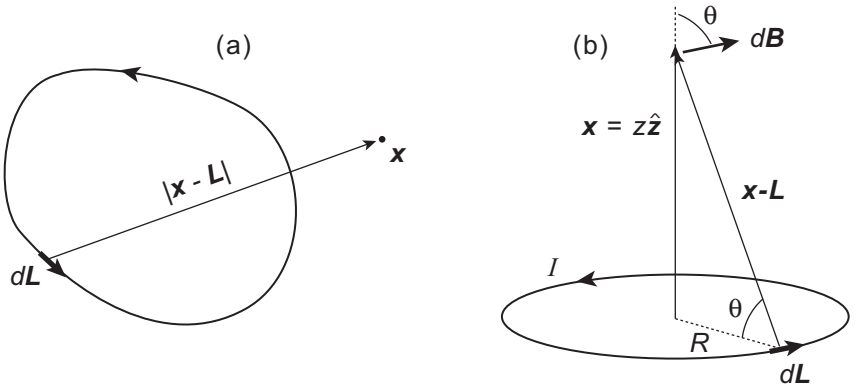


Figure 5.2: (a) Current loop carrying a steady current I. (b) Geometry for Biot–Savart calculation of the magnetic field on the symmetry axis of a circular loop.

5.3 The Law of Biot and Savart

The magnetic field of a steady current is given by the *Biot–Savart law*,

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x',$$

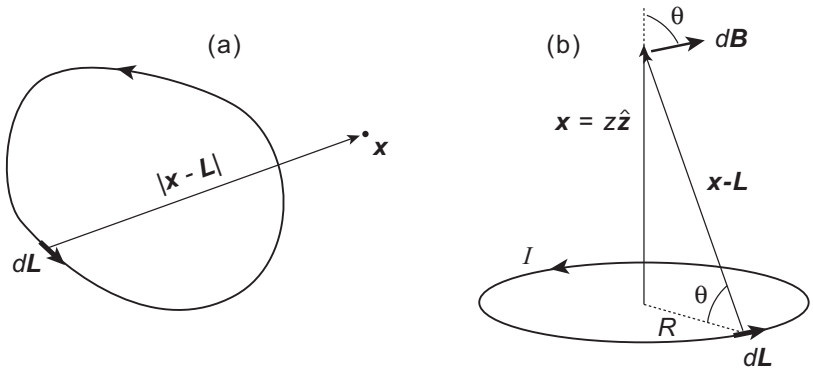
where the *permeability of free space* is

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

- When a *steady current* is flowing in a 1D wire the *magnitude of the current* I *must be constant*.
- If $\boldsymbol{L}$ is a vector that points to a line element $d\boldsymbol{L}$ of the wire as in Fig. 5.2(a),
- substitution of $Id\boldsymbol{L}$ for $\boldsymbol{j}d^3x$ in the equation above yields the *Biot–Savart law* in the form

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0 I}{4\pi}\int \frac{d\boldsymbol{L}\times(\boldsymbol{x}-\boldsymbol{L})}{|\boldsymbol{x}-\boldsymbol{L}|^3}.$$

Example:
Let's use the *Biot–Savart law* to calculate the magnetic field produced by the circular current loop in (b) of the following figure along the z axis.



- The components of $d\boldsymbol{B}$ perpendicular to the z symmetry axis *cancel when the entire loop is traversed* but
- the z components *add with the same magnitude* for each $d\boldsymbol{L}$, giving

$$\boldsymbol{B}(z) = \hat{\boldsymbol{z}}\frac{\mu_0}{4\pi}\frac{\cos\theta}{R^2+z^2}\oint dL = \hat{\boldsymbol{z}}\frac{\mu_0 I}{2}\frac{R^2}{(R^2+z^2)^{3/2}}.$$

- This *non-zero value of* $\boldsymbol{B}$ on the symmetry axis *contrasts with the value of zero for the electric field* $\boldsymbol{E}$ found *for a uniformly charged ring*.
- The difference follows from the *cross product in the Biot–Savart formula*. This causes contributions to $\boldsymbol{B}(z=0)$ from opposite sides of the ring to
  - *add for the magnetic field*, but to
  - *subtract and cancel for the electric field*.

This example is one illustration of the *basic differences* between

- the way that *stationary charges produce electric fields* and
- the way that *charges in motion produce magnetic fields*.
- The *Biot–Savart law* may be viewed as the *starting point for magnetostatics*,
- just as *Coulomb's law* may be viewed as the *starting point for electrostatics*.
- Both exhibit a *one over the square of the distance dependence*, but
- they otherwise *differ substantially* because of the *vector nature of the magnetic law*.

Both *Coulomb's law and the Biot–Savart law are empirical*, with each tailored to account for the corresponding data.

5.4 Differential Form of the Biot–Savart Law

The *Biot–Savart law in integral form*,

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x',$$

contains (in principle) a *complete description of magnetostatics*.

- but is not always the most convenient form for solving problems.
- In many situations a differential equation is more convenient to use.
- Let us find a form of the Biot–Savart law expressed as a differential equation. From the *identity*,

$$\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3} = -\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = \boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right),$$

- which allows *converting the Biot–Savart equation into*

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{x}) &= \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\,d^3x' \\ &= \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',\end{aligned}$$

  where $\boldsymbol{\nabla}$ has been *pulled out of the integral* because it operates on $\boldsymbol{x}$ but not $\boldsymbol{x}'$.

BSdiff1.0

Remember that in these manipulations

- the integrations are over the *primed coordinates*, but
- applied gradient, divergence, and curl operations ($\boldsymbol{\nabla}$, $\boldsymbol{\nabla}\cdot$, and $\boldsymbol{\nabla}\times$, without primes) are taken with respect to the *unprimed coordinates*.
- This is made explicit in the notation exemplified in the Appendix,
- 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.
- Now take the divergence of

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x',$$

  to give

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

- But we have the identity $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{A}) = 0$ so

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0,$$

  which may be termed the *first law of magnetostatics*
- and is the *third Maxwell equation*, corresponding to the absence of magnetic charges.

Next, take the curl of $\boldsymbol{B}$,

$$\boldsymbol{\nabla}\times\boldsymbol{B}=\frac{\mu_0}{4\pi}\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

Using the vector identity $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A})=\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A})-\boldsymbol{\nabla}^2\boldsymbol{A}$, this may be written as

$$\begin{aligned}\boldsymbol{\nabla}\times\boldsymbol{B}&=\frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int\boldsymbol{J}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x'\\&\quad-\frac{\mu_0}{4\pi}\int\boldsymbol{J}(\boldsymbol{x})\,\boldsymbol{\nabla}^2\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x'\\&=-\frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int\boldsymbol{J}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right)d^3x'+\mu_0\boldsymbol{J}(\boldsymbol{x}),\end{aligned}$$

where we have used the identities,

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

(where $\boldsymbol{\nabla}$ operates on $\boldsymbol{x}$ and $\boldsymbol{\nabla}'$ operates on $\boldsymbol{x}'$) in the last step. Integrating the remaining integral by parts then gives

$$\boldsymbol{\nabla}\times\boldsymbol{B}=\mu_0\boldsymbol{J}+\frac{\mu_0}{4\pi}\boldsymbol{\nabla}\int\frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

But for steady-state magnetism $\boldsymbol{\nabla}\cdot\boldsymbol{J}=0$ and we obtain finally

$$\boldsymbol{\nabla}\times\boldsymbol{B}=\mu_0\boldsymbol{J},$$

which may be termed the *second law of magnetostatics* [and is the fourth Maxwell equation if electric fields don't depend on time (Ampère's law)].

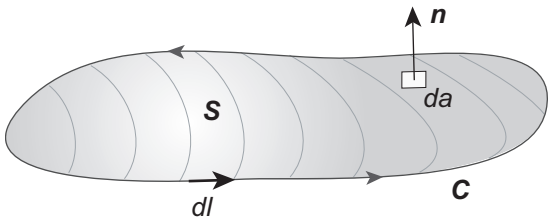


Figure 5.3: Surface S and contour C bounding the surface for Stokes' theorem. An infinitesimal line element on C is indicated by dl and an infinitesimal surface element on S is indicated by da, with the normal to da indicated by $\boldsymbol{n}$. The path in the line integration is traversed in a right-hand screw sense relative to $\boldsymbol{n}$, as indicated by arrows.

The integral equivalent of Ampère's law may be obtained from *Stokes' theorem*,

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

for the vector field $\boldsymbol{A}$ where S is an arbitrary open surface bounded by a closed curve C and where $\boldsymbol{n}$ is the normal to S. Figure 5.3 illustrates. Applying Stokes' theorem to,

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J},$$

gives

$$\int_S (\boldsymbol{\nabla} \times \boldsymbol{B}) \cdot \boldsymbol{n}\, da = \oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \boldsymbol{n}\, da$$

and therefore

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \boldsymbol{n}\, da.$$

Using that the *total current* I *passing through the closed curve* C is given by the *surface integral* on the right side of

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \boldsymbol{n}\, da.$$

the current is

$$I = \int_S \boldsymbol{J} \cdot \boldsymbol{n}\, da,$$

we finally write *Ampère's law in integral form*

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I,$$

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.

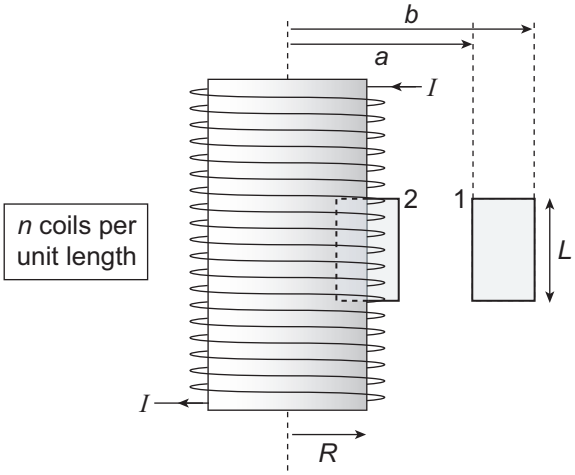


Figure 5.4: Analysis of a long, tightly wound solenoid using Ampère's law. Two rectangular loops are shown, number 1 outside the solenoid and number 2 overlapping the solenoid.

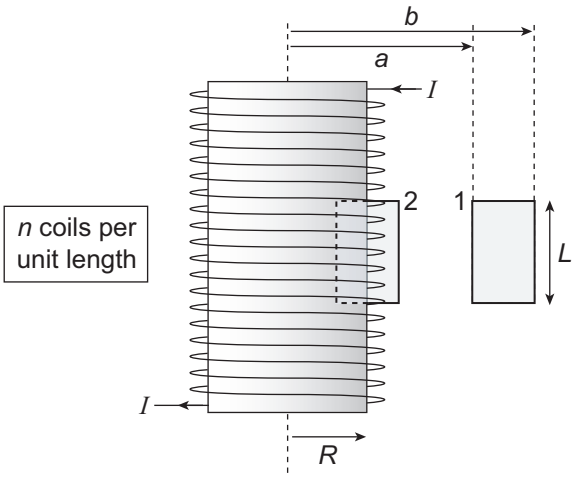
Example:

Let's determine the *magnetic field for the long solenoid* illustrated in Fig. 5.4, with n closely wound turns per unit length on a cylinder of radius R and carrying a steady current I.

- Because the solenoid is *tightly wound*, we assume that *each coil is perpendicular to the cylinder symmetry axis*.
- We expect then on symmetry and general grounds that the *magnetic field field is oriented along the cylinder axis*, and that it *drops to zero at large distance* from the solenoid.
- Let's apply *Ampère's law*

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I,$$

to the *two rectangular loops* shown in the diagram.



- *Loop 1* is completely outside the solenoid and *encloses no current*, $I_{\text{enc}} = 0$, with its left side a distance a and its right side a distance b from the central axis of the cylinder.
- Applying *Ampère's law* to it

$$\oint \boldsymbol{B} \cdot dl = [B(a) - B(b)]L = \mu_0 I_{\text{enc}} = 0,$$

  where $B = |\boldsymbol{B}|$.
- So $B(a) = B(b)$ and the magnetic field outside is independent of the distance from the solenoid.
- But the boundary conditions require that $B = 0$ at infinity, so the magnetic field must *vanish everywhere outside the solenoid*.
- *Loop 2* is halfway inside the solenoid and *Ampère's law* gives

$$\oint \boldsymbol{B} \cdot dl = BL = \mu_0 I_{\text{enc}} = \mu_0 nIL,$$

  where B *is the field inside the solenoid* (there is no contribution from the half rectangle outside the cylinder since $B = 0$ there).

- Thus *inside the solenoid the field is uniform*, $\boldsymbol{B} = \mu_0 n I \hat{\boldsymbol{z}}$, where $\hat{\boldsymbol{z}}$ is a unit vector along the cylinder axis, and
- *outside the solenoid* the magnetic field vanishes.

Like Gauss's law,

- Ampère's law is generally valid (for steady currents), but
- it is *useful* only if a problem has sufficient symmetry to permit
- B to be pulled out of the integral $\oint \boldsymbol{B} \cdot d\boldsymbol{l}$, allowing

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 I,$$

to be *solved easily for the magnetic field*.

5.5 The Vector Potential and Gauge Invariance

We have seen in the preceding section that the *basic laws of magnetostatics* are defined in differential form through

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J},$$
$$\nabla \cdot \boldsymbol{B} = 0,$$

which require *knowledge of the magnetic field* $\boldsymbol{B}$ for their use.

- In electrostatics we found that the electric potential Φ was an extremely useful quantity because
- the electric field can be derived from it by taking the gradient, $\boldsymbol{E} = -\nabla\Phi$.
- The electric potential Φ is a scalar quantity, and it is often termed the *scalar potential*.
- For the special case that *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 $\boldsymbol{B} = -\nabla\Phi_{\mathrm{M}}$.
- Then $\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J}$, with $\boldsymbol{J} = 0$ reduces to the *Laplace equation* for Φ_{M}.
- Therefore, the methods for solving Laplace's equation developed for electrostatics become applicable.

> However, this method has *limited applicability* because it is valid *only if current density vanishes*.

An approach with *more general applicability* may be developed by exploiting the *second magnetostatic equation*, $\nabla \cdot \boldsymbol{B} = 0$.

- This will allow defining a *vector potential* $\boldsymbol{A}$, from which
- the magnetic field $\boldsymbol{B}$ is derived by taking the curl of $\boldsymbol{A}$.
- We begin by noting that if $\nabla \cdot \boldsymbol{B} = 0$ is to hold everywhere,
- then $\boldsymbol{B}$ must be the curl of some vector field $\boldsymbol{A}$ (the vector potential),

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

since then

- the identity $\nabla \cdot (\nabla \times \boldsymbol{A}) = 0$ ensures that the *divergence of* $\boldsymbol{B}$ *vanishes identically* under all conditions.
- In fact, $\boldsymbol{B}$ was already written in this form earlier,

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \nabla \times \int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x',$$

- and upon *comparing the previous two equations*, an $\boldsymbol{A}$ consistent with the phenomenology of magnetism is

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x' + \nabla \psi(\boldsymbol{x}),$$

- where the addition of the *arbitrary scalar function* $\psi(\boldsymbol{x})$ has no effect on $\nabla \cdot \boldsymbol{B} = 0$
- because of the identity $\nabla \times (\nabla f) = 0$ for an arbitrary scalar function f.

Since $\psi(\boldsymbol{x})$ is an *arbitrary scalar function* of $\boldsymbol{x}$, the vector potential $\boldsymbol{A}$ can be *transformed freely* according to

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\psi,$$

- which has no effect on $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$,
- because of the identity $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{\nabla}\chi) = 0$.

Gauge Transformations:

- Adding the gradient of an arbitrary scalar ψ to a vector potential $\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\psi$ is called a *gauge transformation*, and
- the invariance of the laws of electromagnetism under this transformation is called *gauge invariance*.
- The *gauge symmetry* associated with this invariance has large implications for *classical electromagnetism* and *quantum electrodynamics*, and
- a *generalization of this gauge symmetry* is of fundamental importance in *relativistic quantum field theories*, particularly for the *Standard Model* of elementary particle physics.

From the *Helmholtz theorem*,

- a *vector field* with suitable boundary conditions is *specified uniquely by its curl and divergence*.
- From the preceding, 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 electromagnetism.
- Substituting $\boldsymbol{B}=\boldsymbol{\nabla}\times\boldsymbol{A}$ into $\boldsymbol{\nabla}\times\boldsymbol{B}=\mu_0\boldsymbol{J}$, gives

  $$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A})=\mu_0\boldsymbol{J},$$

  which becomes

  $$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A})-\boldsymbol{\nabla}^2\boldsymbol{A}=\mu_0\boldsymbol{J},$$

  upon invoking the identity

  $$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A})=\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A})-\boldsymbol{\nabla}^2\boldsymbol{A}.$$

Now we exploit the freedom to make a *gauge transformation* on

$$\nabla(\nabla\cdot\boldsymbol{A})-\nabla^2\boldsymbol{A}=\mu_0\boldsymbol{J},$$

- by *choosing a gauge* where

$$\nabla\cdot\boldsymbol{A}=0. \qquad \text{(Coulomb gauge condition)},$$

  which defines the *Coulomb gauge* (also called the *radiation gauge* or the *transverse gauge*).

- In Coulomb gauge, $\nabla\cdot\boldsymbol{A}=0$ and

$$\nabla(\nabla\cdot\boldsymbol{A})-\nabla^2\boldsymbol{A}=\mu_0\boldsymbol{J},$$

  is transformed into the *Poisson equation*,

$$\nabla^2\boldsymbol{A}=-\mu_0\boldsymbol{J} \qquad \text{(Coulomb gauge)}.$$

- From application of the Poisson equation in electrostatics we expect that

> The *solution for* $\boldsymbol{A}$ *in Coulomb gauge* is
>
> $$\boldsymbol{A}(\boldsymbol{x})=\frac{\mu_0}{4\pi}\int\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

Gauge transformations will be discussed further in Ch. 7.

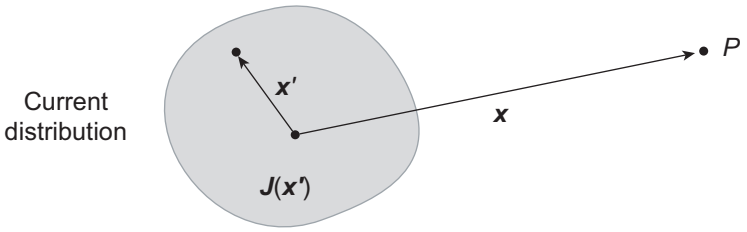


Figure 5.5: A localized current density $\boldsymbol{J}(\boldsymbol{x}')$ that produces a vector potential $\boldsymbol{A}(\boldsymbol{x})$ and corresponding magnetic field $\boldsymbol{B}(\boldsymbol{x})$ at the point $\boldsymbol{x}$, with $\boldsymbol{x} \gg \boldsymbol{x}'$.

5.6 Magnetic Fields of Localized Current Distributions

We now consider a *current distribution* that is *localized in space*.

- Figure 5.5 illustrates.
- A full treatment of this problem by analogy with electric multipole expansions is possible using *vector spherical harmonics*, which are described briefly below.

Vector Spherical Harmonics

Vector spherical harmonics are an *extension of regular (scalar) spherical harmonics* designed for use with vector fields.

- Components of vector spherical harmonics are *complex-valued functions of spherical basis vectors*.
- *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.

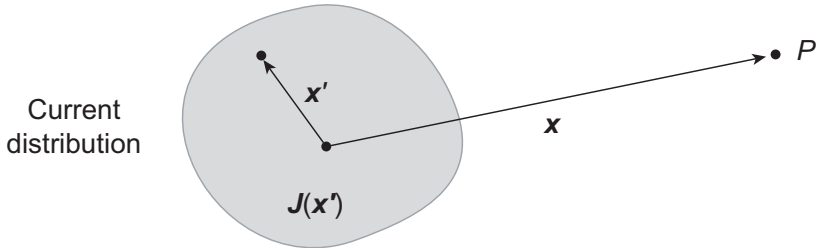
The *radial factors* ensure that the vector spherical harmonics

- have the *same dimensions* as ordinary Y_{lm}, and that
- they *do not depend on the radial coordinate*.
- These new vector fields facilitate *separation of radial from angular coordinates*,
- permitting (for example) a multipole expansion of the electric field,

$$\boldsymbol{E} = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l}\left(E^{r}_{lm}(r)\boldsymbol{Y}_{lm} + E^{(1)}_{lm}(r)\boldsymbol{\Psi}_{lm} + E^{(2)}_{lm}(r)\boldsymbol{\Phi}_{lm}\right),$$

  where 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 $\boldsymbol{r}$).



We shall *not use vector spherical harmonics* and confine ourselves to *lowest-order expansions*.

- Let us assume that $\boldsymbol{x} \gg \boldsymbol{x}'$ and *expand the denominator of*

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x'$$

in powers of $\boldsymbol{x}'$ relative to a suitable origin located in the current distribution in the above figure,

$$\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|} = \frac{1}{|\boldsymbol{x}|} + \frac{\boldsymbol{x}\cdot\boldsymbol{x}'}{|\boldsymbol{x}|^2} + \cdots .$$

- Thus the ith *components of the vector potential* $\boldsymbol{A}(\boldsymbol{x})$ has the expansion

$$A_i(\boldsymbol{x}) = \frac{\mu_0}{4\pi} \left[\frac{1}{|\boldsymbol{x}|} \int J_i(\boldsymbol{x}')\, d^3x' + \frac{\boldsymbol{x}}{|\boldsymbol{x}|^2} \cdot \int J_i(\boldsymbol{x}')\boldsymbol{x}'\, d^3x' + \cdots \right].$$

That the *current* $\boldsymbol{J}(\boldsymbol{x}')$ *is localized* and has *zero divergence* permits simplification of the expansion,

$$A_i(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left[\underbrace{\frac{1}{|\boldsymbol{x}|}\int J_i(\boldsymbol{x}')\,d^3x'}_{\text{No magnetic monopoles}} + \underbrace{\frac{\boldsymbol{x}}{|\boldsymbol{x}|^2}\cdot\int J_i(\boldsymbol{x}')\boldsymbol{x}'\,d^3x'}_{\text{Magnetic dipole}} + \cdots\right].$$

- The *first term vanishes* because the integral of the current vector over all orientations will average to zero.
- (This argument is intuitive; a *more formal proof* that the first term vanishes is given in Jackson, Section 5.6.)
- This reflects the fact that the *first term in the multipole expansion is the monopole term*, but
- there are *no magnetic monopoles* because there is *no magnetic charge* consistent with the Maxwell equations, since $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$.

The *integral in the second term* can be manipulated into

$$\begin{aligned}
\boldsymbol{x}\cdot\int \boldsymbol{x}' J_i(\boldsymbol{x}')\,d^3x' &= \sum_i x_i \int x'_j J_i\,d^3x' \\
&= -\frac{1}{2}\sum_i x_i \int (x'_i J_i - x'_j J_i)\,d^3x' \\
&= -\frac{1}{2}\sum_{j,k} \varepsilon_{ijk} x_j \int (\boldsymbol{x}'\times\boldsymbol{J})_k\,d^3x' \\
&= -\frac{1}{2}\left[\boldsymbol{x}\times\int(\boldsymbol{x}'\times\boldsymbol{J})\,d^3x\right]_i .
\end{aligned}$$

The *magnetic moment density* or *magnetization* is defined by

$$\boldsymbol{M}(\boldsymbol{x}) = \frac{1}{2}\left[\boldsymbol{x} \times \boldsymbol{J}(\boldsymbol{x})\right],$$

and the *magnetic moment* $\boldsymbol{m}$ is defined by

$$\boldsymbol{m} = \frac{1}{2}\int \boldsymbol{x}' \times \boldsymbol{J}(\boldsymbol{x}')\, d^3x'.$$

Then the multipole expansion

$$A_i(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left[\underbrace{\frac{1}{|\boldsymbol{x}|}\int J_i(\boldsymbol{x}')\, d^3x'}_{\text{monopole} = 0} + \underbrace{\frac{\boldsymbol{x}}{|\boldsymbol{x}|^2}\cdot\int J_i(\boldsymbol{x}')\boldsymbol{x}'\, d^3x'}_{\text{dipole}} + \cdots\right].$$

can be written as

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\boldsymbol{x}}{|\boldsymbol{x}|^3} + \cdots.$$

This *first non-vanishing term* in the multipole expansion

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times\boldsymbol{x}}{|\boldsymbol{x}|^3} + \cdots .$$

for a steady-state current distribution

- has the form of a *magnetic dipole*.
- At large distances from the current source the *higher order terms may be ignored* and
- the *magnetic field corresponds to the curl of the dipole term*, which gives

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left(\frac{3\boldsymbol{n}(\boldsymbol{n}\cdot\boldsymbol{m}) - \boldsymbol{m}}{|\boldsymbol{x}|^3}\right),$$

  where $\boldsymbol{n}$ is a unit vector in the direction of $\boldsymbol{x}$.

Because the dipole term typically dominates the magnetic multipole expansion, the *magnetic dipole moment* is commonly called 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 loop shape*,

$$|\boldsymbol{m}| = I \times A,$$

  where A is the *enclosed area* of 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

$$\boldsymbol{m} = \sum_i \frac{q_i}{2M_i} \boldsymbol{L}_i.$$

If all the particles have the *same charge-to-mass ratio* $q_i/M_i = e/M$,

$$\boldsymbol{m} = \frac{e}{2M}\sum_i \boldsymbol{L}_i = \frac{e}{2M}\boldsymbol{L},$$

where $\boldsymbol{L}$ is the *total orbital angular momentum*.

- This is the well-known *classical connection* between *angular momentum* and *magnetic moment*.
- It holds for *orbital angular momentum* even on the atomic scale,
- but *fails for intrinsic moments* arising from *particle spin*.

> This is an *intrinsically quantum effect* that is beyond the scope of our present discussion of classical electromagnetism.

Chapter 6

Magnetic Fields in Matter

Electric fields *polarize matter*; magnetic fields *magnetize matter*.

- In *classical electromagnetism* all magnetic phenomena have their origin in the *motion of electrical charges*.
- At the microscopic level, we may view magnetism as being produced by *small current loops*.

  For example, *electrons in orbits* around nuclei in atoms that act 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*,

- it can cause a *net alignment of the dipoles* so that the matter becomes magnetically polarized (*magnetized*).
- For the *electric polarization of matter* described in Ch. 4, the polarization is usually *in the direction of the* $\boldsymbol{E}$ *field*.
- *Magnetic polarization of matter* is *more complex and varied* than electric polarization of matter.

For example,

1. *paramagnetic materials* acquire a magnetization in the *same direction* as the applied field $\boldsymbol{B}$, while
2. the magnetization of *diamagnetic materials* is in the direction *opposite the applied field*, and
3. *ferromagnetic materials*
   - retain their magnetization after the polarizing field has been removed, with
   - the retained magnetization depending on the *entire magnetic history* of the material.

6.1 Ampère's Law in Magnetically Polarized Matter

In the previous chapter we have dealt with steady-state magnetic fields in a microscopic manner, *assuming the current density* $\boldsymbol{J}$ *to be a known function of position*.

- In macroscopic problems dealing with magnetic effects in materials, *this will often not be true*.
- The atomic currents in matter give rise to *rapidly fluctuating current densities* on a microscopic scale.
- Just as for electric fields in matter, *only averages over macroscopic volumes are known*
- and *only these* enter into the classical equations of electromagnetism.

6.1.1 Macroscopic Averaging

The first step in introducing averaging in matter for *electric fields* in Section 4.7 was to note that

- the averaging procedure *preserves the crucial relation* $\nabla \times \boldsymbol{E} = 0$,
- which ensures that the *macroscopic electric field* is still *derivable from a scalar potential:* $\boldsymbol{E} = -\nabla\Phi$.

In a similar manner, for *magnetic fields*

- the macroscopic averaging procedure leads to the *same equation* $\nabla \cdot \boldsymbol{B} = 0$ *that we found in vacuum*.
- Thus, the averaging procedure *preserves the notion of a vector potential* $\boldsymbol{A}(\boldsymbol{x})$,
- from which we can *derive the macroscopic magnetic field by taking the curl*, $\boldsymbol{B} = \nabla \times \boldsymbol{A}$.

Magnetization establishes *currents within a material and on its surface* that produce the field due to magnetization.

- The average *macroscopic magnetization* or *magnetic (dipole) moment density* is

$$\boldsymbol{M}(\boldsymbol{x}) = \sum_i N_i \langle \boldsymbol{m}_i \rangle,$$

- 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* averaged over a small volume ΔV around $\boldsymbol{x}'$ will take the form

$$\Delta \boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left[\underbrace{\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}}_{\text{free charge}} + \underbrace{\frac{\boldsymbol{M}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}}_{\text{medium dipoles}}\right]\Delta V,$$

- 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

$$A_i(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\left[\frac{1}{|\boldsymbol{x}|}\int J_i(\boldsymbol{x}')\,d^3x' + \frac{\boldsymbol{x}}{|\boldsymbol{x}|^2}\cdot\int J_i(\boldsymbol{x}')\boldsymbol{x}'\,d^3x' + \cdots\right].$$

- Letting $\Delta V \to d^3x'$, the *total vector potential at* $\boldsymbol{x}$ is given by an integral over all space,

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int\left[\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} + \frac{\boldsymbol{M}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\right]d^3x'.$$

The second (magnetization) term in

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \left[\frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} + \frac{\boldsymbol{M}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\right] d^3x'.$$

can be cast in another form by *utilizing the identity*

$$\frac{\boldsymbol{x}-\boldsymbol{x}'}{|\boldsymbol{x}-\boldsymbol{x}'|^3} = -\boldsymbol{\nabla}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) = \boldsymbol{\nabla}'\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right),$$

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

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} d^3x' + \underbrace{\frac{\mu_0}{4\pi}\int \frac{\boldsymbol{M}(\boldsymbol{x}')\times(\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3} d^3x'}_{\text{integrate by parts}}$$

to be written

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\left[\boldsymbol{J}(\boldsymbol{x}')+\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')\right]}{|\boldsymbol{x}-\boldsymbol{x}'|} d^3x',$$

where a *surface term has been discarded* by assuming $\boldsymbol{M}(\boldsymbol{x}')$ to be localized and well behaved.

6.1.2 The Auxiliary Field H and Constituitive Relations

From the results in the preceding section we see that

- the *magnetization* contributes an *effective current density*

$$\boldsymbol{J}_{\mathrm{M}} = \boldsymbol{\nabla} \times \boldsymbol{M}.$$

- Then $\boldsymbol{J} + \boldsymbol{J}_{\mathrm{M}}$ plays the role of the *effective current* such that

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0(\boldsymbol{J} + \boldsymbol{J}_{\mathrm{M}}) = \mu_0(\boldsymbol{J} + \boldsymbol{\nabla} \times \boldsymbol{M}).$$

- It is then convenient to define a *new macroscopic field*

$$\boldsymbol{H} \equiv \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M}.$$

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 calls $\boldsymbol{H}$ the *magnetic field* and $\boldsymbol{B}$ the *magnetic induction*.

This is a question of *terminology, not physics*.

- 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*.

The introduction of $\boldsymbol{H}$ in the presence of *magnetically polarized materials* is a matter of *convenience*.

- It is analogous to the introduction 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* $\boldsymbol{B}$.

This is a *classical statement*. As we shall discuss briefly later, in *quantum mechanics* one finds that

- the *scalar potential* Φ and the *vector potential* $\boldsymbol{A}$ should be viewed as *more fundamental than the electric and magnetic fields*.
- There are two basic reasons:
  1. Some 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).

The *derived fields* $\boldsymbol{D}$ and $\boldsymbol{H}$ are

- introduced as a convenient way to take into account the average contributions to
  - the *charge density* and
  - the *current*

  of the atomic-level charges and currents.
- Then the *macroscopic fields in medium* that replace the microscopic fields are

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J},$$
$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0,$$

- which are analogous to the *macroscopic fields in medium* for electrostatics,

$$\boldsymbol{\nabla} \times \boldsymbol{E} = 0,$$
$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho,$$

  that were derived in earlier chapters.

Just as for electrostatics, the description of *macroscopic magnetostatics* requires *constituitive relationships*.

- For *magnetic effects* , they are 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*,

$$\boldsymbol{B} = \mu \boldsymbol{H},$$

  where the constant μ is *characteristic of the medium* and is called the *magnetic permeability*. Generally,

  - $\mu > 1$ for *paramagnetic materials* and
  - $\mu < 1$ for *diamagnetic materials*,

- with μ/μ_0 differing from unity by a part in $\sim 10^5$ for either case.

In *diamagnetic and paramagnetic materials* the deviation of μ from μ_0 is typically small.

The physical reasons for paramagnetism and diamagnetism are discussed in the following.

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 $\boldsymbol{M}$ that opposes $\boldsymbol{H}$, 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*).

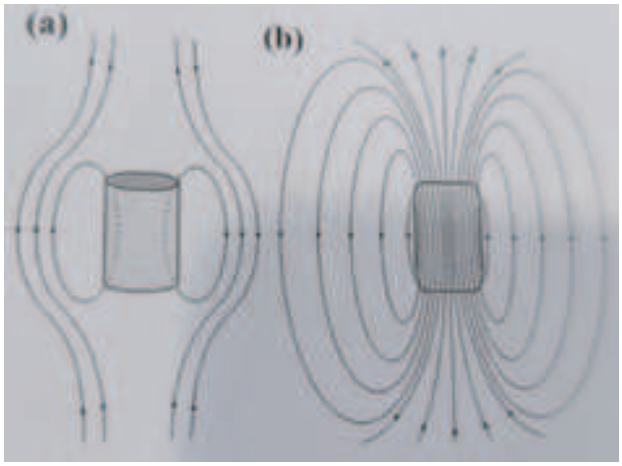
Diamagnetic effects are greatly amplified in a superconductor because the induced currents flow without resistance. It may be shown in quantum field theory that the (normally massless) photon gains an effective mass through interaction with the dielectric medium, which causes it to penetrate the superconductor 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. 9, this *Higgs mode of spontaneous symmetry breaking* is fundamental for the Standard Model of elementary particle physics.

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.



The diamagnetic matter in (a) expels the magnetic field but the paramagnetic matter in (b) enhances the density of field lines within the sample.

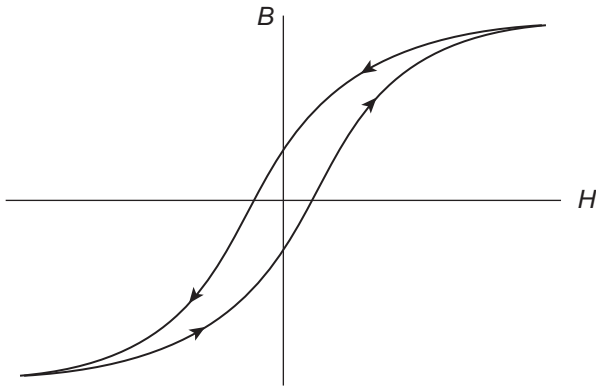


Figure 6.1: Schematic illustration of hysteresis.

The case of *ferromagnetic* materials is *more consequential* and *more complex*,

- leading to a *constituitive relationship*

$$\boldsymbol{B} = \boldsymbol{F}[\boldsymbol{H}],$$

  where $\boldsymbol{F}[\boldsymbol{H}]$ is a *non-linear function of* $\boldsymbol{H}$.

- Ferromagnets may exhibit *hysteresis*, where
- the magnetic field $\boldsymbol{B}$ is *not a single-valued function* of $\boldsymbol{H}$, and
- the state of the system may *depend on its preparation history*; Fig. 6.1 illustrates.
- Clearly the *complex relationship of* $\boldsymbol{B}$ *and* $\boldsymbol{H}$ in ferromagnetic materials means that
- *magnetic boundary-value problems are more difficult* to deal with than corresponding problems in electrostatics.

6.1.3 Magnetic Boundary-Value Conditions

The *boundary conditions* for $\boldsymbol{B}$ and $\boldsymbol{H}$ at the interfaces between media were considered earlier.

- The *normal components of* $\boldsymbol{B}$ and *tangential components of* $\boldsymbol{H}$ on opposite sides of a boundary separating medium 1 and medium 2 are related by

$$(\boldsymbol{B}_2 - \boldsymbol{B}_1) \cdot \boldsymbol{n} = 0,$$
$$\boldsymbol{n} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{K},$$

  - where $\boldsymbol{n}$ is the *unit normal vector* pointing *from region 1 into region 2*, and
  - $\boldsymbol{K}$ is the idealized *surface current density*.

- If the media satisfy a *linear constituitive relation* and have *finite conductivities* so that $\boldsymbol{J} = \sigma \boldsymbol{E}$ and $\boldsymbol{K} = 0$,

  - the boundary conditions can be expressed as,

$$\boldsymbol{B}_2 \cdot \boldsymbol{n} = \boldsymbol{B}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{B}_2 \times \boldsymbol{n} = \frac{\mu_2}{\mu_1} \boldsymbol{B}_1 \times \boldsymbol{n},$$

  - or as

$$\boldsymbol{H}_2 \cdot \boldsymbol{n} = \frac{\mu_1}{\mu_2} \boldsymbol{H}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{H}_2 \times \boldsymbol{n} = \boldsymbol{H}_1 \times \boldsymbol{n}.$$

6.2 Solving Magnetostatic Boundary-Value Problems

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.

Magnetostatic boundary value problems require solving for

- the *charge density* and
- the *current*

of the atomic-level charges and currents.

Then the *macroscopic fields in medium* that replace the microscopic fields are

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{B} = 0,$$

subject to constituitive relations

$$\boldsymbol{B} = \mu \boldsymbol{H} \ \text{(linear)} \quad \text{or} \quad \boldsymbol{B} = \boldsymbol{F}[\boldsymbol{H}] \ \text{(non-linear)},$$

Especially because of the varied constituitive relations a variety of situations are possible and a *survey of possible approaches is useful*. We summarize some approaches following the presentation in Jackson.

6.2.1 Method of the Vector Potential

Because $\nabla \cdot \boldsymbol{B} = 0$,

It is always possible to introduce a *vector potential* $\boldsymbol{A}(\boldsymbol{x})$ such that $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, so that

$$\nabla \cdot \boldsymbol{B} = \nabla \cdot (\nabla \times \boldsymbol{A}) = 0.$$

- If we have a *non-linear constituitive relation* $\boldsymbol{B} = \boldsymbol{F}[\boldsymbol{H}]$, the resulting differential equation

  $$\nabla \times \boldsymbol{H}[\nabla \times \boldsymbol{A}] = \boldsymbol{J}$$

  is generally very *difficult to solve*.

- However, if the *constituitive relation is linear*, $\boldsymbol{B} = \mu \boldsymbol{H}$, the preceding equation becomes

  $$\nabla \times \left(\frac{1}{\mu} \nabla \times \boldsymbol{A} \right) = \boldsymbol{J}.$$

For a region of space in which μ is constant,

- the *vector identity*

$$\nabla \times (\nabla \times \boldsymbol{A}) = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}$$

  may be used to write

$$\nabla \times \left(\frac{1}{\mu} \nabla \times \boldsymbol{A} \right) = \boldsymbol{J}.$$

  as

$$\nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} = \mu \boldsymbol{J}.$$

- Invoking the *Coulomb gauge condition* $\nabla \cdot \boldsymbol{A} = 0$, we obtain the *Poisson equation*

$$\nabla^2 \boldsymbol{A} = -\mu \boldsymbol{J}.$$

- Comparing with

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

  in vacuum,

- this is seen to be a Poisson equation with a *current density modified by the medium* ($\mu \neq \mu_0$),

- which is similar to a Poisson equation for *uniform dielectric media* with an *effective charge density* $(\varepsilon/\varepsilon_0)$.

Matching of solutions for

$$\nabla^2 \boldsymbol{A} = -\mu \boldsymbol{J}.$$

across interfaces between different (linear) media can be implemented using the *boundary conditions* described earlier,

$$\boldsymbol{B}_2 \cdot \boldsymbol{n} = \boldsymbol{B}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{B}_2 \times \boldsymbol{n} = \frac{\mu_2}{\mu_1} \boldsymbol{B}_1 \times \boldsymbol{n},$$

or

$$\boldsymbol{H}_2 \cdot \boldsymbol{n} = \frac{\mu_1}{\mu_2} \boldsymbol{H}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{H}_2 \times \boldsymbol{n} = \boldsymbol{H}_1 \times \boldsymbol{n}.$$

6.2.2 Using a Magnetic Scalar Potential if $\boldsymbol{J} = 0$

As mentioned earlier, for the *special case* $\boldsymbol{J} = 0$ where the current density vanishes in a region of interest,

- $\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J} \quad \longrightarrow \quad \boldsymbol{\nabla} \times \boldsymbol{H} = 0,$
- which suggests introducing a *magnetic scalar potential* Φ_{M} such that

$$\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}}.$$

> Recall that the *electric field* can be derived from $\boldsymbol{E} = -\boldsymbol{\nabla}\Phi$ because $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$.

- If the *medium is linear* and μ *is constant*,
- the *magnetic scalar potential* satisfies the *Laplace equation*

$$\boldsymbol{\nabla}^2\Phi_{\mathrm{M}} = 0,$$

for which the *boundary conditions*

$$\boldsymbol{H}_2 \cdot \boldsymbol{n} = \frac{\mu_1}{\mu_2}\boldsymbol{H}_1 \cdot \boldsymbol{n} \qquad \boldsymbol{H}_2 \times \boldsymbol{n} = \boldsymbol{H}_1 \times \boldsymbol{n}.$$

are appropriate.

- This method can be used *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 below.

6.2.3 Hard Ferromagnets

A *hard ferromagnet* has

- a *magnetization independent of applied field* for moderate field strengths.
- This suggests an *approximation* where a hard ferromagnet can be treated as if it has a *specified fixed magnetization* $\boldsymbol{M}(\boldsymbol{x})$ and $\boldsymbol{J} = 0$,
- using either *magnetic scalar potential* methods or *vector potential* methods.

Solve Using the Magnetic Scalar Potential

Since $\boldsymbol{J} = 0$ for a hard ferromagnetic, the *magnetic scalar potential method* is applicable. Then $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ becomes

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = \mu_0 \boldsymbol{\nabla} \cdot (\boldsymbol{H} + \boldsymbol{M}) = 0,$$

where

$$\boldsymbol{H} \equiv \frac{1}{\mu_0} \boldsymbol{B} - \boldsymbol{M}.$$

was used, and since $\boldsymbol{H} = -\boldsymbol{\nabla} \Phi_{\mathrm{M}}$, this becomes a *magnetic Poisson equation*,

$$\boldsymbol{\nabla}^2 \Phi_{\mathrm{M}} = \boldsymbol{\nabla} \cdot \boldsymbol{M} = -\rho_{\mathrm{M}},$$

where an *effective magnetic-charge density*

$$\rho_{\mathrm{M}} \equiv -\boldsymbol{\nabla} \cdot \boldsymbol{M}$$

has been introduced. By analogy with the *second term of*

$$\Phi(\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0} \int d^3x' \frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \left[\rho(\boldsymbol{x}') - \boldsymbol{\nabla}' \cdot \boldsymbol{P}(\boldsymbol{x}') \right],$$

for *electrostatics in electrically polarized matter*, if there are no boundary surfaces the solution is expected to be

$$\Phi_{\mathrm{M}}(\boldsymbol{x}) = -\frac{1}{4\pi} \int \frac{\boldsymbol{\nabla}' \cdot \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x = -\frac{1}{4\pi} \boldsymbol{\nabla} \cdot \int \frac{\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x,$$

where the second form results from an *integration by parts* and

$$\frac{\boldsymbol{x} - \boldsymbol{x}'}{|\boldsymbol{x} - \boldsymbol{x}'|^3} = -\boldsymbol{\nabla} \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) = \boldsymbol{\nabla}' \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right),$$

which is justified if $\boldsymbol{M}$ is *localized and well behaved*.

Physical magnetization distributions generally don't have discontinuities.

- However, 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

$$\sigma_{\mathrm{M}} = \boldsymbol{n} \cdot \boldsymbol{M},$$

  where $\boldsymbol{n}$ is the outward normal at the surface.
- Then the first form of the potential (6.2.3) is modified to

$$\Phi_{\mathrm{M}}(\boldsymbol{x}) = -\frac{1}{4\pi} \int_V \frac{\boldsymbol{\nabla}' \cdot \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x + \frac{1}{4\pi} \oint_S \frac{\boldsymbol{n}' \cdot \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} da'.$$

6.2.4 Solve Using the Vector Potential

An important special case is for *uniform magnetization of the volume*.

- Then $\boldsymbol{J}' = 0$ inside the volume the first term of

$$\Phi_{\mathrm{M}}(\boldsymbol{x}) = \underbrace{-\frac{1}{4\pi}\int_V \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x}_{=0} + \frac{1}{4\pi}\oint_S \frac{\boldsymbol{n}'\cdot\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da'.$$

  vanishes and

$$\Phi_{\mathrm{M}}(\boldsymbol{x}) = \frac{1}{4\pi}\oint_S \frac{\boldsymbol{n}'\cdot\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da' = \frac{1}{4\pi}\oint_S \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,da',$$

  where $\sigma_{\mathrm{M}} = \boldsymbol{n}\cdot\boldsymbol{M}$, was used in the last step.

Thus, introduction of a *sharp boundary for the volume of a uniformly magnetized object* induces a *surface charge* on the boundary given by σ_{M}.

If we choose to satisfy $\boldsymbol{\nabla}\cdot\boldsymbol{B}$ automatically by introducing a vector potential $\boldsymbol{A}$ and defining $\boldsymbol{B}\equiv\boldsymbol{\nabla}\times\boldsymbol{A}$,

- then $\boldsymbol{\nabla}\times\boldsymbol{H}=\boldsymbol{J}$ becomes

$$\boldsymbol{\nabla}\times\boldsymbol{H}=\boldsymbol{\nabla}\times\underbrace{\left(\frac{1}{\mu_0}\boldsymbol{B}-\boldsymbol{M}\right)}_{\boldsymbol{H}}=0,$$

- 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

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A})=\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A})-\boldsymbol{\nabla}^2\boldsymbol{A}$$

on the left side and

$$\boldsymbol{J}_{\mathrm{M}}\equiv\boldsymbol{\nabla}\times\boldsymbol{M}.$$

on the right side,

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A})-\boldsymbol{\nabla}^2\boldsymbol{A}=\mu_0\boldsymbol{J}_{\mathrm{M}},$$

where $\boldsymbol{J}_{\mathrm{M}}$ is the *effective magnetic current density*.

- Transforming

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \boldsymbol{\nabla}^2\boldsymbol{A} = \mu_0 \boldsymbol{J}_{\mathrm{M}},$$

  to *Coulomb gauge* by setting $\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0$ then leads to a Poisson equation for the vector potential,

$$\boldsymbol{\nabla}^2\boldsymbol{A} = -\mu_0 \boldsymbol{J}_{\mathrm{M}}.$$

- If there are no bounding surfaces, the solution is

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x,$$

- which may be compared with

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\left[\boldsymbol{J}(\boldsymbol{x}')+\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')\right]}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',$$

  with $\boldsymbol{J} = 0$.

If the *magnetization is discontinuous* (as in a uniformly magnetized sphere with sharp surface)

- it is necessary to add a *surface integral* to

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{\nabla}' \times \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x.$$

- For the case of $\boldsymbol{M}$ falling to zero at the surface S bounding the volume V, starting from

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x' + \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{M}(\boldsymbol{x}') \times (\boldsymbol{x}-\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|^3}\, d^3x',$$

- the proper generalization may be shown to be

$$\boldsymbol{A}(\boldsymbol{x}) = \underbrace{\frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{\nabla}' \times \boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d^3x}_{\text{Volume}} + \underbrace{\frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}') \times \boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}\, da'}_{\text{Surface}}.$$

- For the special case that the *magnetization is constant over the volume* V, only the surface integral can contribute and

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}') \times \boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}\, da'.$$

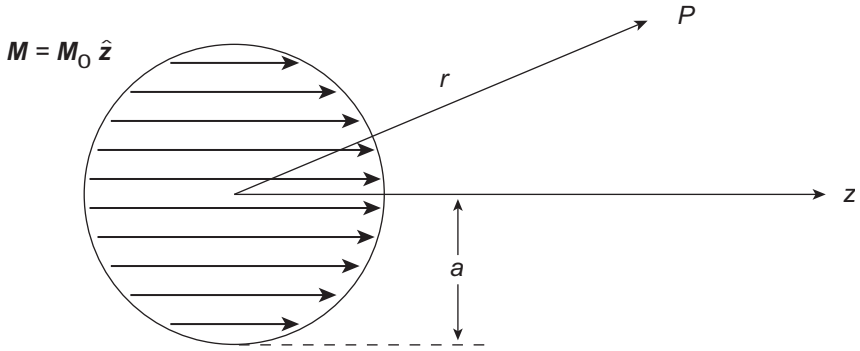


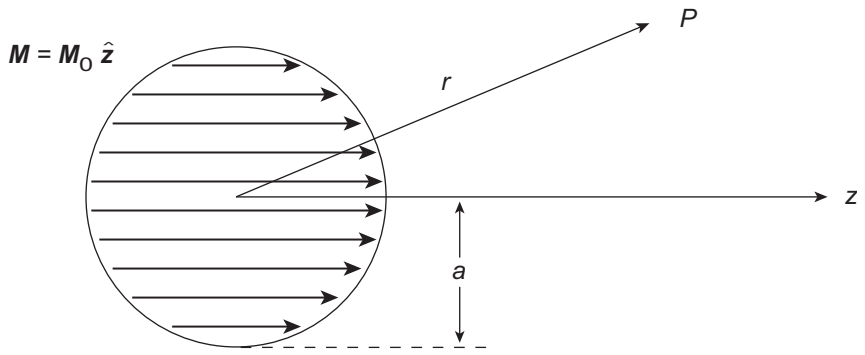
Figure 6.2: Uniformly magnetized sphere of radius a, with a sharp surface and surface magnetic-charge density $\sigma_{\rm M}(\theta)$ with $\boldsymbol{M} = \boldsymbol{M}_0\hat{\boldsymbol{z}}$ parallel to the z axis, so that $\sigma_{\rm M} = \boldsymbol{n}\cdot\boldsymbol{M} = M_0\cos\theta$.

6.2.5 Example: Uniformly Magnetized Sphere

As an example of solving boundary problems in magnetostatics using the methods just discussed,

- let's consider a *sphere with uniform permanent magnetization*,
- with a *sharp transition at the surface* from magnetized to non-magnetized matter.

Fig. 6.2 illustrates.



Solution Using the Magnetic Scalar Potential

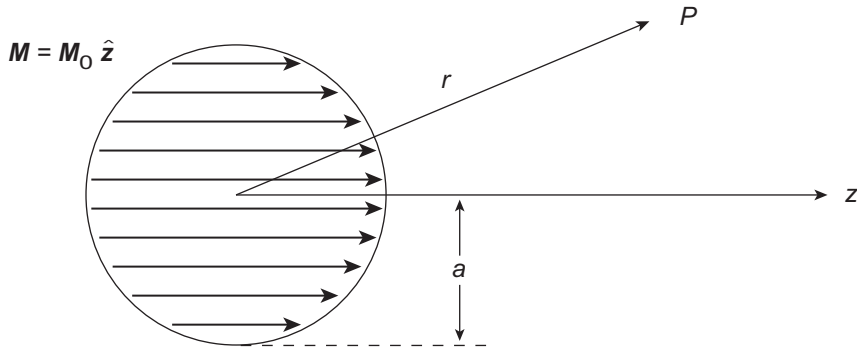
Since the sphere is uniformly magnetized,

- there are *no volume currents* and
- the *scalar magnetic potential method* is applicable.

There will be a *surface charge* $\sigma_{\mathrm{M}} = \boldsymbol{n} \cdot \boldsymbol{M}$ associated with the *sharp transition from magnetized to un-magnetized material*.

- Assuming that $\boldsymbol{M} = M_0 \hat{\boldsymbol{z}}$ so that $\sigma_{\mathrm{M}} = \boldsymbol{n} \cdot \boldsymbol{M} = \boldsymbol{M}_0 \cos\theta$ using spherical coordinates,
- the *scalar magnetic potential* is

$$\Phi_{\mathrm{M}}(r,\theta) = \frac{1}{4\pi} \oint_S \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, da' = \frac{M_0 a^2}{4\pi} \int \frac{\cos\theta}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d\Omega'.$$



Inserting the *multipole expansion*

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

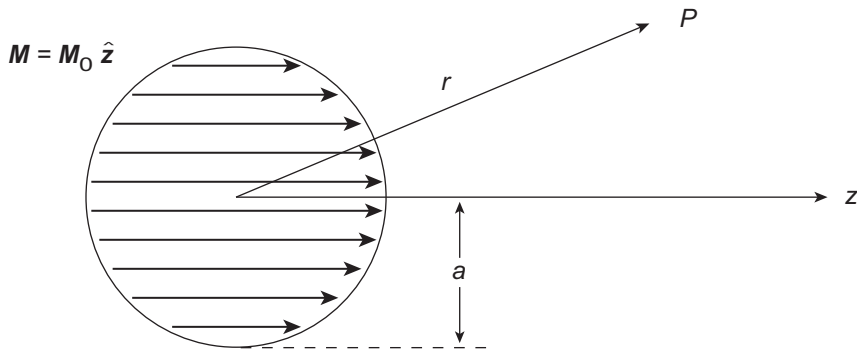
into

$$\Phi_{\mathrm{M}}(r,\theta) = \frac{1}{4\pi} \oint_S \frac{\sigma(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\, da' = \frac{M_0 a^2}{4\pi} \int \frac{\cos\theta}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d\Omega'.$$

and using that $P_1(\cos\theta) = \cos\theta$ gives

$$\begin{aligned}
\Phi_{\mathrm{M}}(r,\theta) &= \frac{M_0 a^2}{4\pi} \int \frac{P_1(\cos\theta)}{|\boldsymbol{x}-\boldsymbol{x}'|}\, d\Omega' \\
&= \frac{M_0 a^2}{4\pi} \sum_{l=0}^{\infty} \frac{r_<^l}{r_>^{l+1}} \int P_l(\cos\theta) P_1(\cos\theta)\, d\Omega' \\
&= \frac{1}{3} M_0 a^2 \frac{r_<}{r_>^2} P_1(\cos\theta) = \frac{1}{3} M_0 a^2 \frac{r_<}{r_>^2} \cos\theta,
\end{aligned}$$

- where the *Legendre orthogonality condition* has been used to eliminate all terms in the sum except for $l = 1$, and
- $(r_<, r_>)$ are the *smaller and larger* of r and a, respectively.



Inside the sphere, $r_< = r$ and $r_> = a$, so the *magnetic scalar potential* is

$$\Phi_{\rm M}^{\rm in} = \frac{1}{3} M_0 r \cos\theta = \frac{1}{3} M_0 z.$$

Then from $\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\rm M}$,

$$\boldsymbol{H}_{\rm in} = -\boldsymbol{\nabla}\Phi_{\rm M}^{\rm in} = -\frac{\partial}{\partial z}\left(\frac{1}{3} M_0 z\right)\hat{\boldsymbol{z}} = -\frac{1}{3} M_0 \hat{\boldsymbol{z}} = -\frac{1}{3}\boldsymbol{M},$$

and from $\boldsymbol{H} \equiv \dfrac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M}$,

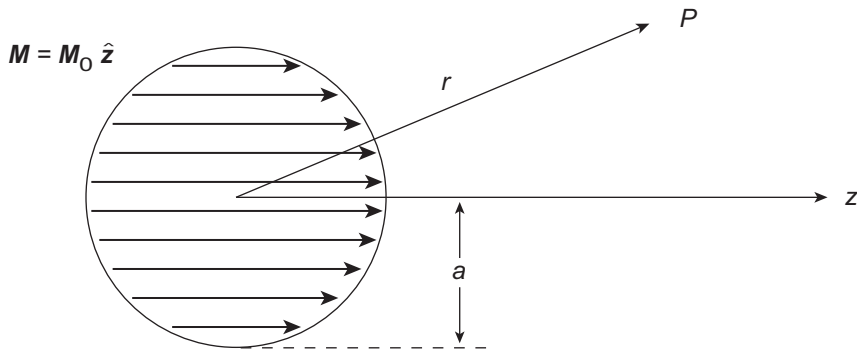
$$\boldsymbol{B}_{\rm in} = \mu_0(\boldsymbol{H}_{\rm in} + \boldsymbol{M}) = \mu_0\left(-\frac{1}{3}\boldsymbol{M} + \boldsymbol{M}\right) = \frac{2\mu_0}{3}\boldsymbol{M}.$$

Therefore, the *interior solution* is

$$\Phi_{\rm M}^{\rm in} = \frac{1}{3} M_0 z \qquad \boldsymbol{H}_{\rm in} = -\frac{1}{3}\boldsymbol{M} \qquad \boldsymbol{B}_{\rm in} = \frac{2\mu_0}{3}\boldsymbol{M},$$

where we see that the fields are *constant inside the sphere*, with

- $\boldsymbol{B}$ *parallel to* $\boldsymbol{M}$ and
- $\boldsymbol{H}$ *antiparallel to* $\boldsymbol{M}$.



For the *exterior solution*, $r_< = a$ and $r_> = r$, which gives from

$$\Phi_{\rm M}(r,\theta) = \frac{1}{3} M_0 a^2 \frac{r_<}{r_>^2} \cos\theta,$$

for the exterior potential,

$$\Phi_{\rm M}^{\rm out} = \frac{1}{3} M_0 a^3 \frac{\cos\theta}{r^2},$$

which is a *dipole potential* with a *dipole moment*

$$\boldsymbol{m} = \frac{4\pi a^3}{3} \boldsymbol{M}.$$

Then, proceeding as above

$$\boldsymbol{H}_{\rm out} = -\nabla \Phi_{\rm M}^{\rm out} = -\frac{1}{3}\frac{a^3}{r^3}\boldsymbol{M},$$

$$\boldsymbol{B}_{\rm out} = \mu_0(\boldsymbol{H}+\boldsymbol{M}) = \left(\frac{3r^3 - a^3}{3r^3}\right)\mu_0 \boldsymbol{M}.$$

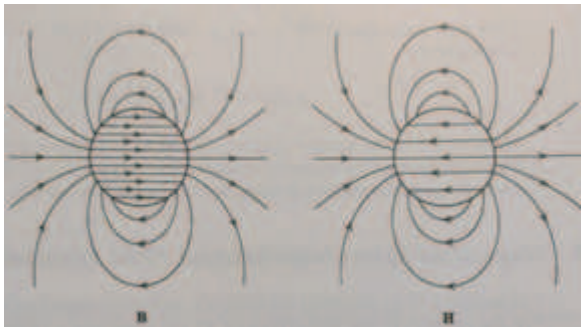
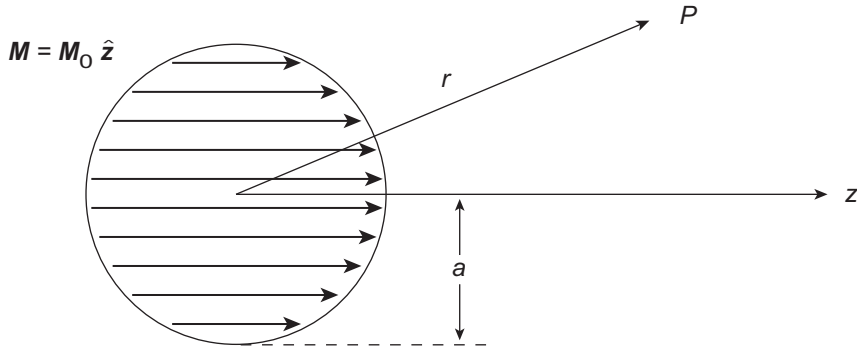


Figure 6.3: Lines of $\boldsymbol{B}$ and $\boldsymbol{H}$ for a uniformly magnetized sphere having a sharp boundary.

The $\boldsymbol{B}$ and $\boldsymbol{H}$ fields are plotted in Fig. 6.3



Solution Using the Vector Potential

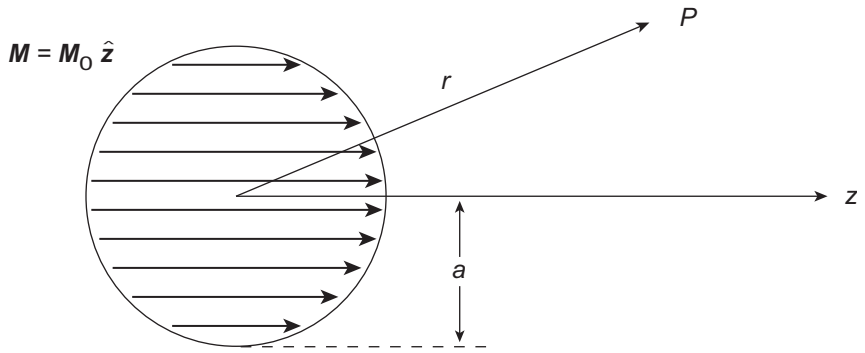
Alternatively, the *vector potential*

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{\nabla}'\times\boldsymbol{M}(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x + \frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}')\times\boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}da',$$

can be used to obtain the solution for the problem posed in above figure.

- The *magnetization is assumed uniform* so the volume current $\boldsymbol{J}_{\mathrm{M}} = 0$ and
- the first (volume) term make *no contribution*.
- However, there is a *surface charge* due to the sharp boundary on magnetization so the *second term will be non-zero* and

$$\boldsymbol{A}(\boldsymbol{x}) = \frac{\mu_0}{4\pi}\oint_S \frac{\boldsymbol{M}(\boldsymbol{x}')\times\boldsymbol{n}'}{|\boldsymbol{x}-\boldsymbol{x}'|}da'.$$



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

$$\begin{aligned} \boldsymbol{M}(\boldsymbol{x}') \times \boldsymbol{n}' &= M_0 \sin\theta' \boldsymbol{\varepsilon}_\phi \\ &= M_0 \sin\theta' (-\sin\phi \boldsymbol{\varepsilon}_1 + \cos\phi' \boldsymbol{\varepsilon}_2). \end{aligned}$$

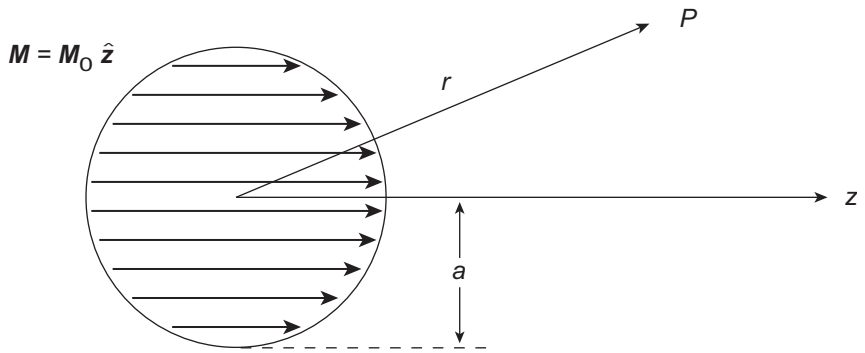
- The problem has azimuthal (ϕ) symmetry about the z-axis.
- If the observing point P is chosen to lie in the x-y plane,
- then only the y component of $\boldsymbol{M} \times \boldsymbol{n}$ survives integration over the azimuth so

$$-\sin\phi' \boldsymbol{\varepsilon}_1 + \cos\phi' \boldsymbol{\varepsilon}_2 \quad \longrightarrow \quad \cos\phi' \boldsymbol{\varepsilon}_2,$$

- which gives an azimuthal component of $\boldsymbol{A}$,

$$A_\phi = \frac{\mu_0}{4\pi} M_0 a^2 \int \frac{\sin\theta' \cos\phi'}{|\boldsymbol{x} - \boldsymbol{x}'|} d\Omega',$$

where the components of $\boldsymbol{x}'$ are $(r' = a, \theta', \phi')$, with $r' = a$ *confining the integration to the surface of the sphere*.



Since

$$Y_{11}(\theta',\phi') = -\sqrt{\frac{3}{8\pi}}\sin\theta' e^{i\phi'} = -\sqrt{\frac{3}{8\pi}}\sin\theta'(\cos\phi' + i\sin\phi'),$$

we may write

$$\sin\theta'\cos\phi' = -\sqrt{\frac{8\pi}{3}}\,\mathrm{Re}\,[\mathrm{Y}_{11}(\theta',\phi')]$$

where $\mathrm{Re}\,(\mathrm{x})$ denotes the *real part of* x. Thus,

$$A_\phi = -\sqrt{\frac{3}{8\pi}}\frac{\mu_0}{4\pi}M_0 a^2 \int \frac{\mathrm{Re}\,[\mathrm{Y}_{11}(\theta',\phi')]}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d\Omega'.$$

Then if the denominator is expanded using ,

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

the *spherical harmonic orthogonality relation*,

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

ensures that *only the* $l = 1, m = 1$ *term survives the summation* and the *vector potential* is

$$A_\phi(\boldsymbol{x}) = \frac{\mu_0}{3} M_0 a^2 \left(\frac{r_<}{r_>^2}\right) \sin\theta.$$

For the *inside solution*, $r_< = r$ and $r_> = a$, so

$$A^{\text{in}}_\phi(\boldsymbol{x}) = \frac{\mu_0}{3} M_0 r \sin\theta$$

You are asked to calculate the corresponding $\boldsymbol{B}$ and $\boldsymbol{H}$ fields for the interior of the sphere in a problem.

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. The following example illustrates.

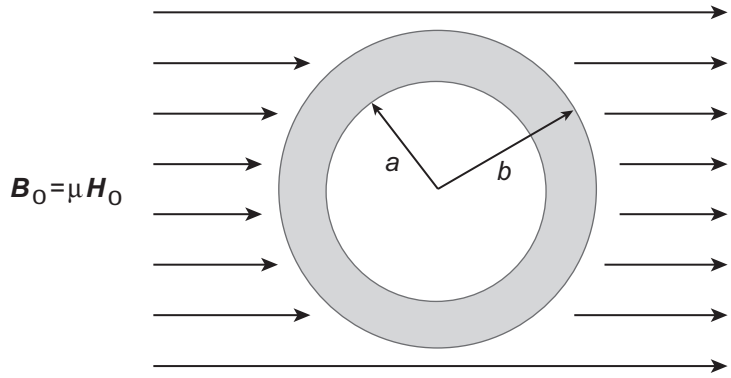


Figure 6.4: A shell of material with permeability μ is placed in a previously uniform magnetic field $\boldsymbol{B}_0 = \mu_0 \boldsymbol{H}_0$. If the shell contains highly permeable material the cavity inside the shell is shielded strongly from the magnetic field.

Shell of Permeable Material

Consider Fig. 6.4 (*shell with permeability* μ).

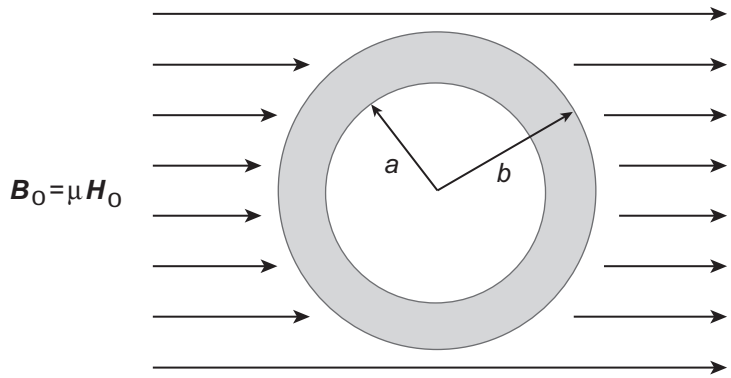
- Let us *find the fields* $\boldsymbol{B}$ and $\boldsymbol{H}$ for this arrangement.
- There are *no currents* so the *magnetic scalar potential method is applicable* and

$$\boldsymbol{H} = -\boldsymbol{\nabla}\Phi_{\mathrm{M}},$$

  and since $\boldsymbol{B} = \mu \boldsymbol{H}$,
- $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \longrightarrow \boldsymbol{\nabla} \cdot \boldsymbol{H} = 0$ in all regions.
- Therefore, $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla}\Phi_{\mathrm{M}}) = \boldsymbol{\nabla} \cdot \boldsymbol{H} = 0$
  and Φ_{M} satisfies the Laplace equation

$$\boldsymbol{\nabla}^2 \Phi_{\mathrm{M}} = 0.$$

- The problem reduces to *solving the Laplace equation* subject to the *proper boundary conditions* at $r = a$ and $r = b$.



- If $r > b$ *(outside the shell)*, the potential is of the form

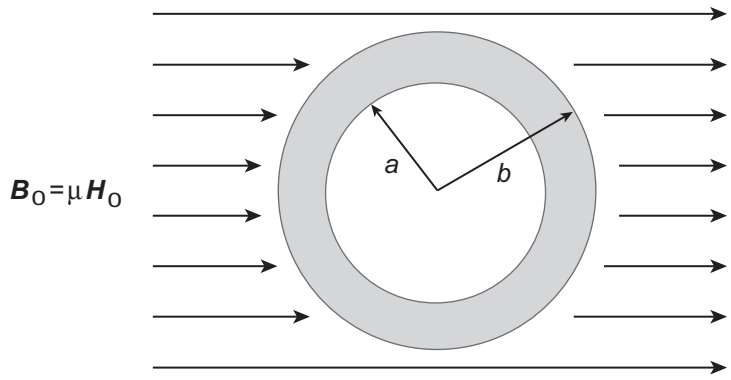
$$\Phi_M = -H_0 r\cos\theta + \sum_{l=0}^{\infty} \frac{\alpha_l}{r^{l+1}} P_l(\cos\theta) \qquad (r > b),$$

  which gives a *uniform field* $\Phi_M = \boldsymbol{H}_0$ *at large distance*.

- Likewise, in the *interior regions* $(r < b)$ the potential must take the form

$$\Phi_M = \sum_{l=0}^{\infty} \left(\beta_l r^l + \gamma_l \frac{1}{r^{l+1}}\right) P_l(\cos\theta) \qquad (a < r < b),$$

$$\Phi_M = \sum_{l=0}^{\infty} \delta_l r^l P_l(\cos\theta) \qquad (r < a).$$



The boundary conditions at $r = a$ and $r = b$

- require the components H_θ and B_r be continuous,
- which requires that

$$\frac{\partial \Phi_M}{\partial \theta}(b_+) = \frac{\partial \Phi_M}{\partial \theta}(b_-) \qquad \frac{\partial \Phi_M}{\partial \theta}(a_+) = \frac{\partial \Phi_M}{\partial \theta}(a_-),$$

$$\mu_0 \frac{\partial \Phi_M}{\partial r}(b_+) = \mu \frac{\partial \Phi_M}{\partial r}(b_-) \qquad \mu \frac{\partial \Phi_M}{\partial r}(a_+) = \mu_0 \frac{\partial \Phi_M}{\partial r}(a_-),$$

- 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

$$\frac{\partial \Phi_{\rm M}}{\partial \theta}(b_+) = \frac{\partial \Phi_{\rm M}}{\partial \theta}(b_-) \qquad \frac{\partial \Phi_{\rm M}}{\partial \theta}(a_+) = \frac{\partial \Phi_{\rm M}}{\partial \theta}(a_-),$$

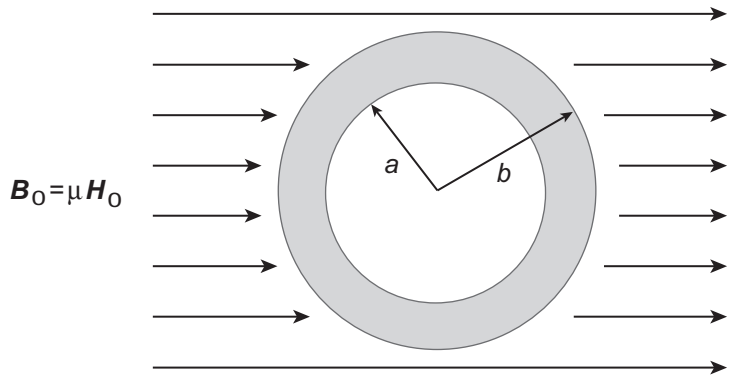
$$\mu_0 \frac{\partial \Phi_{\rm M}}{\partial r}(b_+) = \mu \frac{\partial \Phi_{\rm M}}{\partial r}(b_-) \qquad \mu \frac{\partial \Phi_{\rm M}}{\partial r}(a_+) = \mu_0 \frac{\partial \Phi_{\rm M}}{\partial r}(a_-),$$

determine the unknown constants in

$$\Phi_{\rm M} = -H_0 r \cos\theta + \sum_{l=0}^{\infty} \frac{\alpha_l}{r^{l+1}} P_l(\cos\theta) \qquad (r > b),$$

$$\Phi_{\rm M} = \sum_{l=0}^{\infty} \left(\beta_l r^l + \gamma_l \frac{1}{r^{l+1}} \right) P_l(\cos\theta) \qquad (a < r < b),$$

$$\Phi_{\rm M} = \sum_{l=0}^{\infty} \delta_l r^l P_l(\cos\theta) \qquad (r < a).$$



One finds that

- *all coefficients with* $l \neq 1$ *vanish* and
- *the* $l = 1$ *coefficients satisfy* the simultaneous equations

$$\begin{aligned}
\alpha_1 - b^3\beta_1 - \gamma_1 &= b^3 H_0, \\
2\alpha_1 + \mu' b^3\beta_1 - 2\mu'\gamma_1 &= -b^3 H_0, \\
a^3\beta_1 + \gamma_1 - a^3\delta_1 &= 0, \\
\mu' a^3\beta_1 - 2\mu'\gamma_1 - a^3\delta_1 &= 0,
\end{aligned}$$

where $\mu' \equiv \mu/\mu_0$.

The solutions for α_1 and δ_1 are,

$$\alpha_1 = \left(\frac{(2\mu'+1)(\mu'-1)}{(2\mu'+1)(\mu'+2) - 2\frac{a^3}{b^3}(\mu'-1)^2} \right) (b^3 - a^3) H_0,$$

$$\delta_1 = -\left(\frac{9\mu'}{(2\mu'+1)(\mu'+2) - 2\frac{a^3}{b^3}(\mu'-1)^2} \right) H_0.$$

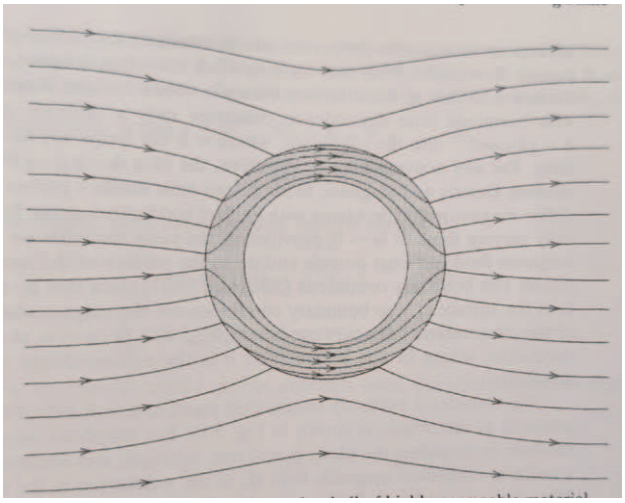
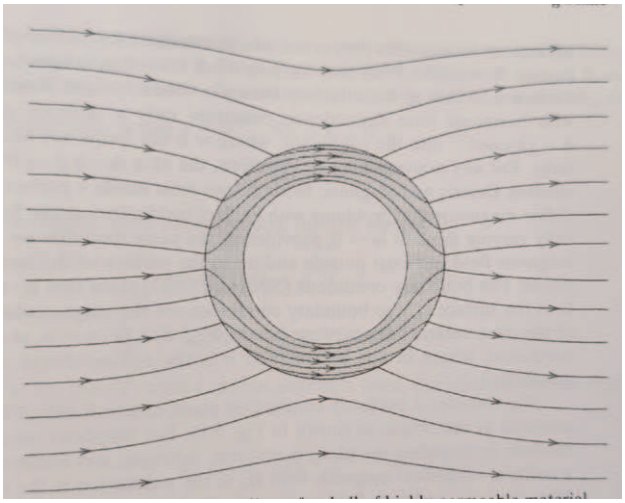


Figure 6.5: The magnetic shielding effect of a shell of highly permeable material. Lines for the magnetic field $\boldsymbol{B}$ are shown. Note the absence of field lines in the cavity inside the shell.

The corresponding *magnetic field lines* are shown in Fig. 6.5.

- For this solution the *potential outside the shell* corresponds to the *original uniform field* $\boldsymbol{H}_0$
- *plus a dipole field* with dipole moment α_1 parallel to $\boldsymbol{H}_0$.
- In the *cavity inside the shell* there is a *uniform field parallel to* $\boldsymbol{H}_0$ and equal in magnitude to $-\delta_1$.
- If $\mu \gg \mu_0$, the dipole moment α_1 and the inner field $-\delta_1$ tend to

$$\alpha_1 \to b^3 H_0 \qquad -\delta_1 \to \frac{9\mu_0}{2\mu(1-a^3/b^3)} H_0.$$



- Thus the *field in the inner cavity is proportional to* μ^{-1} and
- a shield made of highly permeable material causes a *large reduction of the field inside*.

Even *thin shells* of material

- with $\mu/\mu_0 \sim 10^3 - 10^6$
- can *greatly reduce the interior field*,

as is clear from the figure above.

6.3 Faraday and the Law of Induction

We have considered primarily

- *static charges* (no relative motion) as sources of electric fields and
- *steady-state* (no change in time) charge currents as the source of magnetic fields to this point.

However, many important electromagnetic phenomena involve the *motion of charges and/or non-steady electrical currents*.

- *Michael Faraday (1831)* performed quantitative experiments on *time-dependent electric and magnetic fields*.
- *Faraday's essential observations:*
  1. a *transient current* is produced in a test circuit if a *steady current is turned off* in a nearby circuit;
  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*.

SUMMARY: a current flows in the test circuit

- if a the *current in a nearby circuit changes in time*, or
- if the two *circuits move with respect to each* other, or
- a nearby *magnet is moved*.

Faraday interpreted these results as

- originating in a *changing magnetic flux* that
- *produced an electric field* around the test circuit, which leads to an
- *electromotive force* $\mathscr{E}$ (EMF) that *drives a current in the test circuit* governed by *Ohm's law* (see below).

Thus, with Faraday's results and interpretation begins the *unification of electricity and magnetism* in a single theory of *electromagnetism*.

Ohm's Law

To *make a current flow* some force

- must *push on the charges* (continuously if, as is usual, there is any *resistance* to the charge flow).
- For most materials the *current density* $\boldsymbol{J}$ is proportional to the *force per unit charge* $\boldsymbol{f}$,

$$\boldsymbol{J} = \sigma \boldsymbol{f},$$

where the *conductivity* σ depends on the medium.

> Often the *reciprocal of the conductivity*, $\rho = \sigma^{-1}$, which is called the *resistivity*, is used instead.

- If the force is *electromagnetic (Lorentz force)*,

$$\boldsymbol{J} = \sigma(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}),$$

where

  - $\boldsymbol{E}$ is the *electric field*,
  - $\boldsymbol{B}$ is the *magnetic field*, and
  - $\boldsymbol{v}$ is the *velocity*.

Except for special circumstances such as in some plasmas,

- the velocity is small (compared with c) and
- the $\boldsymbol{v} \times \boldsymbol{B}$ term can be neglected, leaving

$$\boldsymbol{J} = \sigma \boldsymbol{E}.$$

- This is called *Ohm's law*.

Ohm's law is *not a true "law"* in the same sense as say the law of gravity.

- It is a *"rule of thumb"* that is often (but not always) obeyed.
- Conductors that obey Ohm's law are called *ohmic conductors*.

Example:

For a cylindrical wire of *cross-sectional area* A and *conductivity* σ,

- if the *potential difference* between the ends is V,
- the *electric field* $\boldsymbol{E}$ *and current density* $\boldsymbol{J}$ *are uniform* and the *total current* is

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

- In this example the total current flow in the wire from one point to another is proportional to the potential difference between the points,

$$V = IR,$$

  where the constant of proportionality R is called the *resistance*.

- This is the *most familiar form of Ohm's law*.
- The resistance depends on the *geometry* and the *conductivity* of the medium;
- in the example given above $R = L/\sigma A$, where L is the length.

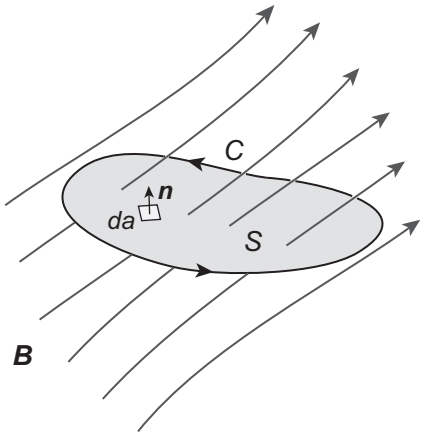


Figure 6.6: Magnetic flux through a circuit C bounding a surface S with a unit normal to the surface $\boldsymbol{n}$ and surface element da. The magnetic field in the neighborhood of the circuit is $\boldsymbol{B}$.

Let us *express Faraday's observations mathematically*.

- Consider *Fig. 6.6*;
- the *magnetic flux* linking the circuit is

$$\mathscr{F} = \int_S \boldsymbol{B} \cdot \boldsymbol{n} da$$

- and the *electromotive force* around the circuit is

$$\mathscr{E} = \oint_C \boldsymbol{E}' \cdot d\boldsymbol{l},$$

- where $\boldsymbol{E}'$ is the electric field at element $d\boldsymbol{l}$ of the circuit C.

Faraday's observations are summarized by the relationship

$$\mathscr{E} = -k\frac{d\mathscr{F}}{dt}.$$

- between the *rate of change of the magnetic flux* $d\mathscr{F}/dt$ and the *EMF* $\mathscr{E}$,
- where k is a constant of proportionality.
- The constant k can be determined by requiring Galilean invariance at low velocities

As we shall see in Ch. 9,

- the *Maxwell equations* are *invariant under Lorentz transformations*, not Galilean transformations.
- But *Lorentz transformations reduce to Galilean transformations* in the limit $v \ll c$.

Thus requiring Galilean invariance for low velocities is a legitimate way to determine k.

This indicates that

- $k = 1$ in SI units (and $k = c^{-1}$ in Gaussian units).
- Therefore, *in SI units*

$$\mathscr{E} = -\frac{d\mathscr{F}}{dt},$$

where the *sign is specified by Lenz's law*.

Lenz's law: The *induced current and magnetic field* are in a direction so as to *oppose changing the flux* through the circuit.

Combining

$$\mathscr{E} = -\frac{d\mathscr{F}}{dt} \qquad \mathscr{F} = \int_S \boldsymbol{B}\cdot\boldsymbol{n}da \qquad \mathscr{E} = \oint_C \boldsymbol{E}'\cdot d\boldsymbol{l},,$$

leads to

$$\oint_C \boldsymbol{E}'\cdot d\boldsymbol{l} = -\frac{d}{dt}\int_S \boldsymbol{B}\cdot\boldsymbol{n}da \qquad \text{(Faraday's law)}$$

indicating that the *induced EMF* is proportional to the *total time derivative* of the flux.

The *flux can be changed* by changing

- the *magnetic field*, or
- the *shape, position, or orientation* of the circuit.
- The *total time derivative* accounts for all such possibilities.

Note that $\boldsymbol{E}'$ is the electric field at $d\boldsymbol{l}$ *in the coordinate system where* $d\boldsymbol{l}$ *is at rest*.

The equation

$$\oint_C \boldsymbol{E}' \cdot d\boldsymbol{l} = -\frac{d}{dt}\int_S \boldsymbol{B}\cdot\boldsymbol{n}\,da \qquad \text{(Faraday's law)}$$

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,
- this equation may be viewed as a *relationship among the fields* $\boldsymbol{B}$ and $\boldsymbol{E}'$.
- If the circuit C is *moving with some velocity*, the total time derivative must take that motion into account, since the *flux through the circuit can change* for two reasons:
  1. the flux is *changing with time at a given point*, or
  2. translation of the circuit *changes location of the circuit* in an external field.

This can be taken into account with the *convective derivative*,

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla},$$

- 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,

$$\frac{d}{dt}\int_S \boldsymbol{B}\cdot\boldsymbol{n}\,da = \int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}\,da + \oint_C (\boldsymbol{B}\times\boldsymbol{v})\cdot d\boldsymbol{l},$$

- which can be used to write Faraday's law in the form

$$\oint_C \left[\boldsymbol{E}' - (\boldsymbol{v}\times\boldsymbol{B})\right]\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}\,da.$$

- This is *Faraday's law applied to the moving circuit* C, but
- if we think of the circuit C and surface S being *instantaneously at a certain point*, then
- application of Faraday's law to that circuit *at fixed location* gives,

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot\boldsymbol{n}\,da,$$

where $\boldsymbol{E}$ is now the *electric field in the laboratory frame*.

Galilean invariance (valid at low velocities) then requires that

- the left sides of the following equations are equivalent

$$\oint_C \left[\boldsymbol{E}' - (\boldsymbol{v}\times\boldsymbol{B})\right]\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot \boldsymbol{n}\, da.$$

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\int_S \frac{\partial \boldsymbol{B}}{\partial t}\cdot \boldsymbol{n}\, da,$$

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 in integral form*

$$\oint_C \boldsymbol{E}'\cdot d\boldsymbol{l} = -\frac{d}{dt}\int_S \boldsymbol{B}\cdot \boldsymbol{n}\, da \qquad \text{(Faraday's law)}$$

  can be *transformed into a differential equation* using *Stoke's theorem*,

$$\oint_C \boldsymbol{E}'\cdot d\boldsymbol{l} = \int_S (\boldsymbol{\nabla}\times\boldsymbol{E}')\cdot \boldsymbol{n}\, da.$$

- Then the *integral form of Faraday's law* may be written as

$$\int_S \left(\boldsymbol{\nabla}\times\boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t}\right)\cdot \boldsymbol{n}\, da = 0.$$

- But the circuit C and surface S are *arbitrary*, so
- the *integrand must vanish identically*, giving

$$\boldsymbol{\nabla}\times\boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0,$$

  which is *Faraday's law in differential form*.

Faraday's law in differential form,

$$\boldsymbol{\nabla} \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0,$$

is the *time-dependent generalization* of the equation $\boldsymbol{\nabla} \times \boldsymbol{E} = 0$ from electrostatics.

Chapter 7

Maxwell's Equations

Preceding chapters have provided a systematic understanding and *validation of the various pieces of electromagnetic theory* that are synthesized into the four concise *Maxwell equations*.

- Up to this point we have treated electricity and magnetism *largely as separate subjects*.
- As *Faraday's discovery of induction* that was discussed earlier makes clear,
- this distinction begins to fail for time-dependent phenomena, and
- we will now begin to address a unified picture of electromagnetism.
- In this chapter we take the Maxwell equations to be the basis for all classical understanding of electromagnetism and
- address systematically their broader scientific and technical implications.

7.1 The Almost-but-Not-Quite Maxwell's Equations

The basic equations of electricity and magnetism that we have studied in Chs. 1-6 can be summarized (in medium, in differential form) as

$$\begin{aligned} \nabla \cdot \boldsymbol{D} &= \rho && \text{(Gauss's law)},\\ \nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\ \nabla \cdot \boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\ \nabla \times \boldsymbol{H} &= \boldsymbol{J} && \text{(Ampère's law)}, \end{aligned}$$

- In modern notation, these were the fundamental equations of electromagnetism as they stood in the mid-1800s
- when James Clerk Maxwell set about his synthesis of electromagnetic understanding.
- A comparison shows that these are *almost, but not quite, the Maxwell equations* (in medium, in differential form).

$$\begin{aligned} \nabla \cdot \boldsymbol{D} &= \rho && \text{(Gauss's law)},\\ \nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\ \nabla \cdot \boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\ \nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)}, \end{aligned}$$

The difference lies in the fourth equation *(Ampère's law)*, where a term $-\partial \boldsymbol{D}/\partial t$ is missing.

7.1.1 Ampère's Law and the Displacement Current

It is important to recognize that

- all but Faraday's law were derived from data taken under steady-state conditions, so
- we should expect that modifications might be required in the face of data for time-dependent fields.

Maxwell first realized that the fundamental equations of electromagnetism in his time were inconsistent as they then stood.

- Since the divergence of a curl vanishes by a basic vector-calculus identity, $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = 0$,
- it follows from Ampère's law

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J}$$

for steady currents that

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = \boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{H}) = 0.$$

- But from the continuity equation ensuring conservation of charge gives,

$$\frac{\partial \rho}{\partial t} = -\boldsymbol{\nabla} \cdot \boldsymbol{J},$$

so $\boldsymbol{\nabla} \cdot \boldsymbol{J} = 0$ can hold *only if the charge density doesn't change with time*.

Maxwell *modified Ampère's law* to accomodate a time dependence by introducing a *displacement current* term $\partial \boldsymbol{D}/\partial t$,

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J} \quad \longrightarrow \boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t},$$

thus leading to what we now call the Maxwell equations (in medium, in differential form),

$$\begin{aligned}
\boldsymbol{\nabla} \cdot \boldsymbol{D} &= \rho && \text{(Gauss's law)},\\
\boldsymbol{\nabla} \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= 0 && \text{(Faraday's law)},\\
\boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0 && \text{(No magnetic charges)},\\
\boldsymbol{\nabla} \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} &= \boldsymbol{J} && \text{(Ampère–Maxwell law)},
\end{aligned}$$

where the last equation is now called the *Ampère–Maxwell law*.

- It still is the same law as before when applied to steady-state phenomena,
- but addition of the displacement current term means that a *changing electric field can generate a magnetic field*, even if there is no current.
- Thus, it is the *converse of Faraday's law*, where a changing magnetic field can produce an electric field.

These *(Maxwell) equations* are now *consistent with the continuity equation.*

- Taking the *divergence of both sides of the Ampère–Maxwell law*,

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial(\nabla \cdot \boldsymbol{D})}{\partial t} \;\rightarrow\; \nabla \cdot (\nabla \times \boldsymbol{H}) = \nabla \cdot \boldsymbol{J} + \frac{\partial(\nabla \cdot \boldsymbol{D})}{\partial t},$$

  and using $\nabla \cdot \boldsymbol{D} = \rho$ from *Gauss's law* and the *identity* $\nabla \cdot (\nabla \times \boldsymbol{B}) = 0$,

- this becomes

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{j} = 0,$$

  which is the *continuity equation*.

Local conservation of charge: Physically the continuity equation requires that

- changes in electrical charge in some arbitrary volume are caused by flow of charge through the surface bounding that volume.
- This requires conservation of charge within a volume of space that can be arbitrarily small,
- which implies that charge is *conserved locally*.

> A charge that disappears from one point in space and instantly reappears at another point is consistent with *global charge conservation*, but not with *local charge conservation*.

- The reason requires relativistic quantum field theory for its full explanation:
- destroying a charge at one point and simultaneously creating it at another would require
- *instantaneous propagation of a signal* between the two points,
- which is *inconsistent with special relativity*.

Maxwell's equations

- supplemented by the *Lorentz force law*

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}),$$

  to describe *electromagnetic forces*, and

- *Newton's laws of motion* to translate force into particle motion

are thought to describe all of classical electromagnetism.

7.1.2 Implications of the Ampère–Maxwell Law

Maxwell's seemingly small change in the Ampère law turns out to have enormous implications for both our classical and quantum understanding of electromagnetism.

- Maxwell's addition of $\partial \boldsymbol{D}/\partial t$ to Ampère's law (and Faraday's induction experiments) effectively brought together the previously separate subjects of electricity and magnetism.
- Ampère's law is only about magnetism, but *both* the polarized electric field $\boldsymbol{D}$ and the polarized magnetic field $\boldsymbol{H}$ appear in the Ampère–Maxwell law,
- while $\boldsymbol{E}$ and $\boldsymbol{B}$ both appear in *Faraday's law*.
- *Changing electric fields produce magnetic fields and changing magnetic fields produce electric fields*, and we may now speak of the unified subject of *electromagnetism*.
- The fundamental equations of electromagnetism are now consistent with *(local) conservation of electrical charge*.
- This modification will lead eventually to the greatest triumph of the classical Maxwell equations:
  - the realization that the *Maxwell equations have a wave solution*, and that
  - the resulting *electromagnetic waves may be interpreted as light*,
  - thus *unifying electricity and magnetism with optics*.

- That Maxwell's equations obey the continuity equation and thus conserve charge locally will lead to the idea of classical *electromagnetic gauge invariance*,
- which will underlie a quantum field theory of electromagnetism (*quantum electrodynamics* or QED).
- Electromagnetic gauge invariance will eventually be generalized to *more sophisticated local gauge invariance* in the *weak and strong interactions*,
- resulting in the relativistic quantum field theory that we term the *Standard Model* of elementary particle physics.

Some of these topics are beyond the scope of classical electromagnetism, but have their historical and scientific antecedents in that discipline.

7.2 Vector and Scalar Potentials

In our discussions of electrostatics and magnetostatics we introduced the scalar potential Φ and the vector potential $\boldsymbol{A}$.

- Maxwell's equations 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 the Maxwell equations.

$$\nabla \cdot \boldsymbol{D} = \rho \qquad \text{(Gauss's law)},$$
$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)},$$
$$\nabla \cdot \boldsymbol{B} = 0 \qquad \text{(No magnetic charges)},$$
$$\nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)},$$

- Since $\nabla \cdot \boldsymbol{B} = 0$, the magnetic field $\boldsymbol{B}$ can be described as the curl of a vector potential $\boldsymbol{A}$,

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}.$$

Recall the reason:

- if $\boldsymbol{B} = \nabla \times \boldsymbol{A}$,
- then by *taking the divergence* of both sides $\nabla \cdot \boldsymbol{B} = \nabla \cdot (\nabla \times \boldsymbol{A}) = 0$.

Thus the condition $\nabla \cdot \boldsymbol{B} = 0$ is guaranteed if $\boldsymbol{B}$ *derives from the curl of a vector potential*.

Then *Faraday's law* may be written as,

$$\boldsymbol{\nabla} \times \left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} \right) = 0.$$

- Since the *curl of* $\boldsymbol{E} + \partial \boldsymbol{A}/\partial t$ *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

$$\boldsymbol{E} = -\boldsymbol{\nabla}\Phi - \frac{\partial \boldsymbol{A}}{\partial t}.$$

Introducing the substitutions

$$\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A} \qquad \boldsymbol{E} = -\boldsymbol{\nabla}\Phi - \partial \boldsymbol{A}/\partial t$$

into the Maxwell equations, we find using using the identities

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = 0 \qquad \boldsymbol{\nabla} \times \boldsymbol{\nabla}\Phi = 0$$

that the homogeneous equations

$$\boldsymbol{\nabla} \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad \text{(Faraday's law)},$$
$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \qquad \text{(No magnetic charges)},$$

are *satisfied identically*, while the inhomogeneous equations

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho \qquad \text{(Gauss's law)},$$
$$\boldsymbol{\nabla} \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J} \qquad \text{(Ampère–Maxwell law)},$$

are transformed into the coupled second-order differential equations,

$$\boldsymbol{\nabla}^2\Phi + \frac{\partial}{\partial t}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) = -\frac{\rho}{\varepsilon_0},$$
$$\boldsymbol{\nabla}^2 \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t}\right) = -\mu_0 \boldsymbol{J},$$

where $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \boldsymbol{\nabla}^2 \boldsymbol{A}$ has been used.

- Thus the problem has been converted into *solving two coupled second-order differential equations*.
- Let us now show that these equations can be *decoupled by a suitable gauge transformation*.

7.2.1 Exploiting Gauge Symmetry

Earlier we showed that

- the magnetic field is invariant under a gauge transformation on the vector potential,

$$\boldsymbol{A} \rightarrow \boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\chi,$$

  where χ is an arbitrary scalar function.

- If the electric field to be unaltered under this transformation the scalar potential must be changed at the same time according to

$$\Phi \rightarrow \Phi' = \Phi - \frac{\partial \chi}{\partial t}.$$

- Thus, a classical gauge transformation on the electromagnetic field is defined by the simultaneous transformations

$$\boldsymbol{A} \rightarrow \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \rightarrow \Phi - \frac{\partial \chi}{\partial t},$$

  on the vector potential $\boldsymbol{A}$ and the scalar potential Φ, for an arbitrary scalar function χ.

Suppose that we now exploit the invariance of electromagnetism under such gauge transformations and choose a set of potentials $\{\Phi, \boldsymbol{A}\}$ that satisfy the *Lorenz condition*

$$\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t} = 0.$$

A constraint like

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}+\frac{1}{c^2}\frac{\partial\Phi}{\partial t}=0.$$

- is termed a *gauge-fixing condition* and
- imposing such a constraint is termed *fixing the gauge*.
- The gauge choice implied by the equation above is called the *Lorenz gauge*.

> The *Lorenz gauge* (named for Ludvig Lorenz) is often mistakenly called the *Lorentz gauge* (after the more famous Hendrik Lorentz).

- If the Lorenz gauge condition

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}+\frac{1}{c^2}\frac{\partial\Phi}{\partial t}=0.$$

  is inserted into the coupled Maxwell equations

$$\boldsymbol{\nabla}^2\Phi+\frac{\partial}{\partial t}(\boldsymbol{\nabla}\cdot\boldsymbol{A})=-\frac{\rho}{\varepsilon_0},$$
$$\boldsymbol{\nabla}^2\boldsymbol{A}-\frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2}-\boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\boldsymbol{A}+\frac{1}{c^2}\frac{\partial\Phi}{\partial t}\right)=-\mu_0\boldsymbol{J},$$

- the equations decouple and
- solving the Maxwell equations has now been reduced to solving two second-order differential equations,

$$\boldsymbol{\nabla}^2\Phi-\frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2}=-\frac{\rho}{\varepsilon_0},$$
$$\boldsymbol{\nabla}^2\boldsymbol{A}-\frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2}=-\mu_0\boldsymbol{J},$$

  which are *uncoupled:*

- solution of the first equation gives the scalar potential Φ independent of the vector potential $\boldsymbol{A}$,
- while solution of the second equation gives the vector potential, independent of the scalar potential.

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 the Maxwell equations, but does not satisfy the gauge condition.
- 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 gauge scalar function χ can be found that satisfies

$$\nabla^2\chi - \frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2} = -\left(\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t}\right),$$

the new potentials will satisfy the Lorenz gauge conditions and the simultaneous Maxwell equations.

The Lorenz condition

$$\boldsymbol{\nabla}\cdot\boldsymbol{A}+\frac{1}{c^2}\frac{\partial\Phi}{\partial t}=0.$$

- does not exhaust the gauge degrees of freedom in Lorenz gauge.
- The *restricted gauge transformation*

$$\boldsymbol{A}\to\boldsymbol{A}+\boldsymbol{\nabla}\chi \qquad \Phi\to\Phi-\frac{\partial\chi}{\partial t},$$

where the scalar function χ satisfies

$$\boldsymbol{\nabla}^2\chi-\frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2}=0,$$

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. The Lorenz gauge is often used for two reasons:

1. It leads to the decoupled Maxwell equations that treat Φ and $\boldsymbol{A}$ on an equal footing.
2. As will be elaborated in Ch. 8, the Lorenz gauge condition is *invariant under Lorentz transformations*, which fits naturally into special relativity.

> Notice: the equations are in (Ludvig) *Lorenz gauge*, but we shall see that they are invariant under (Hendrik) *Lorentz transformations*.

There are infinitely many valid gauge transformations that can be made using

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \Phi \to \Phi - \frac{\partial \chi}{\partial t},$$

with *different choices for the scalar function* χ.

- However, only some prove to be useful.
- One of use is the transformation to *Lorenz* gauge described in this section.
- Another that we have already encountered in Section 5.5 is the transformation to *Coulomb gauge*.

7.2.2 Coulomb Gauge

We have already introduced the *Coulomb gauge condition*

$$\nabla \cdot \boldsymbol{A} = 0 \qquad \text{(Coulomb gauge).}$$

From

$$\nabla^2 \Phi + \frac{\partial}{\partial t}(\nabla \cdot \boldsymbol{A}) = -\frac{\rho}{\varepsilon_0},$$
$$\nabla^2 \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \nabla\left(\nabla \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t}\right) = -\mu_0 \boldsymbol{J},$$

in Coulomb gauge the *scalar potential obeys a Poisson equation*

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

which has a solution

$$\Phi(\boldsymbol{x}, t) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}', t)}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x'.$$

This is the *instantaneous Coulomb potential* caused by the charge density $\rho(\boldsymbol{x})$, which is the source of the name *Coulomb gauge*.

Causality and Coulomb and Gravitational Potentials

The Coulomb potential

$$\Phi(\boldsymbol{x},t) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'.$$

- suggests instantaneous transmission of information in a universe where
- lightspeed c is the speed limit.

> *Causality* (that *cause precedes effects*) is a fundamental principle underlying all of modern science. The Coulomb potential appears to *violate that principle*.

Causality and the Coulomb Potential

The Coulomb potential is said to be *"instantaneous"* because

- the associated force acts without delay at any $\boldsymbol{x}$ corresponding a distance $|\boldsymbol{x}-\boldsymbol{x}'|$ from the source.
- This is *inconsistent with special relativity* and relativistic quantum field theory,
- which require that a force be transmitted by a virtual particle (the photon in this case)
- communicated at a speed less than or equal to the speed of light c.

Causality and the Newtonian Gravitation Potential

The Newtonianian gravitational potential

- has the form of a Coulomb potential with *masses playing the role of charges*, and
- has the same causality problem:

  it implies that the gravitational force acts instantaneously over any distance.

- The solution of this problem in gravitational physics is replacement of Newtonian gravity with general relativity,
- which generalizes special relativity and requires that the transmission speed of the gravitational force be equal to the speed of light.

This prediction of general relativity has been confirmed by observations, as we now describe.

The Speed of Light and the Speed of Gravity

In 2017 the ground-based Ligo–Virgo detectors observed gravitational wave GW170817. But there was more to come:

- 1.7 seconds later the Fermi Gamma-ray Space Telescope (Fermi) and International Gamma-Ray Astrophysics Laboratory (INTEGRAL) in orbit around Earth
- detected a gamma-ray burst from the same portion of the sky as the source of the gravitational wave.
- Detailed observations of the afterglow of the gamma-ray burst at many wavelengths, and
- comprehensive analysis of the gravitational and electromagnetic data sets,
- concluded that the gravitational wave and the gamma-ray burst were caused by a binary neutron star merger in the galaxy NGC 4993,
- at a distance of 40 megaparsecs (130 million lightyears).

- That the gravitational and electromagnetic signals arrived *within 1.7 seconds of each other* after traveling *40 mega-parsecs* implied that
- the *speed of gravity* (for the gravitational waves) and the speed of light c (for the gamma-rays) differ by *at most 3 parts in* 10^{15}.
- Thus *the speed of gravity is* c, just as predicted by the general theory of relativity.

We will discuss briefly the description of gravitational waves in linearized gravity and the similarity with the Maxwell equations in Ch. 8.

In Coulomb gauge the vector potential obeys

$$\nabla^2 \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\mu_0 \boldsymbol{J} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t}.$$

- 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*

$$\boldsymbol{J} = \boldsymbol{J}_{\mathrm{L}} + \boldsymbol{J}_{\mathrm{T}},$$

- where the terms have the following properties:
  1. The component $\boldsymbol{J}_{\mathrm{L}}$ has vanishing curl, $\nabla \times \boldsymbol{J}_{\mathrm{L}} = 0$; it is called the *longitudinal current* or the *irrotational current*.
  2. The component $\boldsymbol{J}_{\mathrm{T}}$ has vanishing divergence, $\nabla \cdot \boldsymbol{J}_{\mathrm{T}} = 0$; it is called the *transverse current* or the *solenoidal current*.

The *longitudinal current* and *transverse current* are explicitly

$$\boldsymbol{J}_{\mathrm{L}} = -\frac{1}{4\pi}\boldsymbol{\nabla}\int \frac{\boldsymbol{\nabla}'\cdot\boldsymbol{J}(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x',$$

$$\boldsymbol{J}_{\mathrm{T}} = \frac{1}{4\pi}\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\int \frac{\boldsymbol{J}(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

Combining the expression for the scalar potential,

$$\Phi(\boldsymbol{x},t) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{x}',t)}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

with the continuity equation, $\partial\rho/\partial t + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0$ leads to

$$\frac{1}{c^2}\boldsymbol{\nabla}\frac{\partial\Phi}{\partial t} = \mu_0\boldsymbol{J}_{\mathrm{T}},$$

which gives when inserted into

$$\boldsymbol{\nabla}^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t}\right) = -\mu_0\boldsymbol{J},$$

and using the identity $\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) = 0$ and $\boldsymbol{J} = \boldsymbol{J}_{\mathrm{L}} + \boldsymbol{J}_{\mathrm{T}}$,

$$\boldsymbol{\nabla}^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J}_{\mathrm{T}}.$$

Thus, *in Coulomb gauge* the source for the $\boldsymbol{A}$ equation can be expressed *entirely in terms of the transverse current* $\boldsymbol{J}_{\mathrm{T}}$.

- Because the vector potential is determined entirely by the transverse current, the Coulomb gauge is sometimes termed the *transverse gauge*.
- The Coulomb gauge is also sometimes called the *radiation gauge*,
- because one finds in quantum electrodynamics that only the vector potential (which is determined by the transverse components) need be quantized.

Notice that

$$\nabla^2 \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\mu_0 \boldsymbol{J}_{\mathrm{T}}.$$

has the form of a *wave equation* with the *speed of the wave equal to* c.

- This is the behavior *expected for electromagnetic waves*.
- However, the discussion above implies that for the *scalar potential* the propagation speed is *infinite*.

The resolution of this seeming paradox requires *relativistic quantum field theory*, but in essence

- the propagating classical field has *only transverse components* and one finds that
- in QED *only the vector potential* (and thus *only the transverse components*) need be quantized.

As we will see in Ch. 8,

- Lorentz covariance requires the 3-vector potential $\boldsymbol{A}$ and the scalar potential Φ
- to be combined into a spacetime 4-vector A^μ with the components of the 4-vector given by

$$A^\mu = (A^0, A^1, A^2, A^3) = (\Phi, \boldsymbol{A}) = (\Phi, A^1, A^2, A^3),$$

  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.

A small *terminology problem* now enters our discussion.

- We have been calling $\boldsymbol{A}$ the vector potential, but
- relativistically it will be more convenient to work with the 4-vector potential A^μ,
- which makes more obvious the requirement of relativity that space and time enter on an equal footing.
- Thus beginning in the next chapter we will be calling the 4-vector A^μ the vector potential.

We adopt a policy that

- where it is clear that we are working non-relativistically we will call $\boldsymbol{A}$ the vector potential, while
- if it is clear that we are working in a relativistic context we will call A^μ the vector potential.
- If there is any chance for confusion we use the explicit names *"3-vector potential"* for $\boldsymbol{A}$ and *"4-vector potential"* for A^μ.

We have asserted above that only the *two transverse components or polarizations* of the vector are required to describe propagating EM waves.

- But a 3-vector like $\boldsymbol{A}$ normally has *three components*, and
- a 4-vector like A^μ has *four components*.

So how can a propagating photon have only *two rather than three or four* degrees of freedom?

- *Relativistic quantum field theory* for electrons and photons (QED) is *beyond our present scope* in these lectures.
- However, the answer is that for QED in a *covariant gauge* like *Lorenz gauge* there are *four states of polarization*,
- but the contributions from *timelike* and *longitudinal* polarizations have *equal magnitudes* but *opposite signs* and
- exactly *cancel each other*,
- leaving *only the transverse polarizations* for a free propagating photon. (This is called the *Gupta–Bleuler mechanism* in QED.)
- A *massive vector field* would have *three spatial polarization components*.
- As we will see, the *reduction to two spatial polarizations* is because the *photon is identically massless:*

> *Massless vector fields* have *two rather than three* states of polarization.

7.3 Retarded Green Function

From the *Maxwell equations* in *Lorenz gauge*,

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = -\frac{\rho}{\varepsilon_0},$$
$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J},$$

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,

$$\Box\psi = -f,$$

where we now introduce for convenience the *d'Alembertian operator* $\Box$ with

$$\Box \equiv -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2.$$

The *d'Alembertian operator* $\Box$ is

- a *Lorentz-invariant* combination of the second derivatives with respect to space and time
- that will play a significant role in discussing the *relationship of the Maxwell equations to special relativity* in Ch. 8.

Since we are *assuming Lorenz gauge*, we must also satisfy

$$\nabla \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \Phi}{\partial t} = 0 \qquad \text{(Lorenz gauge condition)}$$

- But the *retarded solutions* of

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -\frac{\rho}{\varepsilon_0},$$
$$\nabla^2 \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\mu_0 \boldsymbol{J},$$

  with sources that we will find shortly will in fact

- *satisfy the Lorenz gauge condition automatically*, provided that they decrease rapidly enough at infinity.

Generalizing the electrostatics case, a *Green function* $G(t,\boldsymbol{x};t'\boldsymbol{x}')$ can be defined by

$$\Box\, G(t,\boldsymbol{x};t'\boldsymbol{x}') = -\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t-t'),$$

- where the derivative operators in $\Box$ are understood to *operate on the unprimed variables*,
- 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 $\Box\psi = -f$ is given by

$$\psi(t,\boldsymbol{x}) = \int G(t,\boldsymbol{x};t'\boldsymbol{x}')\, f(t',\boldsymbol{x}')\, d^3x'\, dt',$$

if the integral converges.

Let us seek a solution of

$$\Box G(t,\boldsymbol{x};t'\boldsymbol{x}') = -\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t-t'),$$

using *Fourier transforms*.

- For an integrable function $F:\mathbb{R}\to\mathbb{R}$, define its Fourier transform $\hat{F}$ as

$$\hat{F}(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} F(x)e^{-ikx}dx.$$

- Fourier transforms can be *extended to distributions* such as the Dirac delta function.
- For example, the *Fourier transform of a delta function* is,

$$\hat{\delta}_{x_0} = \frac{1}{\sqrt{2\pi}}e^{ilx_0}.$$

Assuming a smooth function falling off fast enough at infinity,

- the original function $F(x)$ can be recovered by the *inverse Fourier transform*,

$$F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \hat{F}(k)e^{+ikx}dk.$$

- This formula *applies also to distributions* and the *inverse Fourier transform for a delta function* is

$$\delta_{x_0}(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \hat{\delta}_{x_0}(k)e^{+ikx}dk = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ikx_0}e^{+ikx}\,dk.$$

An important property of Fourier transforms is that

- differentiation in real space corresponds to multiplication by ik in Fourier transform space.
- For example, as you are asked to show in a problem,

$$\widehat{\frac{dF}{dx}}(k) = ik\hat{F}(k),$$

for the Fourier transform of $dF(x)/dt$.

Thus any partial differential equation with constant coefficients in real space can be converted to an algebraic equation in Fourier transform space.

Now let's try to solve

$$\Box\, G(t,\boldsymbol{x};t'\boldsymbol{x}') = -\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t-t'),$$

for G using Fourier transforms.

- To simplify notation we temporarily set $x' = t' = 0$ and $c = 1$, and
- define the 4D Fourier transform of G as

$$\hat{G}(\omega,\boldsymbol{k}) = \frac{1}{(2\pi)^2}\int_{-\infty}^{+\infty} G(t,\boldsymbol{x})e^{+i\omega t}e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}dt\,d^3x.$$

Note: 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 (Ch. 8).

Taking the Fourier transform of

$$\Box G(t,\boldsymbol{x};t'\boldsymbol{x}') = -\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t-t'),$$

with respect to t and $\boldsymbol{x}$, using

$$\hat{\delta}_{x_0} = \frac{1}{\sqrt{2\pi}} e^{ilx_0}.$$

with $x_0 = 0$, and using

$$\widehat{\frac{dF}{dx}}(k) = ik\hat{F}(k),$$

yields

$$(\omega^2 - k^2)\,\hat{G}(\omega,\boldsymbol{k}) = -\frac{1}{4\pi^2},$$

where $k \equiv |k|$. Naively, this suggest the solution

$$\hat{G}(\omega,\boldsymbol{k}) = -\frac{1}{4\pi^2}\frac{1}{(\omega^2-k^2)} = -\frac{1}{4\pi^2}\frac{1}{(\omega+k)(\omega-k)},$$

but division by $(\omega^2 - k^2)$ is illegal since ω could be equal to 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 $\tilde{G}$,

$$\begin{aligned}\tilde{G}(t,\boldsymbol{k}) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\hat{G}(\omega,\boldsymbol{k})e^{-i\omega t}\,d\omega. \\ &= -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k)(\omega-k)}d\omega,\end{aligned}$$

  where

$$\hat{G}(\omega,\boldsymbol{k}) = -\frac{1}{4\pi^2}\frac{1}{(\omega^2-k^2)} = -\frac{1}{4\pi^2}\frac{1}{(\omega+k)(\omega-k)},$$

  was used.

This has *logarithmic diverges* of the integral if $\omega \to k$.

> The simplest logarithmic divergence occurs in an integral of the form
>
> $$f(x) = \int_{x_0}^{x}\frac{1}{\Lambda}d\Lambda.$$
>
> Such an integral diverges as $x \to \infty$, but rather mildly as $f(x) \sim \log(x)$.

Thus it is *ill-defined*.

The basic reason for the ambiguity in

$$\tilde{G}(t,\boldsymbol{k}) = -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k)(\omega-k)}\,d\omega,$$

is that

- many different Green functions satisfy

  $$\Box\, G(t,\boldsymbol{x};t'\boldsymbol{x}') = -\delta(\boldsymbol{x}-\boldsymbol{x}')\delta(t-t'),$$

  so we cannot expect to solve for G
- without providing further information to identify the Green function that we are after.

To proceed it is necessary to *regularize the integration* in such a way that

1. $(\omega^2-k^2)\,\hat{G}(\omega,\boldsymbol{k}) = -\dfrac{1}{4\pi^2}$ remains valid, and
2. the right side of

   $$\tilde{G}(t,\boldsymbol{k}) = -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k)(\omega-k)}\,d\omega,$$

   becomes well defined.

One way to do this is to

- displace the poles at $\omega = \pm k$ into the complex ω plane
- and evaluate the resulting *contour integral*.

There is more than one way to do this, each leading to a different Green function.

- The Green function that we seek is the *retarded Green function*,
- which will be appropriate for the situation where there is *only outgoing radiation* from a source.
- To get the retarded Green function the poles should be displaced into the lower half of the complex ω-plane;
- thus we define the retarded Green function by

$$\tilde{G}(t,\boldsymbol{k})_{\text{ret}} = -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k+i\varepsilon)(\omega-k+i\varepsilon)}\,d\omega,$$

where ε is positive and the limit $\varepsilon \to 0$ is to be taken after the integral is evaluated.

Viewing the integral

$$\tilde{G}(t,\boldsymbol{k})_{\text{ret}} = -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega+k+i\varepsilon)(\omega-k+i\varepsilon)}d\omega,$$

as a contour integral in the complex ω-plane,

- if $t<0$ the exponential is damped in the upper half-plane of ω and
- the countour integral can be *closed in the upper half plane*.
- Since the poles are in the negative half-plane, the contour does not enclose them and
- by the Cauchy theorem,

$$\tilde{G}(t,\boldsymbol{k})_{\text{ret}} = 0 \qquad (t<0).$$

- This vanishing of the Green function before $t=0$
- is *characteristic of the retarded solution* and corresponds to a solution with *no incoming radiation*.
- This solution is physically relevant when there is *no initial radiation* when the source is turned on.

By similar reasoning,

- for $t > 0$ the contour can be closed in the lower half-plane.
- Now the contour encloses poles at $\omega = \pm k - i\varepsilon$ and
- by the *Cauchy theorem* the integral evaluates to $2\pi i$ *times a sum of residues at the poles*.

> The residue $\operatorname{res}[f(a)]$ of a function $f(z)$ at a pole $z = a$ is given by
>
> $$\operatorname{res}[f(a)] = \frac{1}{(m-1)!}\lim_{z\to 0}\left(\frac{d^{m-1}}{dz^{m-1}}(z-a)^m f(z)\right),$$
>
> where m is the order of the pole and a is the location of the pole ($a \neq \infty$)

Then from

$$\tilde{G}(t,\boldsymbol{k})_{\text{ret}} = -\frac{1}{4\pi^2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}}{(\omega + k + i\varepsilon)(\omega - k + i\varepsilon)}\,d\omega,$$

we obtain

$$\begin{aligned}\tilde{G}(t,\boldsymbol{k})_{\text{ret}} &= \frac{2\pi i}{4\pi^2\sqrt{2\pi}}\left(\frac{e^{-ikt}}{2k} - \frac{e^{+ikt}}{2k}\right)\\ &= \frac{1}{2\pi\sqrt{2\pi}}\frac{\sin kt}{k} \qquad (t > 0).\end{aligned}$$

The retarded Green function in real (position) space then results from taking the inverse transform of this result, which yields (homework problem),

$$G(t,\boldsymbol{x})_{\text{ret}} = \frac{\delta(t-|\boldsymbol{x}|)}{4\pi\,|\boldsymbol{x}|} \qquad (t>0).$$

Restoring t', $\boldsymbol{x}'$, and c, this result becomes

$$G(t,\boldsymbol{x};t',\boldsymbol{x}')_{\text{ret}} = \begin{cases} 0 & (t<t'), \\ \dfrac{\delta(t-t'-|\boldsymbol{x}-\boldsymbol{x}'|/c)}{4\pi\,|\boldsymbol{x}-\boldsymbol{x}'|} & (t>t'). \end{cases}$$

This propagator now has the expected causal behavior.

- Using special relativity language, it is nonvanishing only on the *future lightcone* of the source point $(t',\boldsymbol{x}')$,
- meaning that it is non-vanishing only if $|\boldsymbol{x}-\boldsymbol{x}'| = c(t-t')$.
- That is, if a field satisfying the wave equation $\Box\psi = -f$, vanishes at early times, and
- a δ-function source is placed at $(t',\boldsymbol{x}')$,
- the resulting disturbance of the field will *propagate away from the source* at exactly *the speed of light*.

The *retarded solution* of $\Box\psi = -f$ is the solution obtained using the *retarded Green function* in,

$$\psi(t,\boldsymbol{x}) = \int G(t,\boldsymbol{x};t'\boldsymbol{x}')f(t',\boldsymbol{x}')\,d^3x'\,dt',$$

which gives

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t-|\boldsymbol{x}-\boldsymbol{x}'|/c,\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x'.$$

This can also be written more compactly as

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t',\boldsymbol{x}')_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',$$

where the notation means that $f(t',\boldsymbol{x}')$ *is to be evaluated at the retarded time*

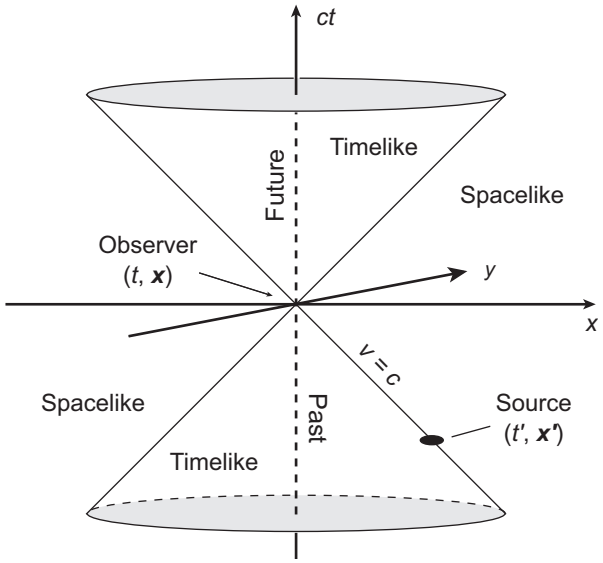
$$t' = t - \frac{|\boldsymbol{x}-\boldsymbol{x}'|}{c},$$

and where t' isn't independent but rather is a function of t, $\boldsymbol{x}$, and $\boldsymbol{x}'$.

In the equation

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t',\boldsymbol{x}')_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',$$

- the integration is *not over all of space* at an instant of time (as the notation might suggest).
- Rather is restricted to the *past lightcone of the spacetime point* $(t,\boldsymbol{x})$,
- since *only points on the past lightcone* can be *causally connected* to $(t,\boldsymbol{x})$ by a signal traveling at lightspeed.



Written in the form

$$\psi(t,\boldsymbol{x}) = \frac{1}{4\pi}\int \frac{f(t',\boldsymbol{x}')_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}\,d^3x',$$

- the retarded solution looks like a *solution of the Poisson equation*, but
- evaluated at the *retarded time*

$$t' = t - \frac{|\boldsymbol{x}-\boldsymbol{x}'|}{c},$$

The Maxwell equations

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = -\frac{\rho}{\varepsilon_0},$$
$$\nabla^2\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\mu_0\boldsymbol{J},$$

for the scalar and vector potentials are of the form,

$$\Box\psi = -f,$$

so we can write immediately the corresponding retarded solutions,

$$\Phi(t,\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{[\rho(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x',$$
$$\boldsymbol{A}(t,\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{J}(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

for the scalar and vector potentials.

Finally, substitution of the solutions

$$\Phi(t,\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{[\rho(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x',$$
$$\boldsymbol{A}(t,\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{J}(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

in the Lorenz gauge condition,

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{1}{c^2}\frac{\partial\Phi}{\partial t} = 0 \qquad \text{(Lorenz gauge condition)}$$

and some manipulation indicates that these solutions are indeed *consistent with the Lorenz gauge condition*. Therefore,

- we may conclude that the *solution of the Maxwell equations* for a charge density ρ and a current density $\boldsymbol{J}$,
- with initial conditions of *no incoming radiation*,
- is given by

$$\Phi(t,\boldsymbol{x}) = \frac{1}{4\pi\varepsilon_0}\int \frac{[\rho(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x',$$
$$\boldsymbol{A}(t,\boldsymbol{x}) = \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{J}(t',\boldsymbol{x}']_{\text{ret}}}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x'.$$

This has been shown here in *Lorenz gauge*, but the Maxwell equations are *invariant under gauge transformations* so this solution may be taken to be valid generally.

Chapter 8

Minkowski Spacetime

These lectures concern *graduate-level classical electromagnetism*.

- The *distinction between classical and modern physics* generally turns on whether dynamics are described by
  - *quantum mechanics and quantum field theory* or *classical mechanics and classical field theory*,
  - and by *non-relativistic mechanics* or *relativistic mechanics*.
- Traditionally *classical electromagnetism excludes quantum mechanics* (except for a few quantum concepts necessary for electromagnetism interacting with the microscopic structure of matter).
- but traditionally *special relativity* has been part of the discussion of classical electromagnetism.

If *strong gravity is important*, as it would be in various electromagnetic phenomena in astrophysics (for example, *black hole accretion disks*),

- 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.

Accordingly, in this chapter we shall introduce *special relativity* and the *Lorentz transformations* on which it is based.

- Then we will demonstrate explicitly the *Lorentz invariance of Maxwell's equations* and therefore,
- their *consistency with special relativity*.

8.1 Maxwell's Equations and Special Relativity

The reasons for the affinity of Maxwell's equations with special relativity are both *scientific* and *historical*. *Scientifically*,

- *electrical charges* may be approximated as *classical objects* (positions and momenta simultaneously well defined)
- so that *quantum mechanics isn't 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.

Historically, classical electromagnetism and special relativity have been intertwined because

- *Einstein was* influenced strongly by the *beauty and symmetry of Maxwell's equations* in formulating the special theory of relativity.
- In particular he was motivated by comparing
  - the *Lorentz invariance* of Maxwell's equations with
  - the *Galilean invariance* of classical mechanics.

This led Einstein to the conclusion that

- *classical mechanics, not electromagnetism*, 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 special relativity*.

Accordingly, let us turn to an understanding of the *special theory of relativity*.

8.2 Minkowski Spacetime and Spacetime Tensors

The most elegant formulation of special relativity is in terms of

- a *4D spacetime manifold* called *Minkowski space*, and in terms of
- *spacetime tensors* that are a consequence of the *differential geometry* of that manifold.

To begin, we must consider *manifolds* (loosely "spaces" to a physicist) and *transformations within that manifold*.

8.3 Coordinate Systems and Transformations

A physical system has a *symmetry under some operation* if the system after the operation is *observationally indistinguishable* from the system before the operation.

Example: A *perfectly uniform sphere* has a symmetry under rotation about any axis because after the rotation the sphere looks the same as before the rotation.

The theory of relativity may be viewed as a *symmetry under coordinate transformations*.

- Two observers, referencing their measurements of the same physical phenomena to two different coordinate systems
- should deduce the *same laws of physics* from their observations.
- In *special relativity* one requires a symmetry under only a *subset of possible coordinate transformations* (between systems that are not accelerated with respect to each other).
- *General relativity* requires that the laws of physics be invariant under the *most general coordinate transformations*.

To understand special and general relativity, we must begin by examining the *transformations that are possible* between different coordinate systems.

8.4 Minkowski Space

The *surface traced out* by allowing coordinates $(x^0,x^1,x^2,x^3) = (ct,x,y,z)$ to range over all possible values

- defines the *manifold of 4-dimensional spacetime*.
- The resulting space is called *Minkowski spacetime,*
- or more briefly *Minkowski space* or just *spacetime.*
- In Minkowski space the *square of the infinitesimal distance* ds^2 between two points

$$(ct,x,y,z) \quad \text{and} \quad (ct+cdt,x+dx,y+dy,z+dz)$$

is given by

$$\begin{aligned} ds^2 &= \sum_{\mu\nu} \eta_{\mu\nu} dx^\mu dx^\nu = -c^2dt^2 + dx^2 + dy^2 + dz^2, \\ &= -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2, \end{aligned}$$

- which is called the *line element* of the Minkowski space.

In this notation

- ds^2 means the square of ds [that is, $(ds)^2$],
- and dx^2 means $(dx)^2$, but
- in (x^0,x^1,x^2,x^3) the superscripts are indices and not powers.

The quantity $\eta_{\mu\nu}$ appearing in the line element

$$ds^2 = \sum_{\mu\nu} \eta_{\mu\nu} dx^\mu dx^\nu = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2,$$

- may be expressed as the diagonal matrix

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

- and is termed the *metric tensor* of the Minkowski space.
- The line element ds^2 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 signs on the diagonal of

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

- defines the *signature of the metric.*
- The *Minkowski space signature* is $(-+++)$.
- A metric such as that of Minkowski space in which the signs in the sign pattern are *not all the same* is termed an *indefinite metric*.
- Some authors use instead the signature $(+---)$ for the sign pattern in $\eta_{\mu\nu}$.
- The important point is that *for Minkowski spacetime*
  - the last three terms have the *same sign* and
  - the sign of the first term is *different* from that of the other three.

Assuming the modern convention of placing the timelike coordinate in the first position and the spacelike coordinates in the following positions.

The *geometry of 4D Minkowski space* differs from that of *4D Euclidean space*:

- 4D Minkowski spacetime is *not* "just like ordinary space but with more dimensions".
- The difference is encoded in the *signature of the metric*,
- which for 4D *euclidean space* is $(++++)$,
- compared with the signature $(-+++)$ for the *Minkowski metric*.
- That is, the metric tensor of *4D euclidean space* is just the 4×4 unit matrix.
- That *change in sign* for the first entry makes all the difference.
- Most of the *unusual features* of special relativity
  - *space contraction*,
  - *time dilation*,
  - *relativity of simultaneity*,
  - the *twin "paradox"*, ...

  follow from this *difference in geometry* between 4D Minkowski spacetime and 4D euclidean space that is encoded in that sign difference.

8.4.1 Coordinate Systems in Euclidean Space

Our goal is to describe transformations between coordinates in a possibly curved space having

- *three space-like coordinates* and
- *one timelike coordinate*.

However, to introduce these concepts we shall begin with the *simpler and more familiar case* of vector fields in 2D and 3D euclidean space.

- Assume a *3D euclidean (flat) space* having a *cartesian coordinate system* (x,y,z), and an associated set of mutually *orthogonal unit vectors* $(\boldsymbol{i},\boldsymbol{j},\boldsymbol{k})$.
- Assume that there is an *alternative coordinate system* (u,v,w), not necessarily cartesian, with the (x,y,z) coordinates related to the (u,v,w) coordinates by

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

- Assume that the *transformation is invertible* so that we can solve for (u,v,w) in terms of (x,y,z).

Example:

Take the (u,v,w) system to be the spherical coordinates (r,θ,ϕ), in which case

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

takes the familiar form

$$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\cos\theta,$$

with the ranges of values $r \geq 0$ and $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$.

- The equations

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

can be combined into a vector equation that gives a *position vector* $\boldsymbol{r}$ *for a point* in the space in terms of the (u,v,w) coordinates:

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k}.$$

- For example, in terms of the *spherical coordinates* (r,θ,ϕ),

$$\boldsymbol{r} = (r\sin\theta\cos\phi)\,\boldsymbol{i} + (r\sin\theta\sin\phi)\,\boldsymbol{j} + (r\cos\theta)\,\boldsymbol{k}.$$

- The second coordinate system in these examples generally is not cartesian but the *space is assumed to be euclidean*.
- In transforming from the (x,y,z) coordinates to the (r,θ,ϕ) coordinates, we are just using a different scheme to label points in a flat space.
- This distinction is important because shortly we shall consider general coordinate transformations in spaces that may not obey euclidean geometry (curved spaces).

8.4.2 Basis Vectors

At any point $P(u_0, v_0, w_0)$ defined for specified coordinates (u_0, v_0, w_0), three surfaces pass. They are defined by $u = u_0$, $v = v_0$, and $w = w_0$, respectively.

- Any two of these surfaces meet in curves.
- From

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k}.$$

  we may obtain general parametric equations for coordinate surfaces or curves by setting one or two of the variables (u, v, w) equal to constants.
- For example, if we set v and w to constant values, $v = v_0$ and $w = w_0$, we obtain a parametric equation for a curve given by the intersection of $v = v_0$ and $w = w_0$,

$$\boldsymbol{r}(u) = x(u,v_0,w_0)\,\boldsymbol{i} + y(u,v_0,w_0)\,\boldsymbol{j} + z(u,v_0,w_0)\,\boldsymbol{k},$$

- This is a parametric equation in which u plays the role of a coordinate along the curve defined by the constraints $v = v_0$ and $w = w_0$.

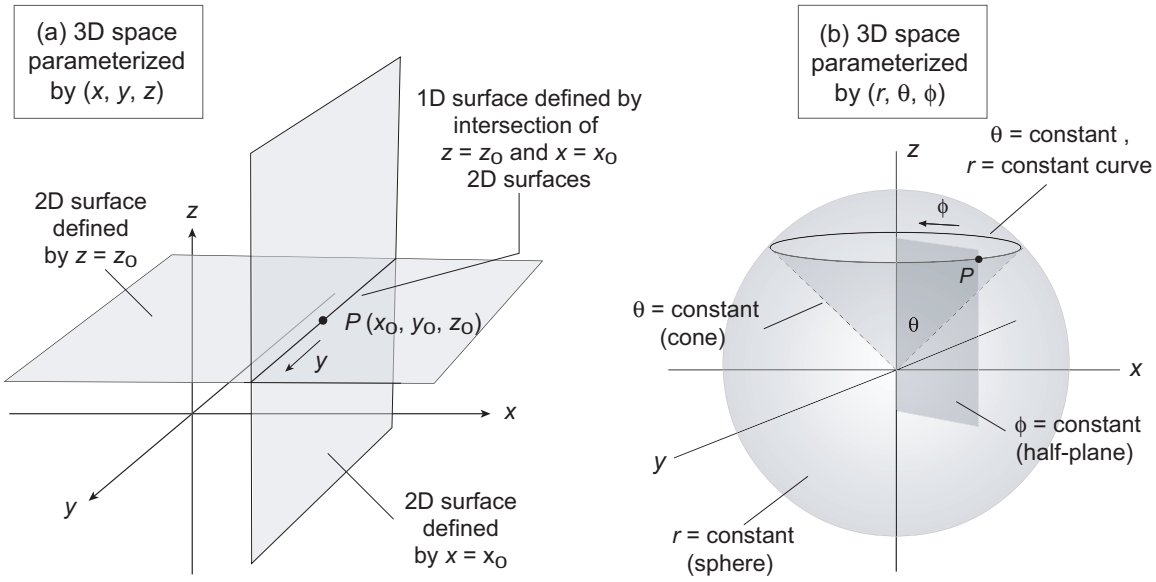


Figure 8.1: Examples of surfaces and curves arising from constraints. (a) In 3D euclidean space parameterized by cartesian coordinates (x, y, z), the constraints $x = x_0$ and $z = z_0$ define 2D planes and the intersection of these planes defines a 1D surface parameterized by the variable y. (b) In 3D space described in spherical coordinates (r, θ, ϕ), the constraint $r =$ 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.

Fig. 8.1(b) illustrates for spherical coordinates (r, θ, ϕ):

- The surface corresponding to r = constant is a *sphere* parameterized by the variables θ and ϕ.
- The constraint θ = constant corresponds to a *cone* parameterized by the variables r and ϕ.
- Setting both r and θ to constants defines a *curve* that is the *intersection of the sphere and the cone*, which is parameterized by the variable ϕ.

- *Partial differentiation* of

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k},$$

  with respect to u, v, and w, respectively, gives *tangents to the coordinate curves* passing though the point P.

- These may be used to define a set of *basis vectors* $\boldsymbol{e}_i$ through

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

  with all partial derivatives *evaluated at the point* $P = (u_0, v_0, w_0)$.

- This basis, generated by the tangents to the coordinate curves, is sometimes termed the *natural basis.* The following example illustrates for a spherical coordinate system.

Example:

Consider the spherical coordinate system defined through

$$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\cos\theta.$$

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 natural basis is obtained from

$$\begin{aligned}
\boldsymbol{e}_1 \equiv \boldsymbol{e}_r &= \frac{\partial \boldsymbol{r}}{\partial r} = (\sin\theta\cos\phi)\,\boldsymbol{i} + (\sin\theta\sin\phi)\,\boldsymbol{j} + (\cos\theta)\,\boldsymbol{k} \\
\boldsymbol{e}_2 \equiv \boldsymbol{e}_\theta &= \frac{\partial \boldsymbol{r}}{\partial \theta} = (r\cos\theta\cos\phi)\,\boldsymbol{i} + (r\cos\theta\sin\phi)\,\boldsymbol{j} - (r\sin\theta)\,\boldsymbol{k} \\
\boldsymbol{e}_3 \equiv \boldsymbol{e}_\phi &= \frac{\partial \boldsymbol{r}}{\partial \phi} = -(r\sin\theta\sin\phi)\,\boldsymbol{i} + (r\sin\theta\cos\phi)\,\boldsymbol{j}.
\end{aligned}$$

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,

$$\begin{aligned}
\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 &= r\sin\theta\cos\theta\cos^2\phi + r\sin\theta\cos\theta\sin^2\phi - r\cos\theta\sin\theta \\
&= r\sin\theta\cos\theta\underbrace{(\cos^2\phi + \sin^2\phi)}_{=1} - r\cos\theta\sin\theta = 0.
\end{aligned}$$

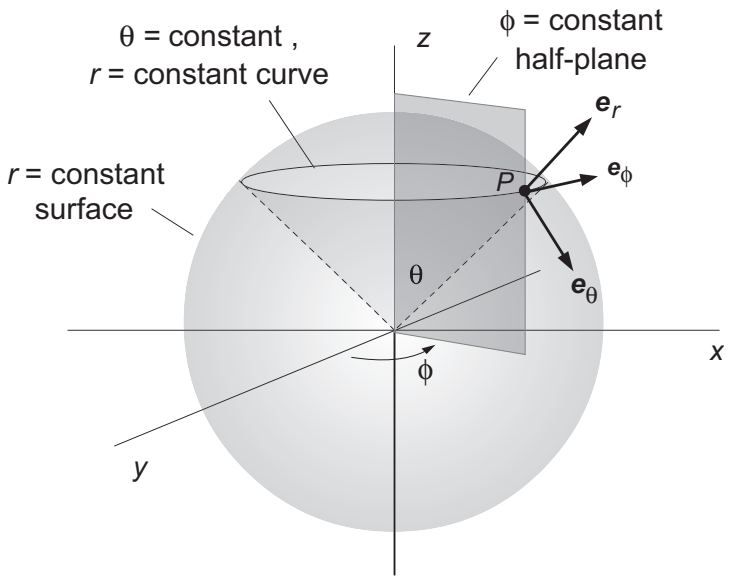


Figure 8.2: Basis vectors for the natural basis in spherical coordinates.

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 we can use these to define a *normalized basis*,

$$\hat{\boldsymbol{e}}_1 \equiv \frac{\boldsymbol{e}_1}{|\boldsymbol{e}_1|} = (\sin\theta\cos\phi)\,\boldsymbol{i} + (\sin\theta\sin\phi)\,\boldsymbol{j} + (\cos\theta)\,\boldsymbol{k}$$
$$\hat{\boldsymbol{e}}_2 \equiv \frac{\boldsymbol{e}_2}{|\boldsymbol{e}_2|} = (\cos\theta\cos\phi)\,\boldsymbol{i} + (\cos\theta\sin\phi)\,\boldsymbol{j} - (\sin\theta)\,\boldsymbol{k}$$
$$\hat{\boldsymbol{e}}_3 \equiv \frac{\boldsymbol{e}_3}{|\boldsymbol{e}_3|} = -(\sin\phi)\,\boldsymbol{i} + (\cos\phi)\,\boldsymbol{j}.$$

These basis vectors are now

- mutually *orthogonal* and
- of *unit length*.

They are illustrated in the figure shown above.

- In many applications it is usual to assume that the coordinate system is orthogonal so that the basis vectors

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

  are mutually orthogonal, and to normalize these basis vectors to unit length.

- In the more general applications that will interest us, the natural basis defined by the partial derivatives in the preceding equation *need not be orthogonal or normalized* to unit length

However, in the simple examples shown so far the natural basis is in fact orthogonal.

8.4.3 Dual Basis

It is also valid to construct a *basis at* P by using *normals to the coordinate surfaces* rather than the tangents to curves.

- We assume that

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

is invertible so we may solve for

$$u = u(x,y,z) \qquad v = v(x,y,z) \qquad w = w(x,y,z),$$

- The *gradients*

$$\nabla u = \frac{\partial u}{\partial x}\boldsymbol{i} + \frac{\partial u}{\partial y}\boldsymbol{j} + \frac{\partial u}{\partial z}\boldsymbol{k}$$
$$\nabla v = \frac{\partial v}{\partial x}\boldsymbol{i} + \frac{\partial v}{\partial y}\boldsymbol{j} + \frac{\partial v}{\partial z}\boldsymbol{k}$$
$$\nabla w = \frac{\partial w}{\partial x}\boldsymbol{i} + \frac{\partial w}{\partial y}\boldsymbol{j} + \frac{\partial w}{\partial z}\boldsymbol{k}$$

are *normal to the three surfaces* through P defined by $u = u_0$, $v = v_0$, and $w = w_0$, respectively.

- Therefore, we may choose as an alternative to the *natural basis*

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

the *dual basis*

$$\boldsymbol{e}^u \equiv \nabla u \qquad \boldsymbol{e}^v \equiv \nabla v \qquad \boldsymbol{e}^w \equiv \nabla w.$$

- This basis $(\boldsymbol{e}^u, \boldsymbol{e}^v, \boldsymbol{e}^w)$, defined in terms of normals, is said to be the *dual* of the normal basis, defined in terms of tangents.
- Notice that we have chosen to distinguish the basis

$$\boldsymbol{e}^u \equiv \boldsymbol{\nabla} u \qquad \boldsymbol{e}^v \equiv \boldsymbol{\nabla} v \qquad \boldsymbol{e}^w \equiv \boldsymbol{\nabla} w.$$

  from the basis

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

  by using *superscript indices* and *subscript indices*, respectively.

These two bases 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 length.
- If the basis vectors are normalized, the normal basis and the dual basis for orthogonal coordinates are equivalent and our preceding distinction is not significant.
- However, for *non-orthogonal coordinate systems* the two bases generally are *not equivalent* and the distinction between upper and lower indices is relevant.

The following example illustrates.

Example:

Define a coordinate system (u,v,w) in terms of cartesian coordinates (x,y,z) through

$$x = u+v \qquad y = u-v \qquad z = 2uv+w.$$

The position vector for a point $\boldsymbol{r}$ is then

$$\boldsymbol{r} = x\boldsymbol{i}+y\boldsymbol{j}+z\boldsymbol{k} = (u+v)\boldsymbol{i}+(u-v)\boldsymbol{j}+(2uv+w)\boldsymbol{k}$$

The natural basis is

$$\boldsymbol{e}_1 \equiv \boldsymbol{e}_u = \frac{\partial \boldsymbol{r}}{\partial u} = \boldsymbol{i}+\boldsymbol{j}+2v\boldsymbol{k}$$
$$\boldsymbol{e}_2 \equiv \boldsymbol{e}_v = \frac{\partial \boldsymbol{r}}{\partial v} = \boldsymbol{i}-\boldsymbol{j}+2u\boldsymbol{k}$$
$$\boldsymbol{e}_3 \equiv \boldsymbol{e}_w = \frac{\partial \boldsymbol{r}}{\partial w} = \boldsymbol{k}.$$

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 thus the dual basis is

$$\boldsymbol{e}^1 \equiv \boldsymbol{e}^u = \boldsymbol{\nabla} u = \frac{\partial u}{\partial x}\boldsymbol{i}+\frac{\partial u}{\partial y}\boldsymbol{j}+\frac{\partial u}{\partial z}\boldsymbol{k} = \tfrac{1}{2}(\boldsymbol{i}+\boldsymbol{j})$$
$$\boldsymbol{e}^2 \equiv \boldsymbol{e}^v = \boldsymbol{\nabla} v = \frac{\partial v}{\partial x}\boldsymbol{i}+\frac{\partial v}{\partial y}\boldsymbol{j}+\frac{\partial v}{\partial z}\boldsymbol{k} = \tfrac{1}{2}(\boldsymbol{i}-\boldsymbol{j})$$
$$\boldsymbol{e}^3 \equiv \boldsymbol{e}^w = \boldsymbol{\nabla} w = \frac{\partial w}{\partial x}\boldsymbol{i}+\frac{\partial w}{\partial y}\boldsymbol{j}+\frac{\partial w}{\partial z}\boldsymbol{k} = -(u+v)\boldsymbol{i}+(u-v)\boldsymbol{j}+\boldsymbol{k}.$$

- What about orthogonality? We can *check by taking scalar products*. For example,

$$\begin{aligned}\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 &= (\boldsymbol{i}+\boldsymbol{j}+2v\boldsymbol{k})\cdot(\boldsymbol{i}-\boldsymbol{j}+2u\boldsymbol{k}) \\ &= \boldsymbol{i}^2-\boldsymbol{j}^2+4uv = 4uv,\end{aligned}$$

where the *orthonormality of the basis* $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$ has been used. For the *natural basis* we find in general

$$\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = 4uv \qquad \boldsymbol{e}_2 \cdot \boldsymbol{e}_3 = 2u \qquad \boldsymbol{e}_3 \cdot \boldsymbol{e}_1 = 2v.$$

Thus the normal basis is *non-orthogonal.*

- Taking the *scalar products* of the natural 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 natural basis is also *not normalized to unit length.*

- It is also clear from the above expressions that generally $\boldsymbol{e}_i$ is not parallel to $\boldsymbol{e}^i$, so in this non-orthogonal case we see that *the normal basis and the dual basis are distinct.*

The preceding example illustrates that for the general case of coordinate systems that are not orthogonal,

$$\boldsymbol{e}^u \equiv \boldsymbol{\nabla} u \qquad \boldsymbol{e}^v \equiv \boldsymbol{\nabla} v \qquad \boldsymbol{e}^w \equiv \boldsymbol{\nabla} w \qquad \text{(dual basis)}$$

and

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w} \qquad \text{(natural basis)}$$

define *different but equally valid bases*, and the placement of indices in upper or lower positions is important.

> Henceforth the reader should assume that the *upper or lower placement of indices in equations is significant.*

8.4.4 Expansion of Vectors

An arbitrary vector $\boldsymbol{V}$ may be expanded in terms of the tangent basis $\{\boldsymbol{e}_i\}$ and an arbitrary dual vector $\boldsymbol{\omega}$ may be expanded in terms of the dual basis $\{\boldsymbol{e}^i\}$:

$$\boldsymbol{V} = V^1\boldsymbol{e}_1 + V^2\boldsymbol{e}_2 + V^3\boldsymbol{e}_3 = \sum_{i=1}^{3} V^i\boldsymbol{e}_i \equiv V^i\boldsymbol{e}_i \qquad \text{(natural basis)}$$

$$\boldsymbol{\omega} = \omega_1\boldsymbol{e}^1 + \omega_2\boldsymbol{e}^2 + \omega_3\boldsymbol{e}^3 = \sum_{i=1}^{3} \omega_i\boldsymbol{e}^i \equiv \omega_i\boldsymbol{e}^i \qquad \text{(dual basis)}$$

where we have introduced in the last step of each equation the *Einstein summation convention:*

- An index appearing twice on one side of an equation, once as a lower index and once as an upper index, implies a *summation on that index.*
- The summation index 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 of the equation.
- If an index appears *more than twice* on the same side of an equation, it *probably indicates a mistake*.
- Since the dummy (repeated) index is summed over, it *doesn't matter what the repeated index is*, as long as it is not equivalent to another index in the equation.

> From this point onward, we shall usually *assume the Einstein summation convention*.

8.4.5 Scalar Product of Vectors and the Metric Tensor

- The upper-index coefficients V^i of the basis vectors $\boldsymbol{e}_i$ in

$$\boldsymbol{V}=V^1\boldsymbol{e}_1+V^2\boldsymbol{e}_2+V^3\boldsymbol{e}_3$$

  are termed the *components of the vector* in the basis $\boldsymbol{e}_i = \{\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3\}$.

- The lower-index coefficients ω_i of the basis vectors $\boldsymbol{e}^i$ in

$$\boldsymbol{\omega}=\omega_1\boldsymbol{e}^1+\omega_2\boldsymbol{e}^2+\omega_3\boldsymbol{e}^3$$

  are termed the *components of the dual vector* in the basis $\boldsymbol{e}^i = \{\boldsymbol{e}^1,\boldsymbol{e}^2,\boldsymbol{e}^3\}$.

- Remember: *Components of vectors and dual vectors generally are distinct* for non-orthogonal coordinate systems.

However,

- Vector and dual vector spaces are *related fundamentally*.
- This permits vector components V^i and dual vector components ω_i to be treated as if they were different components of the same vector.

The first step in establishing this relationship is to introduce a *scalar product* and a *metric*.

8.4.6 Vector Scalar Product and the Metric Tensor

Utilizing the expansions in a vector basis:

- We can write the scalar product of two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as

$$\boldsymbol{A}\cdot\boldsymbol{B} = (A^i\boldsymbol{e}_i)\cdot(B^j\boldsymbol{e}_j) = \boldsymbol{e}_i\cdot\boldsymbol{e}_j A^i B^j = g_{ij}A^iB^j,$$

  where the *metric tensor component* g_{ij} is defined by

$$g_{ij} \equiv \boldsymbol{e}_i\cdot\boldsymbol{e}_j.$$

- Likewise, for the scalar product of dual vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

$$\boldsymbol{\alpha}\cdot\boldsymbol{\beta} = \alpha_i\boldsymbol{e}^i\cdot\beta_j\boldsymbol{e}^j = g^{ij}\alpha_i\beta_j,$$

  where metric tensor components with upper indices are

$$g^{ij} \equiv \boldsymbol{e}^i\cdot\boldsymbol{e}^j,$$

  and the scalar product of dual vectors and vectors is

$$\boldsymbol{\alpha}\cdot\boldsymbol{B} = \alpha_i\boldsymbol{e}^i\cdot B^j\boldsymbol{e}_j = g^i_j\alpha_i B^j,$$

  where the metric tensor component with mixed indices is

$$g^i_j \equiv \boldsymbol{e}^i\cdot\boldsymbol{e}_j.$$

General properties of the metric tensor will be discussed below but first we use it to establish a relationship called *duality* between vector and dual vector spaces.

8.4.7 Relationship of Vectors and Dual vectors

There is little practical difference 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.
- The essential issue is *how to define a vector or dual vector in a curved space,* and what that implies.

The essential mathematics will be discussed in more depth later, but the salient points are that

1. *Vectors* are not specified directly in a curved space, but instead are defined in a *euclidean vector space* attached to the manifold at *each point* called the *tangent space.*
2. Likewise, *dual vectors* are defined in a *euclidean vector space* attached to the 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 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 later, vectors and dual vectors are special cases of *tensors,* and this permits an abstract definition in terms of mappings from vectors and dual vectors to real numbers.

To be specific,

- Dual vectors $\boldsymbol{\omega}$ are linear maps of vectors $\boldsymbol{V}$ to the real numbers: $\boldsymbol{\omega}(\boldsymbol{V}) = \omega_i V^i \in \mathbb{R}$.
- Vectors $\boldsymbol{V}$ are linear maps of dual vectors $\boldsymbol{\omega}$ to the real numbers: $\boldsymbol{V}(\boldsymbol{\omega}) = V^i \omega_i \in \mathbb{R}$.

Expressions like $\boldsymbol{\omega}(\boldsymbol{V}) = \omega_i V^i \in \mathbb{R}$ can be read as

- "Dual vectors $\boldsymbol{\omega}$ act linearly on 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 that the space of vectors and the space of dual vectors are both linear vector spaces.
- 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

$$\boldsymbol{\omega}(\boldsymbol{V}) = \boldsymbol{V}(\boldsymbol{\omega}) = V^i \omega_i \in \mathbb{R}.$$

Notice further that $\boldsymbol{A} \cdot \boldsymbol{B} = g_{ij} A^i B^j$

- Defines a *linear map from vectors to real numbers*, since it takes two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as arguments and returns the scalar product, which is a real number.
- Thus one may write

$$\boldsymbol{A}(\boldsymbol{B}) = \boldsymbol{A} \cdot \boldsymbol{B} \equiv A_i B^i = g_{ij} A^i B^j.$$

- But since in $A_i B^i = g_{ij} A^i B^j$ the vector B is arbitrary,

$$A_i = g_{ij} A^j,$$

- This specifies a correspondence between a vector with components A^i in the tangent space and a dual vector with components A_i in the cotangent space.
- Likewise, the above expression can be inverted using that the inverse of g_{ij} is g^{ij} to give

$$A^i = g^{ij} A_j.$$

- Hence, using the metric tensor to raise and lower indices by summing over a repeated index (*contraction*),
- we see that *vector and dual vector components are related through contraction with the metric tensor.*
- This is the precise sense in which the tangent and cotangent spaces are dual: *they are different, but closely related through the metric tensor.*

The duality of the vector and dual vector spaces may be incorporated concisely by

- Requiring that for the basis vectors $\{\boldsymbol{e}_i\}$ and basis dual vectors $\{\boldsymbol{e}^i\}$ satisfy

$$\boldsymbol{e}^i(\boldsymbol{e}_j) = \boldsymbol{e}^i \cdot \boldsymbol{e}_j = \delta^i_j,$$

  where the Kronecker delta is defined by

$$\delta^i_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

- 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,

$$V^i = \boldsymbol{e}^i \cdot \boldsymbol{V} \qquad V_i = \boldsymbol{e}_i \cdot \boldsymbol{V}.$$

A lot of important mathematics has transpired in the last few equations, so let's take stock.

- For a space with *metric tensor*, vectors and dual vectors are in a *one-to-one relationship* that permits them to be manipulated effectively as if a dual vector were just a vector with a lower index.
- Indices on vectors can be *raised or lowered* as desired by *contraction with the metric tensor*.

Since all spaces of interest here have metrics, this reduces the practical implications of the distinction between vectors and dual vectors to keeping proper track of upper and lower positions for indices.

8.4.8 Properties of the Metric Tensor

- Because it may be defined through scalar products of basis vectors, the metric tensor must be symmetric in its indices:

$$g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}.$$

- Since

$$g_{ij}a^i b^j = g_{ij}b^j a^i = a^i b_i \qquad g^{ij}a_i b_j = g^{ij}b_j a_i = a_i b^i$$

are *valid for arbitrary vector components,* it follows that

$$g_{ij}b^j = b_i \qquad g^{ij}b_j = b^i.$$

That is,

> Contraction with the metric tensor may be used to raise or lower an index on a vector.

- Thus the *scalar product of two vectors* may be written in any of the following equivalent ways,

$$\boldsymbol{a}\cdot\boldsymbol{b} \equiv a^i b_i \equiv a_i b^i = g_{ij}a^i b^j = g^{ij}a_i b_j = g^i_j a_i b^j.$$

- From the preceding expressions

$$b^i = g^{ij} b_j = g^{ij} \underbrace{g_{jk} b^k}_{b_j} \qquad b^i = \delta^i_k b^k,$$

and since this is valid for arbitrary components b^i,

$$g^{ij} g_{jk} = g_{kj} g^{ji} = \delta^i_k.$$

- Viewing g^{ij} as the elements of a matrix G and g_{ij} as the elements of a matrix $\tilde{G}$, the equations

$$g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}.$$

are equivalent to the matrix equations

$$G = G^{\mathrm{T}} \qquad \tilde{G} = \tilde{G}^{\mathrm{T}},$$

where T denotes the transpose. The Kronecker delta is just the 3×3 unit matrix I, implying that

$$g^{ij} g_{jk} = g_{kj} g^{ji} = \delta^i_k$$

may be written as the matrix equations

$$\tilde{G} G = G \tilde{G} = I.$$

The matrix corresponding to the covariant components of the metric tensor is the inverse of the matrix corresponding to the contravariant components of the metric tensor: *one may be obtained from the other by matrix inversion*.

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 : $\tilde{G} G = G \tilde{G} = I \qquad G \equiv [g^{ij}] \qquad \tilde{G} \equiv [g_{ij}]$

Some basic properties of the metric tensor are summarized in the box above.

8.4.9 Line Elements and Distances

- Consider coordinates $u^1(t)$, $u^2(t)$, and $u^3(t)$ parameterized by t.
- As the parameter t varies, the points characterized by the 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 three-dimensional space.

- The position vector for these points as a function of t is

$$\begin{aligned}\boldsymbol{r}(t) = x(u^1(t),u^2(t),u^3(t))\,\boldsymbol{i} + y(u^1(t),u^2(t),u^3(t))\,\boldsymbol{j} \\ + z(u^1(t),u^2(t),u^3(t))\,\boldsymbol{k},\end{aligned}$$

- By the chain rule

$$\begin{aligned}\frac{d\boldsymbol{r}}{dt} &= \frac{\partial \boldsymbol{r}}{\partial u^1}\frac{du^1}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^2}\frac{du^2}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^3}\frac{du^3}{dt} \\ \dot{\boldsymbol{r}} &= \dot{u}^1\boldsymbol{e}_1 + \dot{u}^2\boldsymbol{e}_2 + \dot{u}^3\boldsymbol{e}_3,\end{aligned}$$

where the definitions

$$\dot{\boldsymbol{r}} \equiv \frac{d\boldsymbol{r}}{dt} \qquad \boldsymbol{e}_i \equiv \frac{\partial \boldsymbol{r}}{\partial u^i} \qquad \dot{u}^i \equiv \frac{du^i}{dt}$$

have been used.

- In summation convention the equation

$$\dot{\boldsymbol{r}} = \dot{u}^1 \boldsymbol{e}_1 + \dot{u}^2 \boldsymbol{e}_2 + \dot{u}^3 \boldsymbol{e}_3,$$

  is $\dot{\boldsymbol{r}} = \dot{u}^i \boldsymbol{e}_i$.

- This may be expressed in differential form as $d\boldsymbol{r} = du^i \boldsymbol{e}_i$.

- Thus the squared infinitesimal distance along the curve is

$$\begin{aligned} ds^2 = d\boldsymbol{r} \cdot d\boldsymbol{r} &= du^i \boldsymbol{e}_i \cdot du^j \boldsymbol{e}_j \\ &= \boldsymbol{e}_i \cdot \boldsymbol{e}_j du^i du^j \\ &= g_{ij} du^i du^j, \end{aligned}$$

  where $g_{ij} \equiv \boldsymbol{e}_i \cdot \boldsymbol{e}_j$ has been used.

- Notice that in expressing the line element

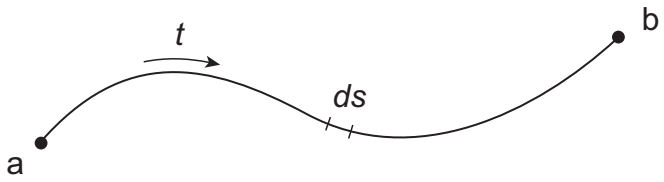
$$ds^2 = g_{ij}du^i du^j$$

we use the usual convention that $d\alpha^2 \equiv (d\alpha)^2$.

- That is, $d\alpha^2$ means the square of $d\alpha$, not the differential of α^2.
- Thus $ds^2 = g_{ij}du^i du^j$ is the *infinitesimal line element* for the space described by the metric g_{ij}.
- The length d of a finite segment between points a and b is obtained from the integral

$$d = \int_a^b ds = \int_a^b \left(g_{ij}du^i du^j\right)^{1/2} = \int_a^b \left(g_{ij}\frac{du^i}{dt}\frac{du^j}{dt}\right)^{1/2} dt,$$

where t parameterizes the position along the segment.



For example:

- The line element for two-dimensional euclidean space in cartesian coordinates (x, y) is given by

$$ds^2 = dx^2 + dy^2,$$

  which is just the Pythagorean theorem for right triangles having infinitesimal sides.

- The corresponding line element expressed in plane polar coordinates (r, ϕ) is then the familiar

$$ds^2 = dr^2 + r^2 d\phi^2.$$

Example:

For plane polar coordinates (r,ϕ) we have

$$x = r\cos\phi \qquad y = r\sin\phi,$$

so the position vector may be expressed as

$$\boldsymbol{r} = (r\cos\phi)\,\boldsymbol{i} + (r\sin\phi)\,\boldsymbol{j}.$$

Then the basis vectors in the natural 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 then follow from $g_{ij} \equiv \boldsymbol{e}_i \cdot \boldsymbol{e}_j$:

$$g_{11} = \cos^2\phi + \sin^2\phi = 1 \qquad g_{22} = r^2(\cos^2\phi + \sin^2\phi) = r^2$$

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

$$ds^2 = g_{ij}\,dx^i du^j = g_{11}(du^1)^2 + g_{22}(du^2)^2 = dr^2 + r^2 d\phi^2,$$

where $u^1 = r$ and $u^2 = \phi$.

This can be expressed as the matrix equation

$$\begin{aligned} ds^2 &= (dr \;\; d\phi) \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \begin{pmatrix} dr \\ d\phi \end{pmatrix} \\ &= (dr \;\; d\phi) \begin{pmatrix} dr \\ r^2 d\phi \end{pmatrix} = dr^2 + r^2 d\phi^2. \end{aligned}$$

The line elements expressed in cartesian and polar coordinates in the preceding two examples

- Correspond to the *same space,* parameterized in terms of different coordinates.
- The form of the line element is different in the two parameterizations, but
- for any two nearby points the distance between them is given by ds, independent of the coordinate system.
- Thus, the line element ds is *invariant under coordinate transformations.*
- Since the distance between any two points that are not nearby can be obtained by integrating ds, we conclude that generally

> The *distance* between any two points is *invariant under coordinate transformations* for metric spaces.

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, we could verify all the axioms of euclidean geometry starting from the line elements if we chose to do so.

8.5 Integration

Integration enters into physical theories in various ways, for example in the formulation of conservation laws.

- It is important to understand how the volume element for integrals behaves under change of coordinate systems.
- Trivial in euclidean space with orthonormal coordinates.
- Non-trivial in curved spaces, or even in flat spaces parameterized in non-cartesian coordinates.

We 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 is in this case

$$dA = \sqrt{\det g}\, dx^1 dx^2,$$

  where $\det g$ is the determinant of the metric tensor g_{ij}.
- For orthonormal coordinates g_{ij} is a unit matrix so $(\det g)^{1/2} = 1$.
- 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.

As we will show later, this 2D example generalizes easily to define invariant integration in 4D spacetime.

8.6 Differentiation

Taking derivatives of vectors in spaces defined by *position-dependent metrics* will be crucial in general relativity.

- First consider the simpler case of taking the derivative of a vector in a euclidean space, but one *parameterized with a vector basis that may depend on the coordinates.*
- We may expand a vector $\boldsymbol{V}$ in a convenient basis $\boldsymbol{e}_i$,

$$\boldsymbol{V} = V^i \boldsymbol{e}_i.$$

- By the usual product rule, the partial derivative is given by a sum of two terms,

$$\frac{\partial \boldsymbol{V}}{\partial x^j} = \underbrace{\frac{\partial V^i}{\partial x^j}\boldsymbol{e}_i}_{\text{component}} + \underbrace{V^i \frac{\partial \boldsymbol{e}_i}{\partial x^j}}_{\text{basis}},$$

  1. The first term represents the *change in the component* V^i.
  2. The second term represents the *change in the basis vectors* $\boldsymbol{e}_i$.

- For the situation where we can choose a basis that is independent of coordinates, the second term is zero and we recover the expected formula.
- However, if the basis depends on the coordinates the second term will generally not be zero.

In the second term of

$$\frac{\partial \boldsymbol{V}}{\partial x^j} = \underbrace{\frac{\partial V^i}{\partial x^j}\boldsymbol{e}_i}_{\text{component}} + \underbrace{V^i \frac{\partial \boldsymbol{e}_i}{\partial x^j}}_{\text{basis}},$$

- The factor $\partial \boldsymbol{e}_i / \partial x^j$ resulting from the action of the derivative operator on the basis vectors is itself a vector and can be expanded in the vector basis,

$$\frac{\partial \boldsymbol{e}_i}{\partial x^j} = \Gamma^k_{ij} \boldsymbol{e}_k.$$

- The expansion coefficients Γ^k_{ij} may be interpreted as specifying the projection on the k axis of the rate of change in the j direction of a basis vector pointing in the i direction.

- The coefficients Γ^k_{ij} are called *connection coefficients* or *Christoffel symbols.*
- Their generalization to 4-dimensional spacetime will be discussed more extensively later.
- There we shall see that the connection coefficients are central to
  1. The definition of derivatives
  2. A prescription for parallel transport of vectors in curved spacetime.

8.7 Transformations

It often proves useful to express physical quantities in more than one coordinate system.

- It therefore becomes necessary to understand how to transform between coordinate systems.
- This issue becomes particularly important in general relativity where it is essential to ensure that the laws of physics are not altered by the most general transformation between coordinate systems.

Let's consider two simple examples.

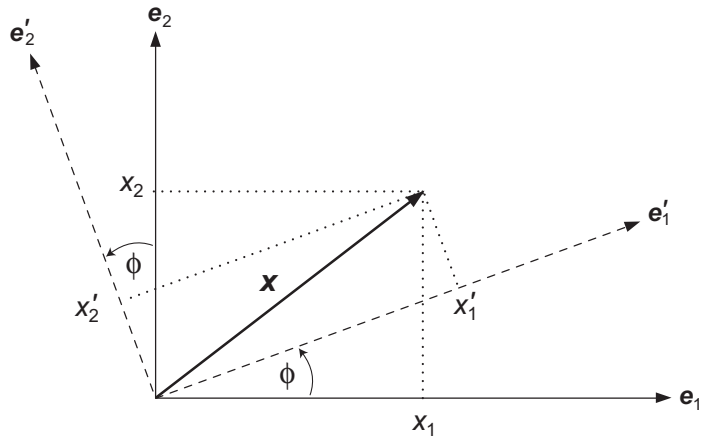


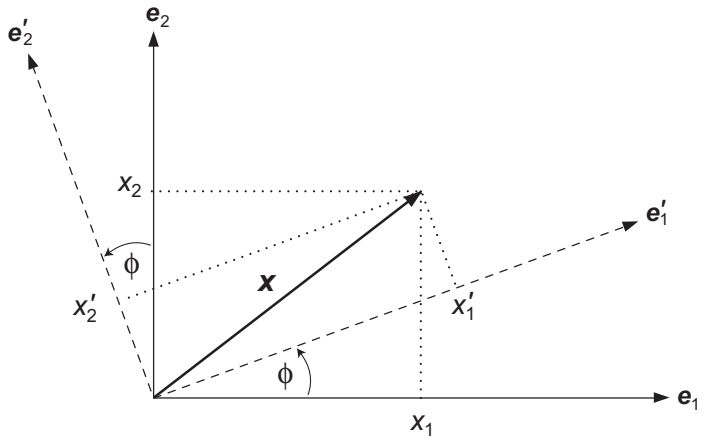
Figure 8.3: Rotation of coordinate system for a vector $\boldsymbol{x}$.

8.7.1 Rotational Symmetries

Consider the familiar example of the description of a vector under rotation of a coordinate system about the z axis by an angle ϕ, as illustrated in Fig. 8.3.

- In terms of the original basis vectors $\{\boldsymbol{e}_i\}$ the vector $\boldsymbol{x}$ has the components x_1 and x_2.
- After rotation of the coordinate system by the angle ϕ 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 of these bases:

$$\boldsymbol{x} = x^i \boldsymbol{e}_i = x'^i \boldsymbol{e}'_i,$$



- We may use the geometry of the above figure to find that the components in the two bases are related by the transformation

$$\begin{pmatrix} x'^1 \\ x'^2 \\ x'^3 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix},$$

  which may also be expressed as

$$x'^i = R^i_j x^j,$$

  where the R^i_j are the elements of the matrix in the preceding equation.

- This transformation law holds for any vector. (We may, in fact, *define* a vector in the x–y plane to be a quantity that obeys this transformation law.)

8.7.2 Galilean Transformations

Another simple example of a transformation is that between inertial frames in classical mechanics.

- Transformations between inertial frames with the same orientation are called *boosts*.
- In Newtonian physics time is considered an absolute quantity and boosts take the Galilean form

$$\boldsymbol{x}' = \boldsymbol{x}'(\boldsymbol{x},t) = \boldsymbol{x} - \boldsymbol{v}t \qquad t' = t'(\boldsymbol{x},t) = t.$$

- The Newtonian version of relativity asserts that the laws of physics are invariant under such Galilean transformations.
- Although the laws of mechanics at low velocity are invariant under Galilean transformations,
- *the laws of electromagnetism (Maxwell's equations) are not.*
- Indeed, the failure of Galilean invariance for the Maxwell equations was a large motivation in Einstein's eventual demonstration that *the laws of mechanics are not invariant with respect to Galilean transformations at high velocity.*

8.8 Spacetime Tensors and Covariance

The term *covariance* implies a formalism in which *the laws of physics maintain the same form* under a specified set of transformations.

EXAMPLE: Lorentz covariance implies equations that are constructed in such a way that they *do not change their form under Lorentz transformations* (three boosts between inertial systems and three rotations).[a]

[a] An inertial system is a frame of reference in which Newton's first law of motion holds. Thus, for example, rotating frames and accelerated frames are not inertial. An inertial system is therefore in uniform translational motion with respect to any other inertial frame.

8.8.1 Spacetime Coordinates and Transformations

We will be concerned with *spacetime,* which is an example of what mathematicians call a *manifold.*

- An *n-dimensional manifold* is
  - *a set* that can be parameterized continuously by
  - *n independent real coordinates* for each point (member of the set).
- We will assume the manifold to be differentiable at each point. Then we have a *differentiable manifold.*
- A *coordinate system*
  - associates n real parameters (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.

> A *cartesian product* $X \times Y$ of two sets X and Y is the set of *all possible ordered pairs* (x, y) with x an element of X and y an element of Y.

- Generally, *more than one overlapping set of coordinates* is required to parameterize an entire manifold uniquely.
  - For example, at least 2 overlapping sets of coordinates are required to parameterize a 2D sphere.
  - A set of coordinates covering one region of the manifold is called a *chart* and
  - a collection of charts sufficient to parameterize a manifold is called an *atlas*.

  Our concern here will be with *flat spacetime* and we won't have to worry about charts and atlases.

- Subsets of points within a manifold define
  - *curves* and
  - *surfaces*,

  which represent *submanifolds* of the full manifold.

- The manifolds that will interest us will be endowed with additional structure: specifically, a *geometry* specified by a *quadratic metric* called *Riemannian geometry*).

We will sometimes use loose physics language and refer to manifolds simply as *spaces*.

Because relativity implies that space and time enter descriptions of nature on comparable footings,

- it will be useful to unify them into a 4-dimensional continuum termed *spacetime*.
- Spacetime is an example of a *differentiable manifold.*
- In spacetime points will be defined by *coordinates having four components*,
  - the first labeling the time multiplied by the speed of light c,
  - the other three labeling the spatial coordinates:

$$x \equiv x^\mu = (x^0, x^1, x^2, x^3) = (ct, \boldsymbol{x}),$$

    where $\boldsymbol{x}$ denotes a vector with three 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 the earlier discussion, the *placement of indices in upper or lower positions is meaningful*.
- Bold symbols will be used to denote (ordinary) vectors defined in the three spatial degrees of freedom,
- with 4-component vectors in spacetime denoted in non-bold symbols.
- For spacetime the modern convention is to *number the indices beginning with zero* rather than one.

The coordinate systems of interest will be assumed to be quite general, subject only to the requirement that

- they *assign a coordinate uniquely to every point* of spacetime, and that they be
- *differentiable to sufficient order* for the task at hand at every spacetime point.

8.8.2 Covariance and Tensor Notation

- We shall be concerned generically with a transformation between one set of spacetime coordinates, denoted by

$$x \equiv x^\mu = (x^0, x^1, x^2, x^3)$$

and a new set

$$x'^\mu = x'^\mu(x) \qquad \mu = 0,1,2,3$$

where $x = x^\mu$ denotes the original coordinates.

- This notation is an economical form of

$$x'^\mu = \xi^\mu(x^1, x^2, x^3, x^4) \qquad (\mu = 1,2,\ldots)$$

- where the single-valued, continuously differentiable functions ξ^μ
- assign a new (primed) coordinate (x'^1, x'^2, x'^3, x'^4) to a point of the manifold with old coordinates (x^1, x^2, x^3, x^4).

> This transformation may be abbreviated to $x'^\mu = \xi^\mu(x)$ and, even more tersely, to $x'^\mu = x'^\mu(x)$.

- Coordinates are just *labels,* so laws of physics cannot depend on them. Hence the system x'^μ is not privileged and this transformation should be *invertible.*
- Notice carefully that we are talking about the *same point* described in *two different coordinate systems.*

As introduced in Chapter 2, we shall generally use the Einstein summation convention for 4D spacetime:

- An index that is repeated, once as a superscript and once as a subscript, implies a summation over that index.
- Such an index is a dummy index that is removed by the summation and should not appear on the other side of the equation.
- A repeated (dummy) index may be replaced by any other index not already in use without altering equation: $A_\alpha B^\alpha = A_\beta B^\beta$.
- A superscript (subscript) in a denominator counts as a subscript (superscript) in a numerator.
- *Greek indices* ($\alpha, \beta, \ldots$) denote the full set of spacetime indices running over 0, 1, 2, 3.
- *Roman indices* ($i, j, \ldots$) denote the indices 1, 2, 3 running only over the spatial coordinates.
- Placement of indices matters: generally x^α and x_α will be different quantities.

At all stages of manipulating equations, the indices on the two sides of an equation (including their up or down placement) must match.

As a minimum, we must consider the transformations of

- Fields
- Derivatives of fields
- Integrals of fields.

The first two are necessary to formulate *equations of motion,* and the latter enter into various *conservation laws.*

To facilitate this, we shall introduce a set of mathematical quantities called *tensors* that are a generalization of the idea of scalars and vectors to more components.

As a starting point, we must look more carefully at *how to define vectors in a non-Euclidean (possibly curved) space.*

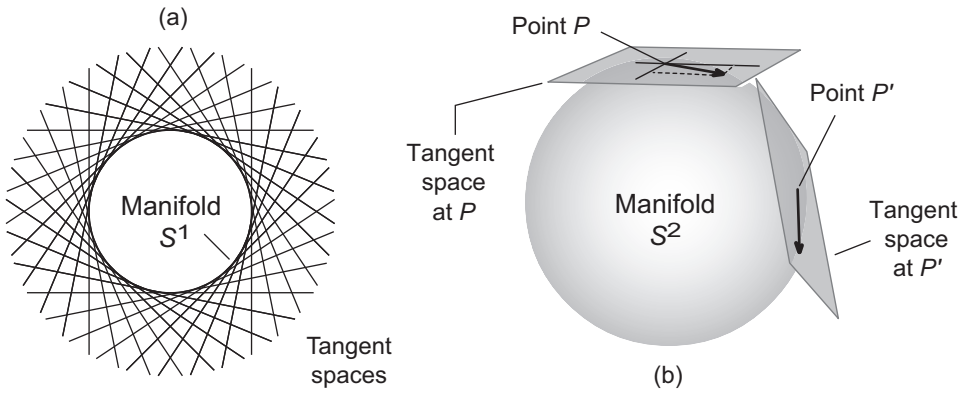
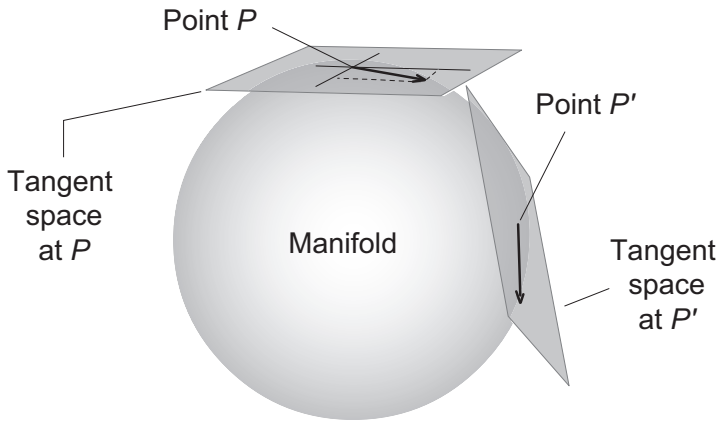


Figure 8.4: Tangent spaces in curved manifolds, illustrated (a) for the manifold S^1 and (b) for the manifold S^2. As illustrated for S^2, vectors (indicated by arrows) are defined in the tangent spaces at each point, not in the curved manifold. Embedding the 1D tangent-space manifold in 2D euclidean 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.

Spacetime is characterized by a manifold that is *not euclidean* [does not obey the rules of high school (Euclidean) geometry].

- *Two reasons:* The manifold may be *curved*, and it has a *non-euclidean metric*, even if it is flat.
- In euclidean space we are used to representing vectors as directed line segments of finite length.
- This fails in curved spacetime, which is *not globally euclidean*, so extended straight lines have no meaning.
- Thus the first question is *how to define a vector* at some spacetime point in a way that is *valid for euclidean and non-euclidean manifolds*.
- *Answer:* Vectors live, not in the manifold, but rather in a *tangent space* that may be visualized as a plane tangent to the point *P* on the curved surface, as illustrated in Fig. 8.4.



- The idea conveyed by the above figure in which planes tangent to a 2D surface are shown embedded in a 3D space is *useful conceptually* but it is *potentially misleading*.
- 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 later.

8.8.3 Tangent and Cotangent Spaces

An n-dimensional Riemannian manifold has at each point P

- An n-dimensional euclidean vector space T_P with a basis defined by the directional derivatives evaluated at P for coordinate curves passing through P.
- This is termed the *tangent space* and *vectors* at P are defined within that space.
- An intrinsically-defined n-dimensional euclidean vector space with a basis defined by viewing the coordinate curves as scalar fields and evaluating their gradients at P.
- This is termed the *cotangent space* T_P^* and *dual vectors* P are defined within this space.
- The tangent space T_P and the cotangent space T_P^* are *dual to each other*.

The definitions given above make clear that

- Vectors and dual vectors are *local to a point*.
- The tangent and cotangent spaces in which they are defined may be *constructed from the properties of the manifold alone*.

> Thus the tangent space and cotangent space at each point of a manifold have an *intrinsic meaning*, independent of embedding in higher dimensions.

8.8.4 Coordinates in Spacetime

A universal coordinate system can be chosen in flat space,

- Basis vectors can be chosen that are mutually *orthogonal and constant*.
- Furthermore, these constant basis vectors can be *normalized to unit length* once and for all.
- Much of ordinary physics is conveniently described using such *orthonormal bases*.
- The situation is more complicated in curved manifolds (or in uncurved manifolds expressed in non-cartesian coordinates).
- Because of the position-dependent metric of curved spacetime it is most convenient in general relativity to choose basis vectors that
  - *depend on position* and that
  - *need not be orthogonal*.
- Since such basis vectors are position-dependent, it usually is *not useful to normalize them*.

8.8.5 Coordinate and Non-Coordinate Bases

The standard conception of a *vector as a directed line segment* is *ill-defined in a curved manifold*.

- The key to specifying vectors in curved space is to *separate the "directed" part from the "line segment" part* of the usual definition.
- This is because the *direction* for vectors of infinitesimal length can be defined consistently in curved or flat spaces using *directional derivatives.*

  - The book-formatted version of these lectures explains *position-dependent coordinates* based on directional derivatives, which are valid in *curved or flat* spacetime.
  - Here in the presentation version I will skip that, since we are going to need only *special relativity* in *flat spacetime*.

8.8.6 Tensors and Coordinate Transformations

In formulating special or general relativity we are interested in how quantities that enter the physical description of the Universe change when the spacetime coordinates are transformed.

- This requires understanding the transformations of
  - *fields*,
  - their *derivatives*, and
  - their *integrals*,
- To facilitate this task, it is useful to introduce a set of mathematical objects called *tensors.*
  - These have a *fundamental definition without reference to specific coordinate systems* (linear maps from vectors and dual vectors to the real numbers).
  - However, for physical applications it often proves convenient to view tensors as *components expressed in a basis that transform in a precise way* if the coordinate system is changed.
- These two approaches were introduced in the earlier discussion of euclidean space.
- The *book-formatted version* of these lectures describes both approaches in the spacetime manifold.
- Here in the *lecture formatted version* we will skip the approach in terms of linear maps and treat tensors as *components in a basis defined by transformation laws*.

The *rank* of a tensor may be given a more fundamental definition, but practically it is the *total number of indices* required to specify its components in some 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.
- As for vectors and dual vectors, the indices may either be upper (*contravariant*) or lower (*covariant*).
  - Tensors carrying only lower indices are termed *covariant tensors*.
  - Tensors carrying only upper indices are termed *contravariant tensors*.
  - Tensors carrying both lower and upper indices are termed *mixed tensors*.
- It is convenient to indicate the *type* of a tensor by the ordered pair (p,q), where when evaluated in a basis
  - p is the number of contravariant (upper) indices,
  - q is the number of covariant (lower) indices,

  and the rank of the tensor is $p+q$.

  Thus a dual vector is a rank-1 tensor of type $(0,1)$ having one covariant index.

Not all quantities with indices are tensor components; it is their *mathematical properties* that mark objects as tensors, not merely that they carry indices when evaluated in a basis.

8.8.7 Tensors Specified by Transformation Laws

In the book-formatted version of these lectures, tensors are introduced at a fundamental level through *linear maps from vectors and dual vectors to the real numbers*.

- However, it is shown there also that
  - when tensors are expressed in an arbitrary basis,
  - 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
  - It is less abstract and requires less new mathematics.
  - A physical interpretation often requires expression of the problem in a well-chosen basis anyway.
  - The component index formalism has a handy built-in error checking mechanism:

    Failure of indices to balance on the two sides of an equation is a sure sign of an error.
- The next sections will summarize the use of tensors to formulate invariant equations by exploiting the transformation properties of their components.

Scalar Transformation Law

Tensors may be viewed as generalizing the idea of scalars and vectors, so let's begin with these more familiar quantities.

Simplest possibility: A field has a single component (magnitude) at each point that is unchanged by the transformation

$$\phi'(x') = \phi(x).$$

Quantities such as $\phi(x)$ that are unchanged under the coordinate transformation are called *scalars*.

> *EXAMPLE:* Value of the temperature at different points on the surface of the Earth.

Vectors and Dual Vectors

Recall also that we can classify tensors by a notation (n, m), where n is the number of upper indices and m is the number of lower indices when evaluated in a basis.

- Thus a scalar is a tensor of type $(0, 0)$, since it carries no indices.
- The sum of n and m is the rank of the tensor. A scalar is a tensor of rank zero.

There are *two kinds of rank-1 tensors,* having the index pattern $(0, 1)$ and $(1, 0)$, respectively. The first is called a *dual vector*, *covariant vector*, or *1-form:*

DUAL VECTOR:

The gradient of a scalar field $\phi(x) = \partial\phi(x)/\partial x$ transforms under change of coordinates as

$$\left(\frac{\partial\phi(x)}{\partial x'^{\mu}}\right) = \frac{\partial x^{\nu}}{\partial x'^{\mu}}\left(\frac{\partial\phi(x)}{\partial x^{\nu}}\right).$$

Remember in such expressions:

- the Einstein summation convention, and
- that all partial derivatives are *understood implicitly* to be evaluated at some point $P = x$.

A tensor having a transformation law that *mimics that of the scalar field gradient*,

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)}$$

is of type $(0,1)$ and is termed a *dual vector* (also *1-form*, *covariant vector*, or *covector*).

ECONOMY OF NOTATION: The preceding equation

$$A'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu(x) \qquad \text{(dual vector)}$$

really means four equations:

$$A'_\mu = \frac{\partial x^0}{\partial x'^\mu} A_0 + \frac{\partial x^1}{\partial x'^\mu} A_1 + \frac{\partial x^2}{\partial x'^\mu} A_2 + \frac{\partial x^3}{\partial x'^\mu} A_3 \quad (\mu = 0,1,2,3)$$

each containing four terms. It is equivalent to the matrix equation

$$\begin{pmatrix} A'_0 \\ A'_1 \\ A'_2 \\ A'_3 \end{pmatrix} = \begin{pmatrix} \frac{\partial x^0}{\partial x'^0} & \frac{\partial x^1}{\partial x'^0} & \frac{\partial x^2}{\partial x'^0} & \frac{\partial x^3}{\partial x'^0} \\ \frac{\partial x^0}{\partial x'^1} & \frac{\partial x^1}{\partial x'^1} & \frac{\partial x^2}{\partial x'^1} & \frac{\partial x^3}{\partial x'^1} \\ \frac{\partial x^0}{\partial x'^2} & \frac{\partial x^1}{\partial x'^2} & \frac{\partial x^2}{\partial x'^2} & \frac{\partial x^3}{\partial x'^2} \\ \frac{\partial x^0}{\partial x'^3} & \frac{\partial x^1}{\partial x'^3} & \frac{\partial x^2}{\partial x'^3} & \frac{\partial x^3}{\partial x'^3} \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix}.$$

VECTORS: A differential dx transforms like

$$dx'^{\mu} = \frac{\partial x'^{\mu}}{\partial x^{\nu}} dx^{\nu},$$

which suggests a second rank-1 transformation rule

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}.$$

A tensor that behaves in this way is of type $(1,0)$ and is termed a *vector* or *contravariant vector*.

Notice carefully the difference between the transformation laws for a vector and a dual vector,

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)},$$

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}.$$

The transformation rules are similar, differing only in

- the vertical placement of the old coordinates x and new coordinates x' in the partial derivatives, and
- the corresponding vertical placement of indices required for consistency in the summation convention.

The dual vector and vector transformation laws,

$$A'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu(x) \qquad \text{(dual vector)}$$

$$A'^\mu(x') = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu(x) \qquad \text{(vector).}$$

may be viewed as matrix equations,

$$A'_\mu(x') = \hat{U}^\nu_\mu A_\nu(x) \qquad A'^\mu(x') = U^\mu_\nu A^\nu(x),$$

with the matrices defined by

$$U \equiv \frac{\partial x'}{\partial x} \qquad \hat{U} \equiv \frac{\partial x}{\partial x'} \qquad \hat{U}U = I,$$

where I is the 4×4 unit matrix. In these transformations

- the matrix U is called the *Jacobian matrix* and
- the matrix $\hat{U}$ is called the *inverse Jacobian matrix*.

Summarizing, we expect the possibility of two rank-1 tensors:

1. *Dual vectors* (also called *one-forms*, *covariant vectors*, or *covectors*), which carry a lower index and transform like the gradient of a scalar:

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)}.$$

2. *Vectors* (also called *contravariant vectors*), which carry an upper index and transform like the coordinate differential:

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}.$$

In the general case they must be distinguished (*by placement—upper or lower—of their indices*).

Scalar Product

Covariant and contravariant indices on vectors permit a scalar product to be defined as

$$A \cdot B \equiv A_\mu B^\mu = A^\mu B_\mu .$$

This transforms as a scalar because from

$$A'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu(x) \qquad A'^\mu(x') = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu(x)$$

we have that

$$\begin{aligned} A' \cdot B' = A'_\mu B'^\mu &= \frac{\partial x^\nu}{\partial x'^\mu} A_\nu \frac{\partial x'^\mu}{\partial x^\alpha} B^\alpha = \frac{\partial x^\nu}{\partial x'^\mu} \frac{\partial x'^\mu}{\partial x^\alpha} A_\nu B^\alpha \\ &= \frac{\partial x^\nu}{\partial x^\alpha} A_\nu B^\alpha = \delta^\nu_\alpha A_\nu B^\alpha \\ &= A_\alpha B^\alpha = A \cdot B, \end{aligned}$$

where the *Kronecker delta* is given by

$$\delta^\mu_\nu = \frac{\partial x'^\mu}{\partial x'^\nu} = \frac{\partial x^\mu}{\partial x^\nu} = \begin{cases} 1 & (\mu = \nu) \\ 0 & (\mu \neq \nu) \end{cases} .$$

Thus $A' \cdot B' = A \cdot B$ and *the scalar product is invariant*.

> Eliminating indices by summing over repeated ones is called *contraction*. The scalar product has no tensor indices left so it is *fully contracted*.

Rank-2 Tensors

Three kinds of rank-2 tensors transform as

$$T'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta},$$

$$T'^{\nu}_{\ \mu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x'^\nu}{\partial x^\beta} T^{\beta}_{\ \alpha},$$

$$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}.$$

This pattern may be generalized to tensors of any rank.

- *Covariant Tensors:* carry only lower indices
- *Contravariant Tensors:* carry only upper indices
- *Mixed Tensors:* carry both upper and lower indices

EXAMPLE: the Kronecker delta δ^ν_μ is a rank-2 mixed tensor.

Handy to recall:

- Upper index μ on left side requires right-side "factor" $\partial x'^\mu / \partial x^\nu$ (prime in numerator).
- Lower index ν on left side requires right-side "factor" $\partial x^\mu / \partial x'^\nu$ (prime in denominator).
- "Vertical position of index on left = vertical position of primed coordinate on right"

Not all quantities carrying indices are tensors! It is

- the *transformation laws* for components in a basis, or
- that they provide *linear maps to the real numbers*

that define tensors.

NOTE: We often employ a standard shorthand by using

- "a tensor $T_{\mu\nu}$" to mean
- "a tensor with components $T_{\mu\nu}$" when evaluated in a basis.

8.8.8 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 part* and an *antisymmetric part:*

$$T_{\alpha\beta} = \tfrac{1}{2}(T_{\alpha\beta} + T_{\beta\alpha}) + \tfrac{1}{2}(T_{\alpha\beta} - T_{\beta\alpha}),$$

  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 we have

$$T_{\alpha\beta} = \pm T_{\beta\alpha} \qquad T^{\alpha\beta} = \pm T^{\beta\alpha} \qquad T_{\alpha}{}^{\beta} = \pm T^{\beta}{}_{\alpha},$$

  where the plus sign holds if the tensor is symmetric and the minus sign if it is antisymmetric.

- More generally, we say that a tensor of rank two or higher is
  - *Symmetric* in any two of its indices if exchanging those indices leaves the tensor invariant and
  - *Antisymmetric* (sometimes termed *skew-symmetric*) in any two indices if it changes sign upon switching those indices.

Metric Tensor

A rank-2 tensor of particular importance is the *metric tensor* $g_{\mu\nu}$ because it is associated with the line element

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu$$

that defines distances in metric spaces. It is symmetric ($g_{\mu\nu} = g_{\nu\mu}$) and satisfies the usual rank-2 transformation rule

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta}.$$

The contravariant components of the metric tensor $g^{\mu\nu}$ are defined by

$$g_{\mu\alpha} g^{\alpha\nu} = \delta^\nu_\mu.$$

(That is, $g_{\mu\nu}$ 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,

$$A^\mu = g^{\mu\nu}A_\nu \qquad A_\mu = g_{\mu\nu}A^\nu$$

$$T^\mu{}_\nu = g_{\nu\alpha}T^{\mu\alpha} \qquad T^\alpha{}_{\beta\gamma} = g^{\alpha\mu}g_{\gamma\varepsilon}T_{\mu\beta}{}^\varepsilon.$$

Thus, the scalar product of vectors may also be expressed

$$A\cdot B = g_{\mu\nu}A^\mu B^\nu \equiv A_\nu B^\nu.$$

In mixed-tensor expressions as above 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.

- This distinction is immaterial if the tensor is *symmetric under exchange of indices* (see following pages).

- However, which index is lowered or raised matters for tensors that are *antisymmetric under index exchange:*

  $$T^\mu{}_\nu = g_{\nu\alpha}T^{\mu\alpha} \qquad T_\nu{}^\mu = g_{\nu\alpha}T^{\alpha\mu}$$

  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.

- However, the preceding discussion 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.

> 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,
- But secretly we will remember that they really are different, and that it is only for metric spaces that this conflation is not likely to land us in trouble.

8.8.9 Summary of Algebraic Tensor Operations

Various algebraic operations are permitted for tensors:

- *Multiplication by a scalar:* For example,

$$aA^{\mu\nu} = B^{\mu\nu},$$

  where a is a scalar and $A^{\mu\nu}$ and $B^{\mu\nu}$ are rank-2 tensors.

- *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.

- *Multiplication:* Tensors may be multiplied by forming products of components. The rank of the resultant tensor will be the sum of the ranks of the factors. Example:

$$A_{\mu\nu} = U_{\mu}V_{\nu},$$

- *Contraction:* For a tensor of type (n,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.

8.8.10 Tensor Calculus in Spacetime

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
  - a prescription to differentiate them.
- *Tensor calculus* is mostly an obvious generalization of normal calculus but complications arise for two reasons:
  - It must be ensured that integration and differentiation *preserve any physical symmetries.*
  - It must be ensured that operations on tensor equations *preserve the tensor structure.*
- We will see that
  - *tensor integration* requires a simple modification of the standard integration rules, but
  - *derivatives of tensors* require a less-simple modification with far-reaching mathematical and physical implications.

  In spacetime, a new *covariant derivative* must be defined. But for *flat spacetime*, the covariant derivative is *just the ordinary partial derivative*.
- Thus we will *omit discussion of the covariant derivative* in these lectures. (It is described in the book-format version.)

Invariant Integration

Change of volume elements for spacetime integration:

$$d^4x = \det\left(\frac{\partial x}{\partial x'}\right) d^4x',$$

where $\det(\partial(x)/\partial(x'))$ is the *Jacobian determinant* of the transformation between the coordinates.

- The metric tensor transforms as

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta} \qquad \text{(triple matrix product).}$$

- Therefore, since:

$$\text{determinant of a product} = \text{product of determinants},$$

- the determinant of the metric tensor $g \equiv \det g_{\mu\nu}$ transforms as

$$g' = \det\left(\frac{\partial x}{\partial x'}\right)\det\left(\frac{\partial x}{\partial x'}\right) g \;\rightarrow\; \det\left(\frac{\partial x}{\partial x'}\right) = \frac{\sqrt{|g'|}}{\sqrt{|g|}},$$

which gives when inserted into the first equation

$$\sqrt{|g|}\, d^4x = \sqrt{|g'|}\, d^4x',$$

($|g|$ because g can be negative in 4-D spacetime).

Therefore, in integrals we shall employ

$$dV = \sqrt{|g|}\, d^4x,$$

as an *invariant volume element*.

Example: The metric for a *2-dimensional spherical manifold (2-sphere)* is specified by the line element

$$ds^2 = g_{ij}dx^i dx^j,$$

which is explicitly in spherical coordinates

$$d\ell^2 = R^2 d\theta^2 + R^2 \sin^2\theta d\phi^2.$$

This may be written as the matrix equation

$$d\ell^2 = (d\theta \quad d\phi) \underbrace{\begin{pmatrix} R^2 & 0 \\ 0 & R^2\sin^2\theta \end{pmatrix}}_{g_{ij}} \begin{pmatrix} d\theta \\ d\phi \end{pmatrix}.$$

The area of the 2-sphere may then be expressed as the *"invariant volume integration"*

$$\begin{aligned} A &= \int \sqrt{|g|}\, d^2x = \int_0^{2\pi} d\phi \int_0^{\pi} \sqrt{\det g_{ij}}\, d\theta \\ &= \int_0^{2\pi} d\phi \int_0^{\pi} R^2 \sin\theta d\theta = 4\pi R^2. \end{aligned}$$

where the metric tensor g_{ij} is the 2×2 matrix in the preceding 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.

8.8.11 Invariant Equations

The properties of tensors elaborated above ensure that

> *Any equation will be invariant under general coordinate transformations* provided that it equates tensors of the same type (*equates components having the same upper and lower indices* when expressed in a basis).

EXAMPLES:

- If $A^{\mu}{}_{\nu}$ and $B^{\mu}{}_{\nu}$ each transform as mixed rank-2 tensors 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 *not in all coordinate systems* because they equate tensors of different kinds:
  - a mixed rank-2 tensor with a scalar in the first case;
  - 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*.
3. Replace all partial derivatives with *covariant derivatives*.
4. Take care to remember that a *covariant differentiation generally does not commute* with a second covariant differentiation.

As will be demonstrated in subsequent chapters, 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

Chapter 9

Special Relativity

To go beyond Newtonian gravitation we must consider, with Einstein, the *intimate relationship between the curvature of space and the gravitational field*.

- Mathematically, this extension is bound inextricably to the *geometry of spacetime*, and in particular to the aspect of geometry that permits quantitative measurement of distances.
- Let us first consider these ideas within the 4-dimensional spacetime termed *Minkowski space*.

As we shall see, requiring *covariance within Minkowski space* will lead us to the *special theory of relativity*.

9.1 The Indefinite Metric of Spacetime

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.

- Familiar examples of metrics were introduced earlier.
- 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 introduced earlier, *fundamentally new features* will enter.

Many of these new features are associated with the *indefinite metric* of Minkowski space.

- Minkowski space is *flat* but it is *not euclidean*, for it does not possess a euclidean metric.
- Many metrics employed in earlier chapters (e.g. for euclidean space) could be put into a diagonal form in which the *signs of the diagonal entries were all positive*.
- Such a metric is termed *positive definite.*
- In contrast, we will see that the Minkowski metric
  - can be put into diagonal form but
  - it is an essential property of Minkowski space that the *diagonal entries cannot all be chosen positive*.

Such a metric is termed *indefinite,* and it leads to properties differing fundamentally from those of euclidean spaces.

9.2 Minkowski Space

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. Any 4-vector A may be expressed in the form,

$$A = A^0 e_0 + A^1 e_1 + A^2 e_2 + A^3 e_3.$$

and the scalar product of 4-vectors is given by

$$A \cdot B = B \cdot A = (A^\mu e_\mu) \cdot (B^\nu e_\nu) = e_\mu \cdot e_\nu A^\mu B^\nu.$$

Note that generally we shall use

- non-bold symbols to denote 4-vectors
- bold symbols for 3-vectors.

We sometimes use a notation such as b^μ to stand generically for all components of a 4-vector.

Defining the metric tensor $\eta_{\mu\nu}$ in Minkowski space,

$$\eta_{\mu\nu} \equiv e_\mu \cdot e_\nu,$$

(it is conventional to denote the metric by $\eta_{\mu\nu}$ rather than $g_{\mu\nu}$ in Minkowski space) the scalar product may be expressed as

$$A \cdot B = \eta_{\mu\nu} A^\mu B^\nu,$$

and the Minkowski-space line element is

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 = \eta_{\mu\nu}dx^\mu dx^\nu,$$

where the *metric tensor of flat spacetime* may be expressed as

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \text{diag}\,(-1,1,1,1).$$

Thus $ds^2 = \eta_{\mu\nu}dx^\mu dx^\nu$ corresponds to the matrix equation

$$ds^2 = (cdt \;\; dx \;\; dy \;\; dz) \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cdt \\ dx \\ dy \\ dz \end{pmatrix},$$

where ds^2 is the spacetime interval between x and $x+dx$ with

$$x = (x^0,x^1,x^2,x^3) = (ct,x^1,x^2,x^3) = (ct,\boldsymbol{x}).$$

- The Minkowski metric is indefinite and is sometimes termed *pseudo-euclidean.*
- As explained below, such metrics are also said to have a *Lorentzian signature.*

Example:

Given a Minkowski vector with components

$$A^\mu = (A^0, A^1, A^2, A^3),$$

what are the components of the corresponding dual vector? Using $\eta_{\mu\nu}$ for the metric tensor, the indices may be lowered through the contraction

$$A_\mu = \eta_{\mu\nu} A^\nu.$$

Therefore, using the Minkowski metric tensor

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

the elements of the dual vector corresponding to the vector A^μ are

$$\begin{aligned} A_\mu &= (\eta_{00} A^0,\, \eta_{11} A^1,\, \eta_{22} A^2,\, \eta_{33} A^3,\,) \\ &= (-1 \times A^0,\, 1 \times A^2,\, 1 \times A^3,\, 1 \times A^3) \\ &= (-A^0, A^1, A^2, A^3). \end{aligned}$$

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.

The Minkowski-space metric

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

is diagonal, with relative sign of the diagonal terms $(-+++)$.

- This sign pattern is termed the *signature of the metric.*
- (Some authors instead define the signature to be an *integer* equal to the difference of the number of positive signs and number of negative signs.)
- It is also common in the literature to see the opposite signature, corresponding to the pattern $(+---)$ that results from multiplying the metric above by -1.
- This choice is *conventional* (no physics depends on it). Metrics with the signature $(-+++)$ or $(+---)$ are sometimes said to be *Lorentzian.*
- However, it is an *essential property of Minkowski space* that it is *not possible to have the same sign* for all terms in the signature of the metric.

The Minkowski metric is indefinite, in contrast to the positive definite metric of euclidean spaces.

9.2.1 Invariance of the Spacetime Interval

Special relativity follows from two assumptions:

- The speed of light is constant for all observers.
- The laws of physics can't depend on coordinates.

The postulate that the speed of light is a constant is equivalent to a statement that

> The spacetime interval ds^2 is an invariant that is *unchanged by transformations between inertial systems* (the Lorentz transformations; see below).

- This invariance does not hold for the euclidean spatial interval $dx^2 + dy^2 + dz^2$,
- nor does it hold for the time interval c^2dt^2.
- Only the particular combination of spatial and time intervals defined by

$$ds^2 = \overbrace{\overbrace{-c^2dt^2}^{\text{not invariant}} \; \overbrace{+dx^2 + dy^2 + dz^2}^{\text{not invariant}}}^{\text{invariant}}$$

  is (Lorentz) invariant.

> Because of this invariance, *Minkowski space is the natural manifold* for special relativity.

Example:

Let's use the metric to determine the relationship between the time coordinate t and the proper time τ, with $\tau^2 \equiv -s^2/c^2$. From

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2,$$

we may write

$$\begin{aligned} d\tau^2 &= \frac{-ds^2}{c^2} = \frac{1}{c^2}(c^2dt^2 - dx^2 - dy^2 - dz^2) \\ &= dt^2 \left\{ 1 - \frac{1}{c^2} \underbrace{\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 \right]}_{v^2} \right\} \\ &= \left(1 - \frac{v^2}{c^2}\right) dt^2. \end{aligned}$$

where v is the magnitude of the ordinary 3-velocity. Therefore, the *proper time* τ that elapses between *coordinate times* t_1 and t_2 is

$$\tau_{12} = \int_{t_1}^{t_2} \left(1 - \frac{v^2}{c^2}\right)^{1/2} dt.$$

The proper time interval τ_{12} *is shorter than the coordinate time interval* $t_2 - t_1$ because the square root is always less than one. If the velocity is constant, this reduces to

$$\Delta\tau = \left(1 - \frac{v^2}{c^2}\right)^{1/2} \Delta t,$$

which is the usual statement of *time dilation in special relativity*.

Table 9.1: Rank 0, 1, and 2 tensor transformation laws

Tensor	Transformation law
Scalar	$\phi' = \phi$
Covariant vector	$A'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu$
Contravariant vector	$A'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu$
Covariant rank-2	$T'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta}$
Contravariant rank-2	$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}$
Mixed rank-2	$T'^{\nu}{}_{\mu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x'^\nu}{\partial x^\beta} T^{\beta}{}_{\alpha}$

9.2.2 Tensors in Minkowski Space

In Minkowski space *transformations between coordinate systems are independent of spacetime.* Thus derivatives appearing in the general definitions of Table 9.1 for tensors are constants and for flat spacetime we have the simplified tensor transformation laws

$$\begin{aligned}
\phi' &= \phi && \text{Scalar}\\
A'^\mu &= \Lambda^\mu{}_\nu A^\nu && \text{Contravariant vector}\\
A'_\mu &= \Lambda_\mu{}^\nu A_\nu && \text{Covariant vector}\\
T'^{\mu\nu} &= \Lambda^\mu{}_\gamma \Lambda^\nu{}_\delta T^{\gamma\delta} && \text{Contravariant rank-2 tensor}\\
T'_{\mu\nu} &= \Lambda_\mu{}^\gamma \Lambda_\nu{}^\delta T_{\gamma\delta} && \text{Covariant rank-2 tensor}\\
T'^{\mu}{}_{\nu} &= \Lambda^\mu{}_\gamma \Lambda_\nu{}^\delta T^{\gamma}{}_{\delta} && \text{Mixed rank-2 tensor}
\end{aligned}$$

where the $\Lambda^\mu{}_\nu$ contain partial derivative factors that *don't depend on the spacetime coordinates*.

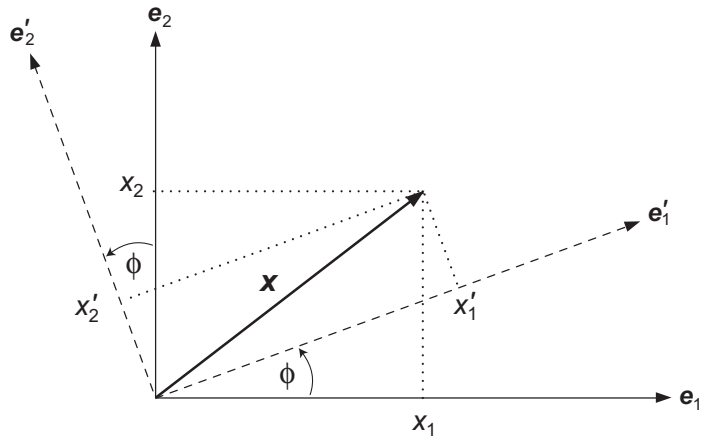
In addition, for flat spacetime we may use a coordinate system for which *covariant derivatives are equivalent to partial derivatives*.

In the Minkowski transformation laws the $\Lambda^{\mu}{}_{\nu}$ are elements of *Lorentz transformations,* to which we now turn our attention.

9.3 Lorentz Transformations

In 3-dimensional euclidean space, rotations are a particularly important class of transformations because they *change the direction for a 3-vector but preserve its length*.

- We wish to generalize this idea to investigate abstract rotations in the 4-dimensional Minkowski space that *change the direction but not the length of 4-vectors*.
- Such rotations in Minkowski space are termed *Lorentz transformations.*



Consider rotation of a 2D euclidean coordinate system (above)

- The length of an arbitrary vector $\boldsymbol{x}$ will be unchanged by this transformation if we require that $\boldsymbol{x} \cdot \boldsymbol{x} = \boldsymbol{x}' \cdot \boldsymbol{x}'$.
- Since $\boldsymbol{x} \cdot \boldsymbol{x} \equiv g_{ij} x^i x^j$, this requires that the transformation matrix R implementing the rotation $x'^i = R^i_{\ j} x^j$ act on the metric tensor g_{ij} in the following way

$$R g_{ij} R^{\mathrm{T}} = g_{ij},$$

  where R^{T} is the transpose of R (switch rows and columns).
- For euclidean space the metric tensor is just the unit matrix so the above requirement reduces to $RR^{\mathrm{T}} = 1$, which is the condition that R be an *orthogonal matrix*.

Thus, we obtain by this somewhat pedantic route the well-known result that *rotations in euclidean space are implemented by orthogonal matrices*.

- But the requirement $Rg_{ij}R^{\mathrm{T}} = g_{ij}$ is *valid generally*, not just for euclidean spaces. Thus, let's use it as guidance to constructing *generalized rotations in Minkowski space*.
- By analogy with the above discussion of rotations in euclidean space, we seek a set of transformations that *leave the length of a 4-vector invariant* in the Minkowski space.
- We write the coordinate transformation in matrix form,

$$dx'^{\mu} = \Lambda^{\mu}{}_{\nu} dx^{\nu},$$

where we expect the transformation matrix $\Lambda^{\mu}{}_{\nu}$ to satisfy the analog of $Rg_{ij}R^{\mathrm{T}} = g_{ij}$ for the Minkowski metric $\eta_{\mu\nu}$,

$$\Lambda\eta_{\mu\nu}\Lambda^{\mathrm{T}} = \eta_{\mu\nu},$$

or explicitly in terms of matrix components,

$$\Lambda_{\mu}{}^{\rho}\Lambda^{\sigma}{}_{\nu}\eta_{\rho\sigma} = \eta_{\mu\nu}.$$

- Let us now use this property to construct the elements of the transformation matrix $\Lambda^{\mu}{}_{\nu}$. These will include
  - *rotations about the spatial axes* (corresponding to rotations within inertial systems) and
  - *transformations between inertial systems* moving at different constant velocities (*Lorentz boosts*).

  We consider first *rotations about the z axis*.

9.3.1 Rotations

For rotations about the z axis

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = R \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix},$$

where a, b, c, and d parameterize the transformation matrix.

- Rotations about a single axis are a 2D problem with euclidean metric, so the condition $Rg_{ij}R^{\mathrm{T}} = g_{ij}$ is

$$\underbrace{\begin{pmatrix} a & b \\ c & d \end{pmatrix}}_{R} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{g_{ij}} \underbrace{\begin{pmatrix} a & c \\ b & d \end{pmatrix}}_{R^{\mathrm{T}}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{g_{ij}},$$

- Carrying out the matrix multiplications 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 comparing the two sides of the equation implies that

$$a^2+b^2 = 1 \qquad c^2+d^2 = 1 \qquad ac+bd = 0.$$

- These requirements are satisfied by the choices

$$a = \cos\phi \qquad b = \sin\phi \qquad c = -\sin\phi \qquad d = \cos\phi,$$

  and we obtain the expected result for an ordinary rotation,

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = R \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}.$$

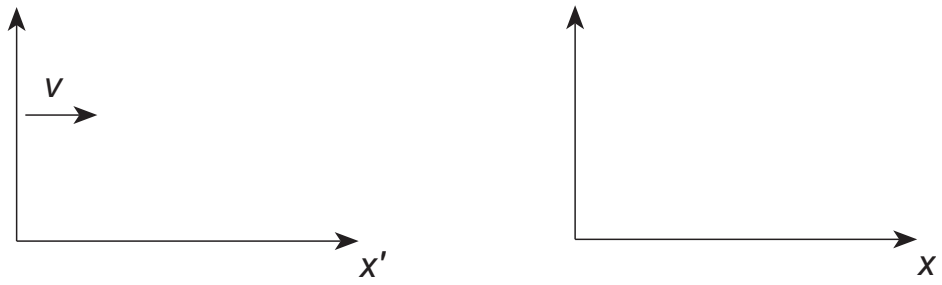


Figure 9.1: A Lorentz boost along the positive x axis.

Now, let's apply this same technique to determine the elements of a *Lorentz boost transformation*.

9.3.2 Lorentz Boosts

Consider a boost from one inertial system to a 2nd one moving at uniform velocity along the x axis (Fig. 9.1).

- The y and z coordinates are unaffected, so this also is effectively a 2-dimensional transformation on t and x,

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} cdt \\ dx \end{pmatrix}$$

- We can write the condition $\Lambda\eta_{\mu\nu}\Lambda^{\mathrm{T}} = \eta_{\mu\nu}$ out as

$$\underbrace{\begin{pmatrix} a & b \\ c & d \end{pmatrix}}_{\Lambda} \underbrace{\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}}_{\eta_{\mu\nu}} \underbrace{\begin{pmatrix} a & c \\ b & d \end{pmatrix}}_{\Lambda^{\mathrm{T}}} = \underbrace{\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}}_{\eta_{\mu\nu}},$$

(identical to rotations, except for the *indefinite metric*).

- Multiplying the matrices on the left side and comparing with the matrix on the right side in

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

gives the conditions

$$a^2 - b^2 = 1 \qquad -c^2 + d^2 = 1 \qquad -ac + bd = 0,$$

- These are satisfied if we choose

$$a = \cosh\xi \qquad b = \sinh\xi \qquad c = \sinh\xi \qquad d = \cosh\xi,$$

where ξ is a hyperbolic variable with $-\infty \leq \xi \leq \infty$.

- Therefore, the boost transformation may be written as

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix} \begin{pmatrix} cdt \\ dx \end{pmatrix} \quad \text{(Lorentz boost).}$$

Which may be compared with the rotational result

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} \quad \text{(Spatial rotation).}$$

The respective derivations make clear that

- the appearance of hyperbolic functions in the boosts, rather than trigonometric functions as in rotations, traces to the role of the indefinite metric

$$g_{\mu\nu} = \text{diag}\,(-1,1)$$

  in the boosts.
- The hyperbolic functions suggest that the boost transformations are *"rotations" in Minkowski space*.
- But these rotations
  - *mix space and time*, and
  - will have unusual properties since they correspond to *rotations through imaginary angles*.
- These unusual properties follow from the metric:
  - The conserved invariant interval is not the length of spatial vectors or time intervals separately.
  - Rather it is the specific *mixture of time and space intervals* implied by the Minkowski line element with *indefinite metric*:

$$ds^2 = (cdt \;\; dx \;\; dy \;\; dz) \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cdt \\ dx \\ dy \\ dz \end{pmatrix}.$$

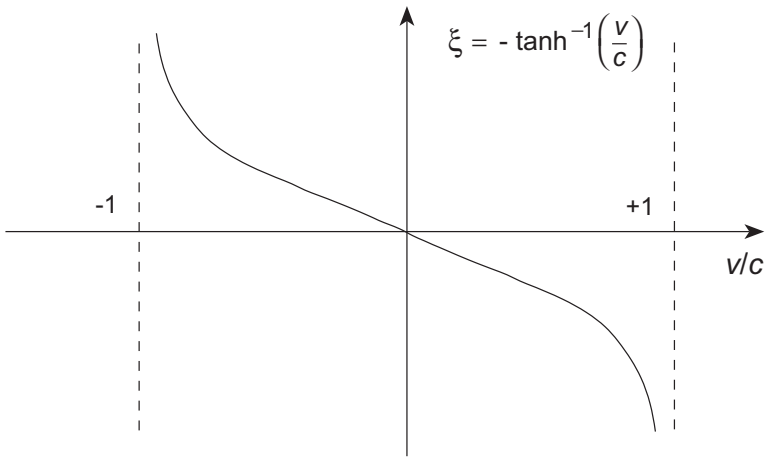


Figure 9.2: Dependence of the Lorentz parameter ξ on $\beta = v/c$.

We can put the Lorentz boost transformation into a more familiar form by relating the boost parameter ξ to the boost velocity.

- Let's work with finite space and time intervals by replacing $dt \to t$ and $dx \to x$ in the preceding equations.
- The velocity of the boosted system is $v = x/t$. From

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix},$$

  the origin ($x' = 0$) of the boosted system is

$$x' = ct\sinh\xi + x\cosh\xi = 0 \quad \to \quad x\cosh\xi = -ct\sinh\xi.$$

- Therefore, $x/t = -c\sinh\xi/\cosh\xi$, so

$$\beta \equiv \frac{v}{c} = \frac{x}{ct} = -\frac{\sinh\xi}{\cosh\xi} = -\tanh\xi.$$

- This relationship between ξ and β is plotted in Fig. 9.2.

- Using the identity $1 = \cosh^2\xi - \sinh^2\xi$, and the definition

$$\gamma \equiv \left(1 - \frac{v^2}{c^2}\right)^{-1/2} = \frac{1}{\sqrt{1 - v^2/c^2}}$$

of the *Lorentz γ factor,* we may write

$$\begin{aligned}\cosh\xi &= \sqrt{\frac{\cosh^2\xi}{1}} = \sqrt{\frac{\cosh^2\xi}{\cosh^2\xi - \sinh^2\xi}} \\ &= \sqrt{\frac{1}{1 - \sinh^2\xi/\cosh^2\xi}} = \frac{1}{\sqrt{1-\beta^2}} \\ &= \frac{1}{\sqrt{1 - v^2/c^2}} = \gamma,\end{aligned}$$

- From this result and

$$\beta = -\frac{\sinh\xi}{\cosh\xi}$$

we obtain

$$\sinh\xi = -\beta\cosh\xi = -\beta\gamma.$$

- Thus, inserting $\cosh\xi = \gamma$ and $\sinh\xi = -\beta\gamma$ in the Lorentz transformation for finite intervals gives

$$\begin{aligned}\begin{pmatrix} ct' \\ x' \end{pmatrix} &= \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix}\begin{pmatrix} ct \\ x \end{pmatrix} \\ &= \begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{pmatrix}\begin{pmatrix} ct \\ x \end{pmatrix} \\ &= = \gamma\begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix}\begin{pmatrix} ct \\ x \end{pmatrix}.\end{aligned}$$

- Writing the matrix expression

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}.$$

  out explicitly for finite intervals gives the Lorentz boost equations (for the specific case of a positive boost along the x axis) in standard textbook form,

$$\begin{aligned} t' &= \gamma\left(t - \frac{vx}{c^2}\right) \\ x' &= \gamma(x - vt) \\ y' &= y \qquad z' = z. \end{aligned}$$

- The inverse transformation corresponds to the replacement $v \to -v$.

- Clearly these reduce to the Galilean boost equations

$$\boldsymbol{x}' = \boldsymbol{x}'(\boldsymbol{x}, t) = \boldsymbol{x} - \boldsymbol{v}t \qquad t' = t'(\boldsymbol{x}, t) = t.$$

  in the limit that v/c vanishes (so $\gamma \to 1$), as we would expect.

It is easily verified that *Lorentz transformations leave invariant the spacetime interval* ds^2.

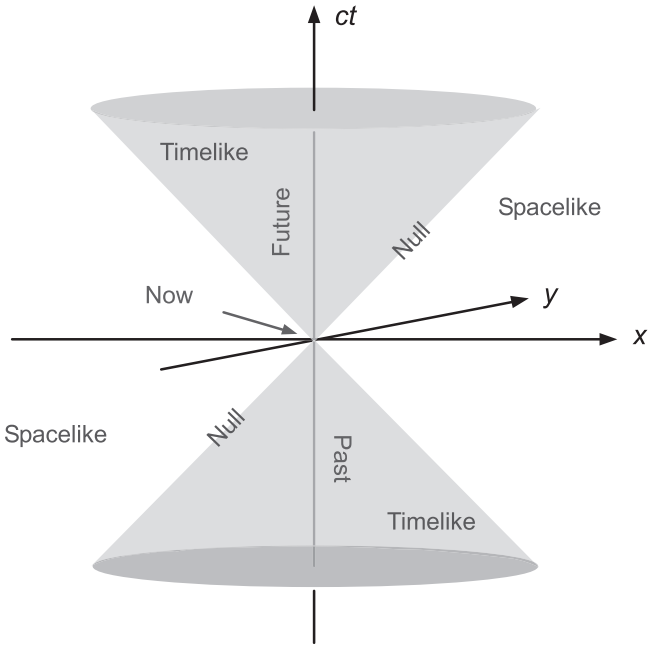


Figure 9.3: The lightcone diagram for two space and one time dimensions.

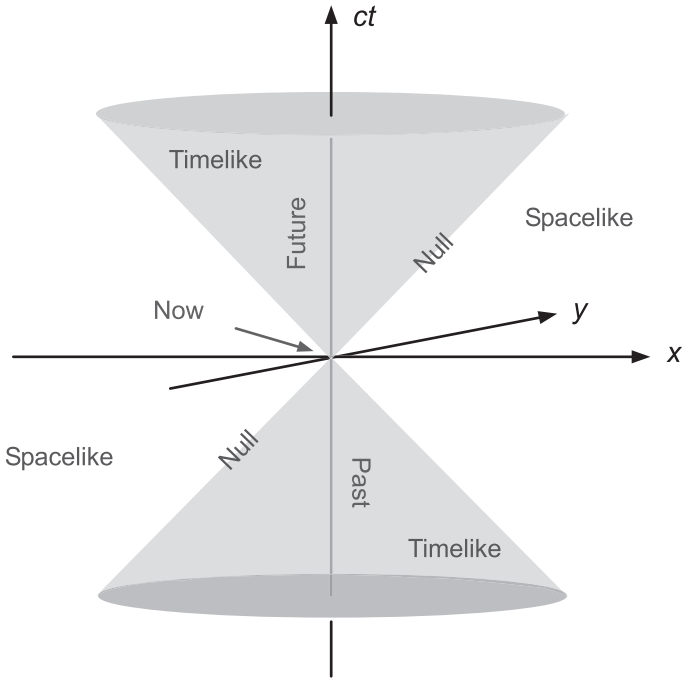
9.3.3 Lightcone Diagrams

By virtue of the line element (which defines a *cone*)

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2,$$

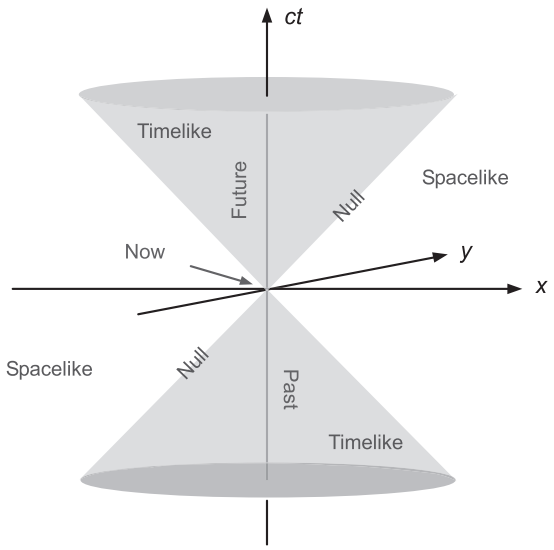
the Minkowski spacetime may be classified according to the lightcone diagram exhibited in Fig. 9.3.

> The lightcone is a 3D surface in 4D spacetime and events and intervals in spacetime may be characterized according to whether they lie inside of, outside of, or on the lightcone.



The standard terminology [*assuming our* $(-1,1,1,1)$ *metric signature*]:

- 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 sometimes *null*).



The lightcone clarifies the distinction between Minkowski spacetime and a 4D euclidean space:

- Two points in Minkowski spacetime are separated by the interval ds defined through

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2.$$

This interval could be

  - positive,
  - negative, or
  - zero,

which embodies *impossibilities for a euclidean space*.

In particular, lightlike particles have worldlines on the lightcone and *the square of the separation of any two points on a lightlike worldline is ZERO.*

Example:

The Minkowski line element in one space and one time dimension [often termed $(1+1)$ dimensions] is $ds^2 = -c^2dt^2 + dx^2$. Thus, if $ds^2 = 0$

$$-c^2dt^2 + dx^2 = 0 \quad \longrightarrow \quad \underbrace{\left(\frac{dx}{dt}\right)^2}_{v^2} = c^2 \quad \longrightarrow \quad v = \pm c.$$

We can generalize this result easily to the full 4D spacetime and we conclude 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 (ct) and space axes have the same scales, the world-line of a freely propagating photon (or any massless particle) always make $\pm 45^\circ$ angles in the lightcone diagram.
- 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 signals with $v > c$ (which would *violate causality*).

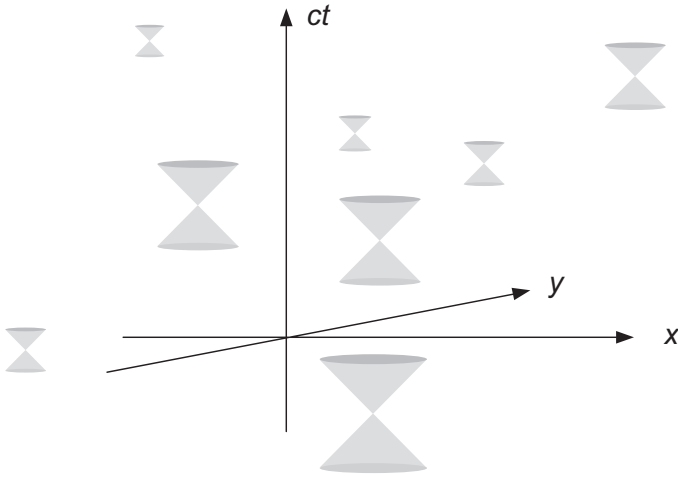


Figure 9.4: Each spacetime point *has its own lightcone*.

We have placed the lightcone in the earlier illustration at the origin of our coordinate system, but we must imagine *a light-cone attached to each point in the spacetime*, as illustrated in Fig. 9.4.

> Lightcones are defined *locally* and *each point of spacetime* has a local lightcone.

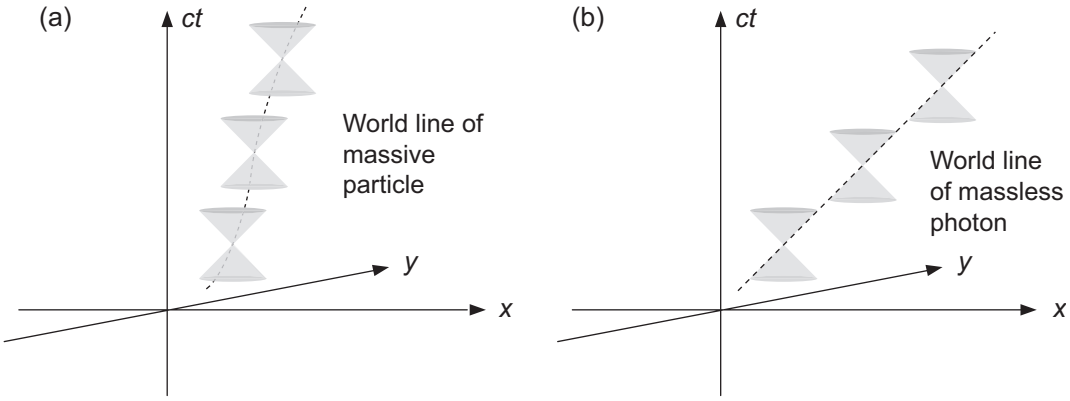


Figure 9.5: Worldlines for (a) massive particles and for (b) massless particles such as photons.

A tangent to the worldline of any particle defines the *local velocity* of the particle and *constant velocity implies straight worldlines*. Therefore, as illustrated in Fig. 9.5,

- In Minkowski space light always travels in straight lines (*not true in in curved space!*), and *always on the lightcone*, since $v = c =$ constant.
- Thus photons have *constant local velocities* equal to c.
- Worldlines for any massive particle lie inside the local lightcone since $v \leq c$ (*timelike trajectory*, since always within the lightcone).
- The worldline for the massive particle in this particular example is curved *(acceleration)*.
- For non-accelerated massive particles the worldline would be straight, but always *within the lightcone.*

Invariance and Simultaneity

- In Galilean relativity, 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 *the Galilean concept of 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 classification of events relative to local lightcones, because *the speed of light is invariant for all observers*.

As we shall now discuss, *The local lightcones define an invariant spacetime structure that may be used to classify events.*

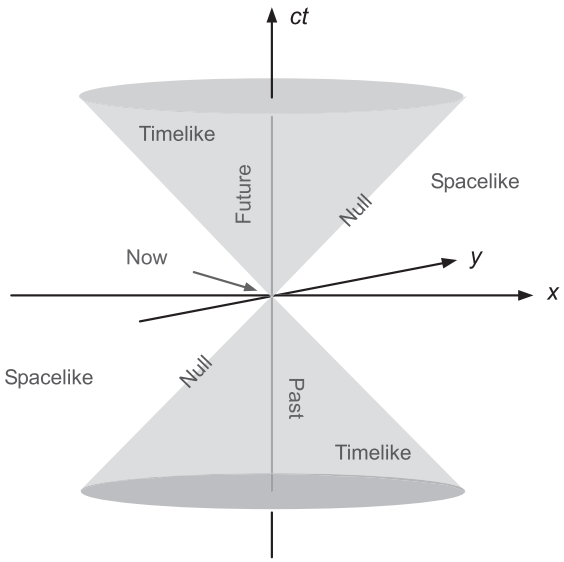


Figure 9.6: The lightcone diagram for two space and one time dimensions.

9.3.4 Causal Structure of Spacetime

The *causal properties* of Minkowski spacetime are *encoded in its light cone structure*, which requires that $v \leq c$ for all signals.

- Each point in spacetime may be viewed as lying at the apex of a lightcone (*"Now"*).
- An event at the origin of a lightcone may influence any event in its forward lightcone (*the "Future"*).
- The event at the origin of the lightcone may be influenced by events in its backward lightcone (*the "Past"*).
- Events at *spacelike separations* are *causally disconnected* from the event at the origin.
- Events *on the lightcone* are *connected by* $v = c$ *signals*.

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 $v \leq c$.

- Velocities are *defined and measured locally*.
- Hence covariant field theories in either flat or curved space are *guaranteed to respect the speed limit* $v \leq c$.
- This is true irrespective of whether *globally velocities* appear to exceed c.

EXAMPLE: In the Hubble expansion of the Universe,

- Galaxies beyond a certain distance (the horizon) would *appear* to recede from us at velocities in excess of c.
- However, all *local measurements* in that expanding, possibly curved, space would determine the velocity of light to be c.

9.3.5 Time Machines and Causality Paradoxes

When time travel comes up it is usually about going *backward* in time.

- Traveling *forward* in time requires no special talent.
- It is easy to arrange various scenarios consistent with relativity where a person could travel into a future time even faster than normal.
- For example, in the *twin paradox* it is possible to arrange for a traveler to arrive back at Earth centuries in the future relative to clocks that remain on Earth.
- Similar options exist using the *gravitational time dilation* of general relativity.

However, the real question is, could you go *back in time* to explore your earlier history?

- No! Not according to current understanding.
- To bend a forward-going timelike worldline continuously into a backward-going one requires *going outside the local lightcone*, requiring that $v > c$.
- If *closed timelike loops* were permitted, travel to earlier times might be possible.
- However, they are forbidden if energy densities are never negative and the Universe has the topology in evidence.

Thus, the determined time traveler has two options:

- Find some negative energy, or
- Find structures with an exotic spacetime topology allowing closed timelike loops.

Unfortunately for the aspiring time traveler,

- negative energy is probably forbidden in classical gravity and
- there is no evidence at present for exotic spacetime topologies with closed timelike loops.

  - These statements are based entirely on *classical gravity considerations;*
  - it is unknown at present whether they could be modified by some future understanding of *quantum gravity*.

From the preceding discussion we may conclude that

- the axioms of special relativity are fundamentally at odds with the Newtonian *concept of absolute simultaneity*, since
- the demand that light have the *same speed for all observers* necessarily means that
- the apparent temporal order of two events *depends upon the observer*.

However, the abolishment of absolute simultaneity

- introduces *no causal ambiguity* because
- all observers agree on the *lightcone structure of spacetime*.
- Thus, for example, all observers will agree that

> Event A can cause event B *only* if A lies in the *past lightcone* of B.

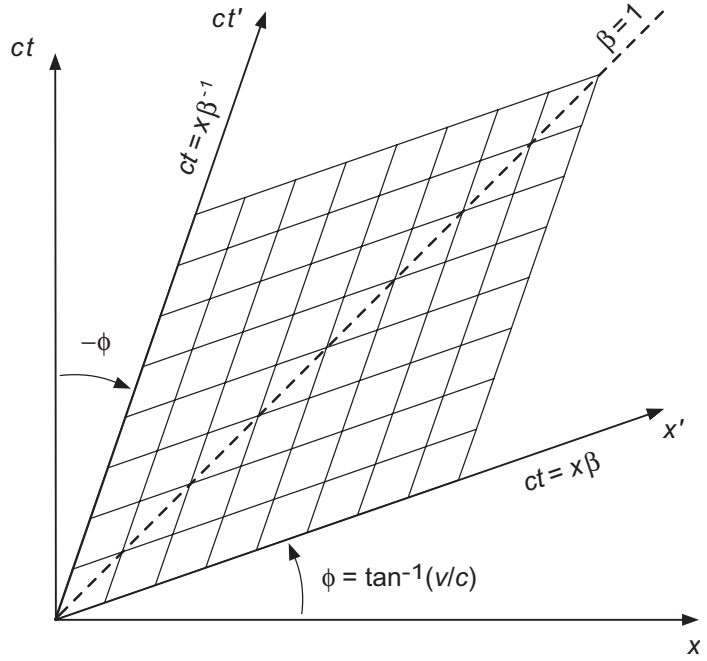
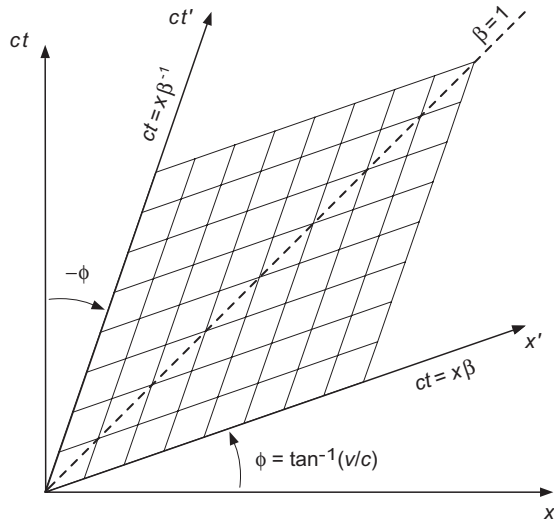


Figure 9.7: Lorentz boost transformation in a spacetime diagram.

9.3.6 Lorentz Transformations in Spacetime Diagrams

It is instructive to look at the action of Lorentz transformations in the spacetime (lightcone) diagram. If we consider 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 in the figure above.



Let's ask what happens to these axes under the Lorentz boost

$$ct' = c\gamma\left(t - \frac{vx}{c^2}\right) \qquad x' = \gamma(x - vt).$$

- *t' axis corresponds to $x' = 0$.* From Eq. 2 with $x' = 0$

$$x = vt \longrightarrow x/c = (v/c)t = \beta t,$$

  so the equation for the t' axis is $ct = x\beta^{-1}$, with $\beta = v/c$.

- Likewise, *the x' axis corresponds to $t' = 0$*, which implies from the 1st equation that $ct = (v/c)x = x\beta$.

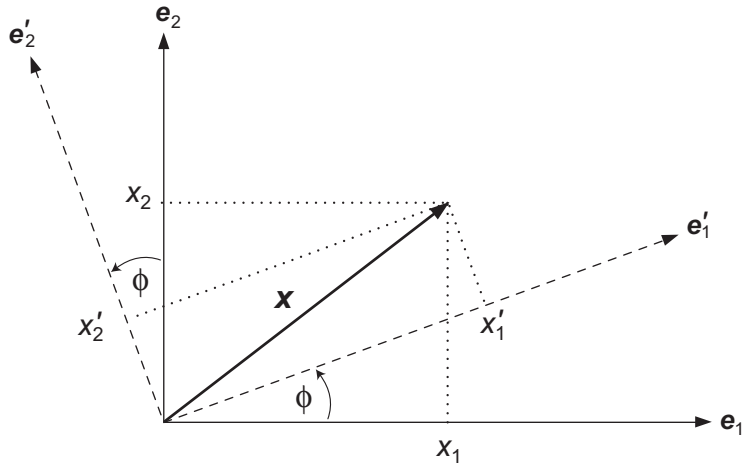
- Thus, the equations of the x' and t' axes [in the (x, ct) coordinate system] are $ct = x\beta$ and $ct = x\beta^{-1}$, respectively.

- The x' and t' axes for the boosted system are also shown in the figure for a positive value of β.

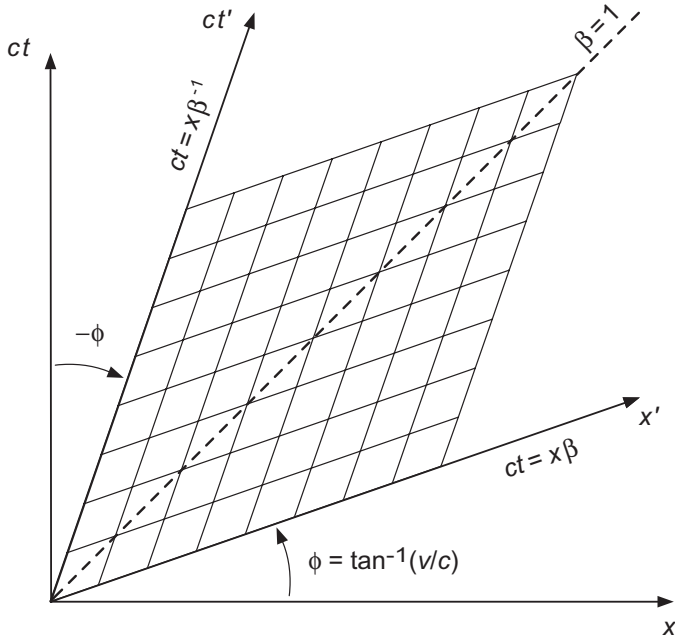
- Time and space axes rotated by same angle, but in *opposite directions* (Cause: the *indefinite Minkowski metric*).

- The rotation angle is given by $\tan\phi = v/c$.

Ordinary rotations: the two axes rotate by the same angle in the *same direction*



Lorentz boost "rotations": the two axes execute *"scissors motion"*, rotating by the *same angle* but in *opposite directions*



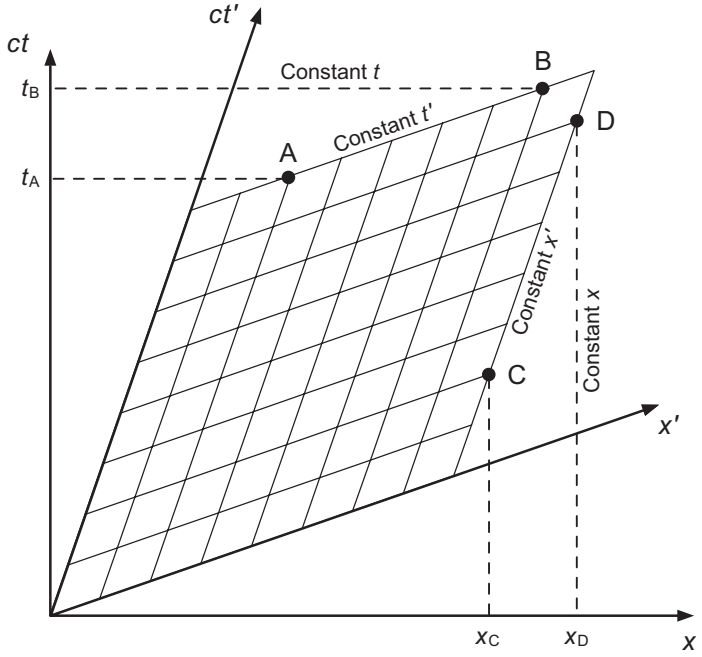


Figure 9.8: Comparison of events in boosted and unboosted frames.

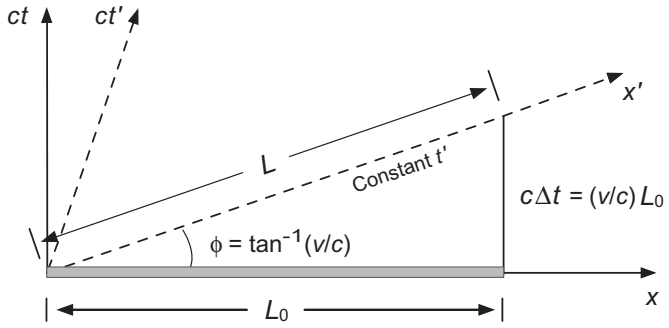
Most of special relativity follows from this diagram.

- For example, relativity of simultaneity is illustrated above.
- 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, *event A occurs before event B in the unboosted frame*.
- 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* follow rather directly from these observations.

Space Contraction

Consider a rod of *proper length* L_0, as measured in its own rest frame (ct, x), that is oriented along the x axis.

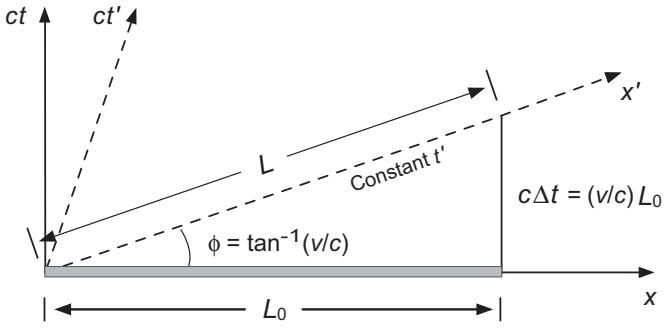


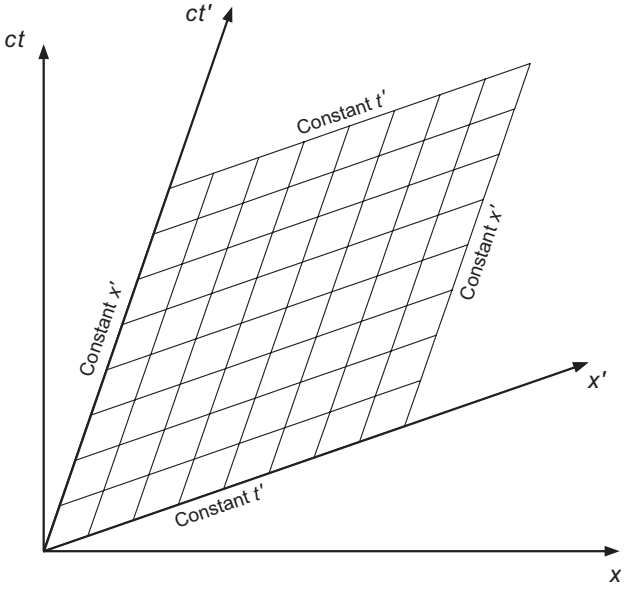
Note: *"Proper"* in relativity denotes a quantity *measured in the rest frame of an object* (proper time, proper length, ...).

What is the length of the rod L as observed in the *boosted* (ct', x') frame? *Fundamental measurement issues:*

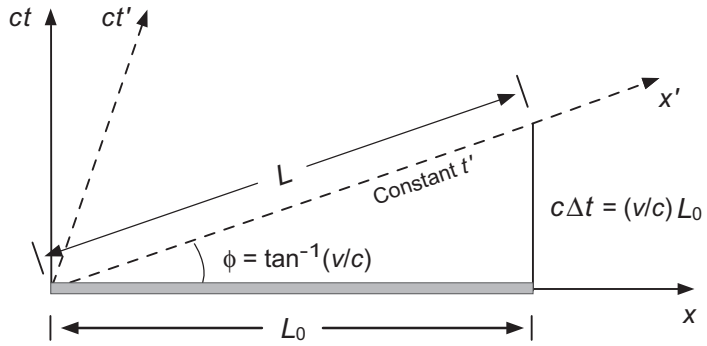
- *Distances* must be measured between spacetime points *at the same time.*
- *Elapsed times* must be measured between spacetime points *at the same place.*

Example: Length of an arrow in flight is not given by the difference between the location of its tip at one time and its tail at a different time. The measurements must be made *at the same time*.





- The frame (ct', x') is boosted by a velocity v along the positive x axis relative to the (ct, 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' has constant t' (bottom figure), so the distance L (top figure) is the *length in the primed frame*.

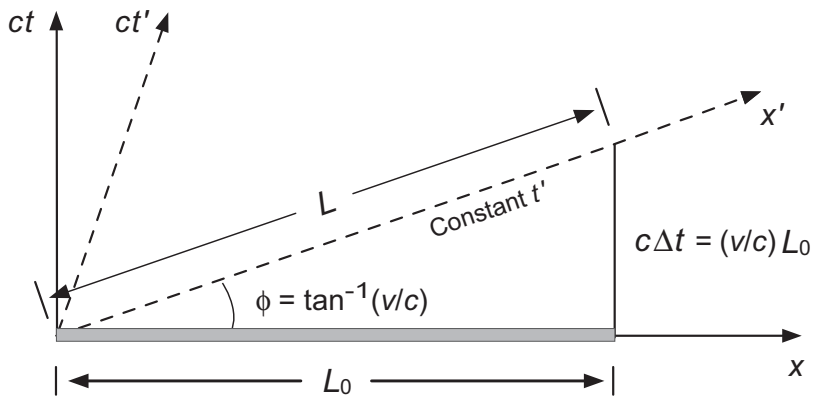


- The distance L *seems* longer than L_0, but this is deceiving because *the diagram is in Minkowski spacetime* but our brain has a *euclidean-space bias*.
- We are familiar with perceived distances being different from actual distances from flat map projections.

Map Projections

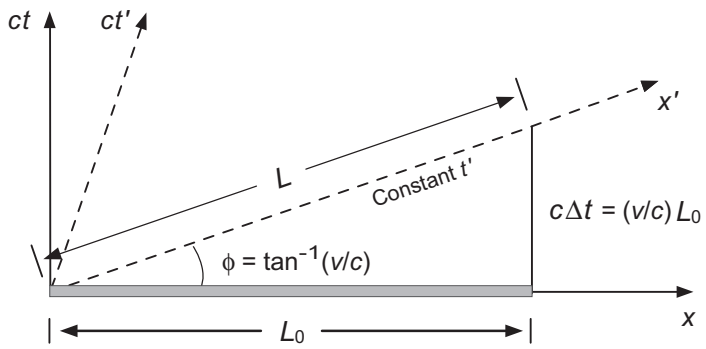
Mercator
(preserves angles, distorts sizes)

Source: *http://www.culturaldetective.com/worldmaps.html*



- A Mercator projection of the globe onto a euclidean sheet of paper gives misleading distance information—Greenland isn't really larger than Brazil, for example,

TRUST THE METRIC,
NOT YOUR LYING EYES!



- From the Minkowski indefinite-metric line element

$$ds^2 = -c^2dt^2 + dx^2.$$

and from the triangle *(Minkowski Pythagorean theorem),*

$$L^2 = L_0^2 - (c\Delta t)^2 \qquad \text{(Minus: indefinite metric)}$$

- But $c\Delta t = (v/c)L_0$ (because from the diagram $\tan\phi = c\Delta t/L_0$ and $\tan\phi = v/c$). Therefore,

$$\begin{aligned} L &= (L_0^2 - (c\Delta t)^2)^{1/2} \\ &= \left(L_0^2 - \left(\frac{v}{c}L_0\right)^2\right)^{1/2} \\ &= L_0\left(1 - \frac{v^2}{c^2}\right)^{1/2} \qquad \text{(SR Length contraction)} \end{aligned}$$

- L is *shorter than* L_0, even though it seems longer in the figure. *TRUST THE METRIC, NOT YOUR LYING EYES!*

Special relativistic *length-contraction* is just the *Pythagorean theorem for Minkowski space*.

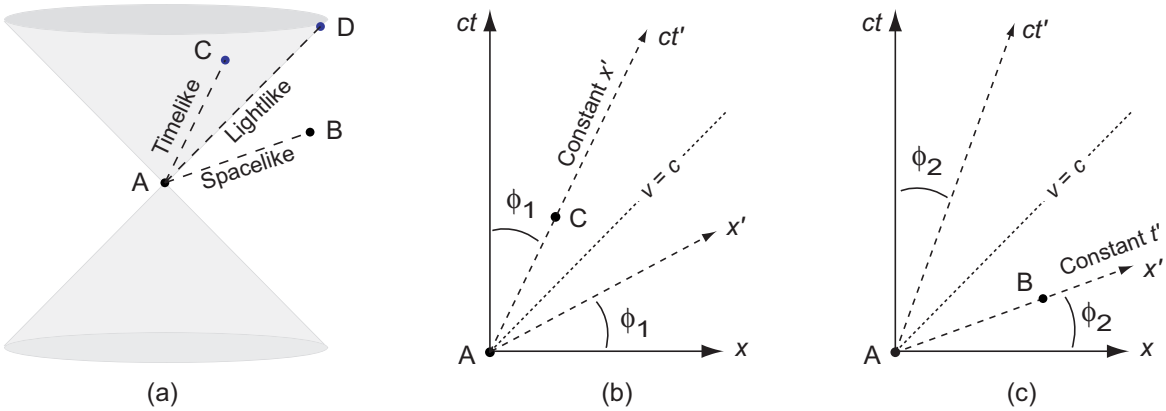
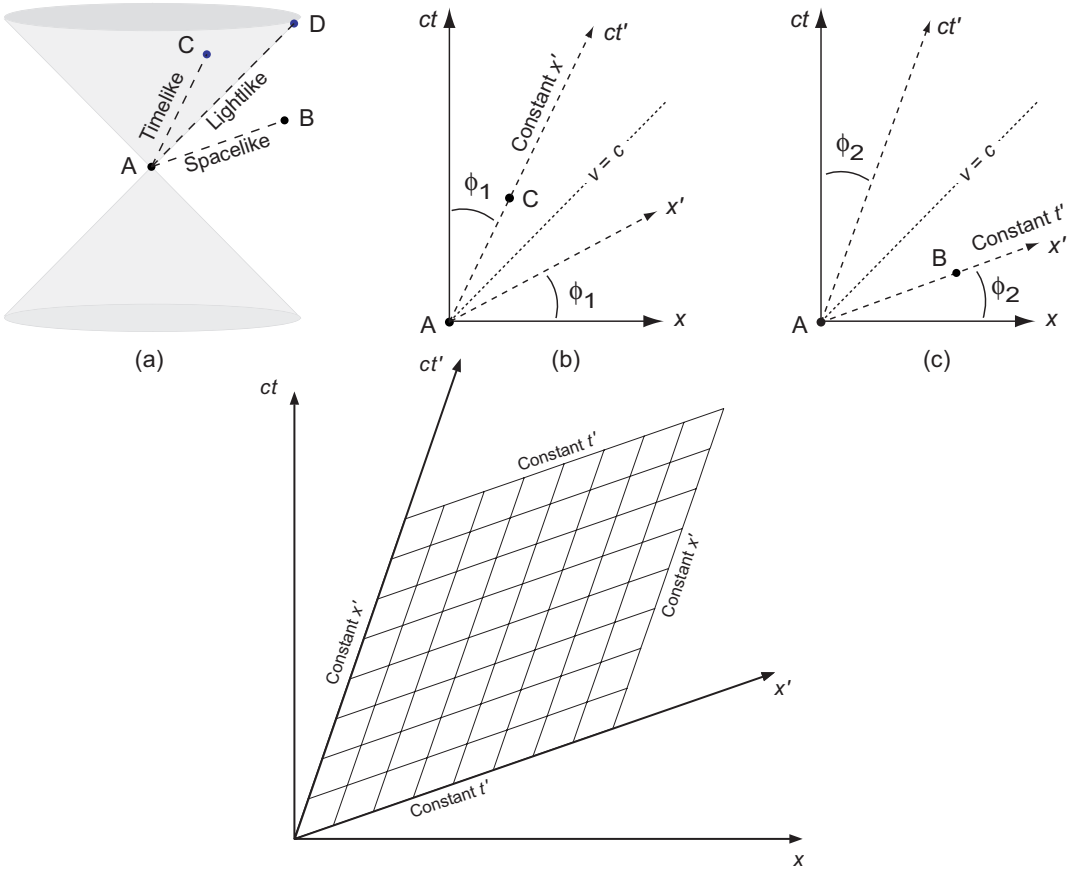


Figure 9.9: (a) Timelike, lightlike (null), and spacelike separations. (b) Lorentz transformation that brings the timelike separated points A and C of (a) into spatial congruence (they lie along a line of constant x' in the primed system). (c) Lorentz transformation that brings the spacelike separated points A and B of (a) into coincidence in time (they lie along a line of constant t' in the primed system.

As we have seen, the spacetime separation between any two events (*spacetime interval*) may be classified in a relativistically invariant way as

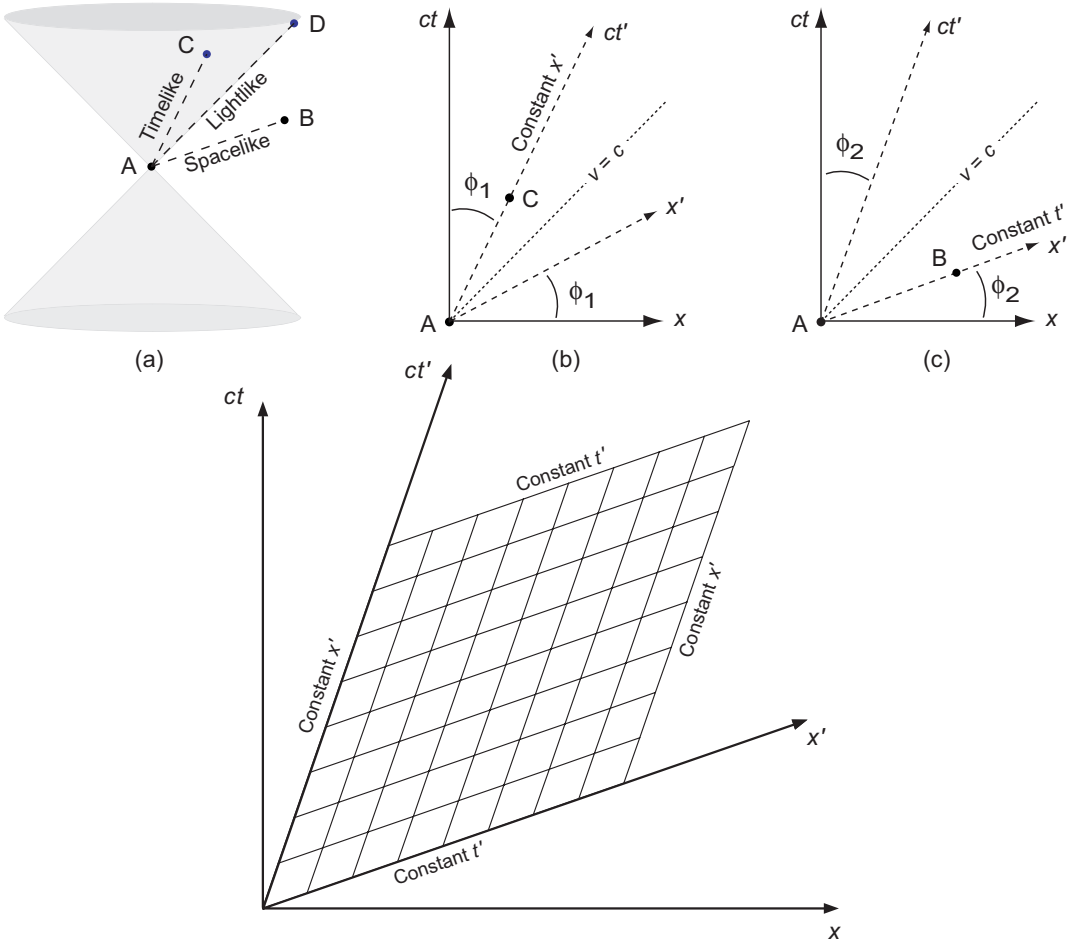
1. *timelike*,
2. *lightlike*, or
3. *spacelike*

by constructing the lightcone at one of the points, as illustrated in Fig. 9.9(a).



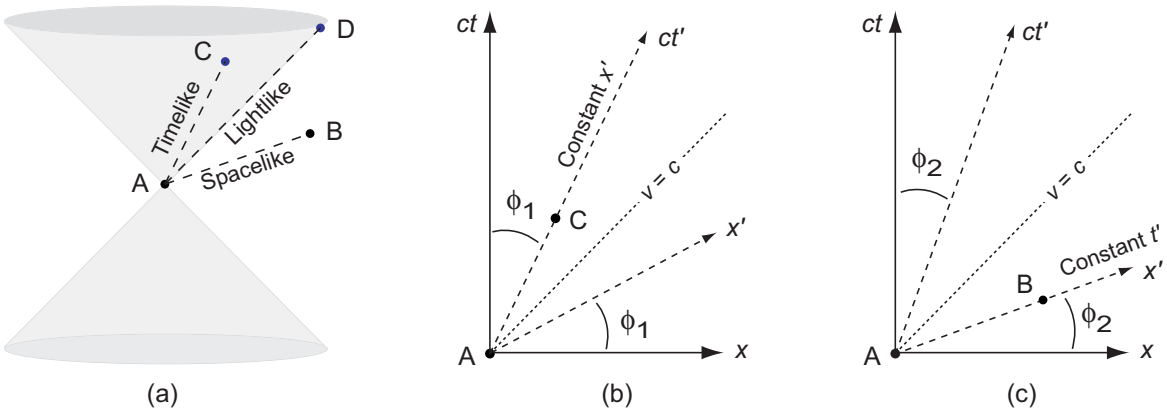
The geometry of the above figures suggests another important distinction between points at spacelike separations [the line AB in Fig. (a)] and timelike separations [the line AC in Fig. (a)]:

- If two events have *timelike separation*, a Lorentz transformation can *bring them into spatial congruence*.
- Figure (b) illustrates a coordinate system (ct', x').
  - in which *A and C have the same coordinate* x'.
  - It is related to the original system by an x-axis Lorentz boost by $v/c = \tan\phi_1$,



On the other hand

- If two events have a *spacelike separation*, a Lorentz transformation exists that can *synchronize the two points*.
- Figure (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'.



Notice that the maximum values of ϕ_1 and ϕ_2 are *limited by the* $v = c$ *line*.

- Thus, the Lorentz transformation to bring point A into spatial congruence with point C
  - exists only if point C lies to the *left of the* $v = c$ *line*
  - and thus is separated by a *timelike interval* from point A.
- Likewise, the Lorentz transformation to synchronize point A with point B
  - exists only if B lies to the *right of the* $v = c$ *line*,
  - meaning that it is separated by a *spacelike interval from A.*

9.4 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.

- It provides a nice example of how useful Lorentz invariance and Lorentz tensors can be.
- The properties of the Maxwell equations influenced Einstein strongly in his development of the special theory of relativity.
- There are many useful parallels between general relativity and the Maxwell theory, particularly for weak gravity where the Einstein field equations may be linearized.

Understanding covariance of the Maxwell equations proves particularly important in discussing gravitational waves.

9.4.1 Maxwell Equations in Non-covariant Form

In free space, using Heaviside–Lorentz, $c = 1$ units, the Maxwell equations may be written as

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{E} &= \rho, \\ \frac{\partial \boldsymbol{B}}{\partial t} + \boldsymbol{\nabla} \times \boldsymbol{E} &= 0, \\ \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0, \\ \boldsymbol{\nabla} \times \boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} &= \boldsymbol{j}, \end{aligned}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, with the charge density ρ and current vector $\boldsymbol{j}$ required to satisfy the equation of continuity

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{j} = 0.$$

Maxwell's equations are consistent with special relativity.

- However, in the above form 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 Minkowski tensors.
- It proves useful to reformulate the Maxwell equations in a manner that is manifestly covariant with respect to Lorentz transformations.

The usual route to accomplishing this begins by replacing the electric and magnetic fields by new variables.

9.4.2 Scalar and Vector Potentials

The electric and magnetic fields may be eliminated in favor of a *vector potential* $\boldsymbol{A}$ and a *scalar potential* ϕ through the definitions

$$\boldsymbol{B} \equiv \boldsymbol{\nabla} \times \boldsymbol{A} \qquad \boldsymbol{E} \equiv -\boldsymbol{\nabla}\phi - \frac{\partial \boldsymbol{A}}{\partial t}.$$

The vector identities

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = 0 \qquad \boldsymbol{\nabla} \times \boldsymbol{\nabla}\phi = 0,$$

may then be used to show that the second and third Maxwell equations are satisfied identically, and the identity

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \boldsymbol{\nabla}^2 \boldsymbol{A},$$

may be used to write the remaining two Maxwell equations as the coupled second-order equations

$$\begin{aligned} \boldsymbol{\nabla}^2\phi + \frac{\partial}{\partial t}\boldsymbol{\nabla} \cdot \boldsymbol{A} &= -\rho \\ \boldsymbol{\nabla}^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{\partial \phi}{\partial t}\right) &= -\boldsymbol{j}. \end{aligned}$$

These equations may then be *decoupled* by exploiting a fundamental symmetry of electromagnetism termed *gauge invariance*.

9.4.3 Gauge Transformations

Because of the identity $\nabla \times \nabla\phi = 0$, the simultaneous transformations

$$\boldsymbol{A} \to \boldsymbol{A} + \nabla\chi \qquad \phi \to \phi - \frac{\partial\chi}{\partial t}$$

for an arbitrary scalar function χ do not change the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields; thus, *they leave the Maxwell equations invariant*.

- These are termed (classical) *gauge transformations.*
- This freedom of gauge transformation may be used to decouple the Maxwell equations.
- For example, if a set of potentials $(\boldsymbol{A}, \phi)$ that satisfy

  $$\nabla \cdot \boldsymbol{A} + \frac{\partial\phi}{\partial t} = 0,$$

  is chosen, the equations decouple to yield

  $$\nabla^2\phi - \frac{\partial^2\phi}{\partial t^2} = -\rho \qquad \nabla^2\boldsymbol{A} - \frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\boldsymbol{j},$$

  which may be solved independently for $\boldsymbol{A}$ and ϕ.
- Such a constraint is called a *gauge condition* and imposing the constraint is termed *fixing the gauge.*

> The particular choice of gauge in the above example is termed the *Lorenz gauge*.

Another common gauge is the *Coulomb gauge*, with a gauge-fixing condition

$$\nabla \cdot \boldsymbol{A} = 0,$$

which leads to the decoupled Maxwell equations

$$\nabla^2 \phi = -\rho \qquad \nabla^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial^2 t} = \nabla \frac{\partial \phi}{\partial t} - \boldsymbol{j}.$$

Let's utilize the shorthand for derivatives introduced earlier:

$$\partial^\mu \equiv \frac{\partial}{\partial x_\mu} = (\partial^0, \partial^1, \partial^2, \partial^3) = \left(-\frac{\partial}{\partial x^0}, \boldsymbol{\nabla}\right),$$
$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = (\partial_0, \partial_1, \partial_2, \partial_3) = \left(\frac{\partial}{\partial x^0}, \boldsymbol{\nabla}\right),$$

where, for example, $\partial^1 = \partial/\partial x_1$ and the 3-divergence is

$$\boldsymbol{\nabla} \equiv (\partial^1, \partial^2, \partial^3).$$

A *covariant formalism* then results from introducing

- the *4-vector potential* A^μ,
- the *4-current* j^μ, and
- the *d'Alembertian operator* $\Box$

through the definitions

$$A^\mu \equiv (\phi, \boldsymbol{A}) = (A^0, \boldsymbol{A}) \qquad j^\mu \equiv (\rho, \boldsymbol{j}) \qquad \Box \equiv \partial_\mu \partial^\mu.$$

Then a gauge transformation takes the form

$$A^\mu \to A^\mu - \partial^\mu \chi \equiv A'^\mu$$

and the preceding examples of gauge-fixing constraints become

$$\partial_\mu A^\mu = 0 \ \text{(Lorenz gauge)} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{A} = 0 \ \text{(Coulomb gauge)}.$$

> The *Lorenz condition is covariant* (formulated in terms of 4-vectors); the *Coulomb gauge condition is not covariant* (formulated in terms of 3-vectors).

The operator $\Box$ is *Lorentz invariant* since

$$\Box' = \partial'_\mu \partial'^\mu = \Lambda^\nu{}_\mu \Lambda_\lambda{}^\mu \partial_\nu \partial^\lambda = \partial_\mu \partial^\mu = \Box.$$

Thus, the Lorenz-gauge wave equation may be expressed in the *manifestly covariant form*

$$\Box A^\mu = j^\mu$$

and the continuity equation becomes

$$\partial_\mu j^\mu = 0.$$

The Maxwell wave equations in Lorenz gauge are *manifestly covariant*.

- This, coupled with the *gauge invariance of electromagnetism*, ensures that the Maxwell equations are covariant in all gauges.
- However—as was seen in the example of the Coulomb gauge—the covariance may not be manifest for a particular choice of gauge.

Let's now see how to formulate the Maxwell equations in a *manifestly covariant form*.

9.4.4 Maxwell Equations in Manifestly Covariant Form

The Maxwell equations may be cast in a manifestly covariant form by constructing the components of the electric and magnetic fields in terms of the potentials (Problems).

- 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*

$$F^{\mu\nu} = -F^{\nu\mu} = \partial^\mu A^\nu - \partial^\nu A^\mu,$$

  which may be expressed in matrix form as

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & -B^3 & B^2 \\ E^2 & B^3 & 0 & -B^1 \\ E^3 & -B^2 & B^1 & 0 \end{pmatrix}.$$

- That is, the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$
  - are *vectors* in 3D euclidean space but
  - their six components together form an antisymmetric *rank-2 tensor* in Minkowski space.

Now let's employ the *Levi–Civita symbol* $\varepsilon_{\alpha\beta\gamma\delta}$, where

- $\varepsilon_{\alpha\beta\gamma\delta}$ 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 use it to define the *dual field tensor* $\mathscr{F}^{\mu\nu}$ by

$$\mathscr{F}^{\mu\nu} \equiv \tfrac{1}{2}\varepsilon^{\mu\nu\gamma\delta}F_{\gamma\delta} = \begin{pmatrix} 0 & -B^1 & -B^2 & -B^3 \\ B^1 & 0 & E^3 & -E^2 \\ B^2 & -E^3 & 0 & E^1 \\ B^3 & E^2 & -E^1 & 0 \end{pmatrix}.$$

- Then *two of the four Maxwell equations* may be written (homework)

$$\partial_\mu F^{\mu\nu} = j^\nu,$$

- and the *other two Maxwell equations* may be written as (homework)

$$\partial_\mu \mathscr{F}^{\mu\nu} = 0.$$

The Maxwell equations in this form are *manifestly covariant* (under Lorentz transformations) because they are formulated *exclusively in terms of Lorentz tensors*.