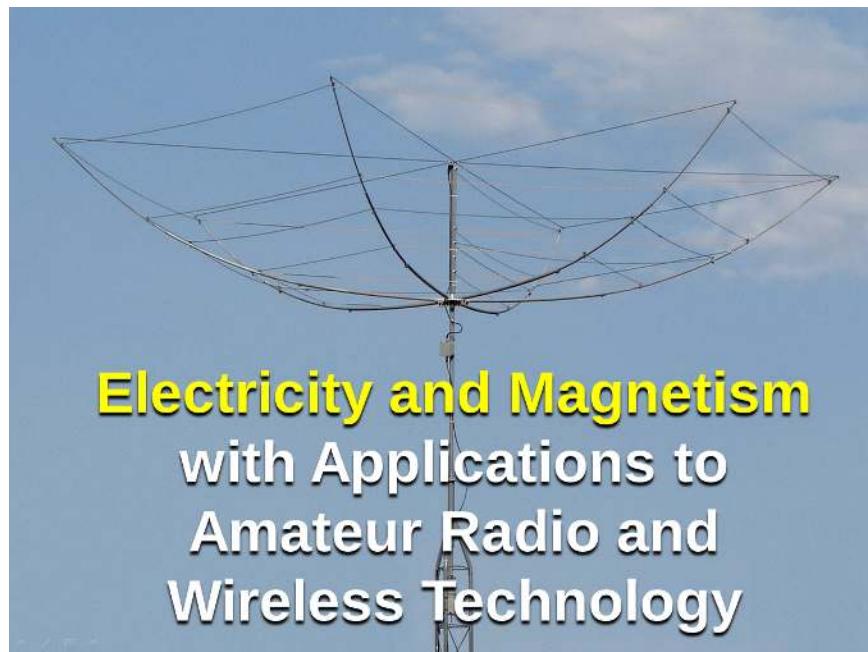


ELECTRICITY AND MAGNETISM WITH APPLICATIONS TO AMATEUR RADIO AND WIRELESS TECHNOLOGY



Ronald Kumon
Kettering University



**Electricity and Magnetism
with Applications to
Amateur Radio and
Wireless Technology**

written and edited by Ronald E. Kumon

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (<https://LibreTexts.org>) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by [NICE CXone Expert](#) and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

This text was compiled on 12/11/2025

TABLE OF CONTENTS

Licensing

Preface

1: Preliminary Concepts

- [1.1: What is Electricity and Magnetism?](#)
- [1.2: Wireless Technology and Amateur Radio - What and Why?](#)
- [1.3: Units](#)
- [1.4: Electromagnetic Spectrum](#)
- [1.5: Amateur Radio Equipment Basics](#)
- [1.6: Notation](#)
- [1.7: Coordinate Systems](#)
- [1.8: Where Do We Go from Here?](#)

2: The Electric Field

- [2.1: Introduction](#)
- [2.2: Electric Charge Model](#)
- [2.3: Conduction and Charging](#)
- [2.4: Electric Fields and Forces](#)
- [2.5: Electric Fields and Forces with Multiple Charges](#)
- [2.6: Electric Field Diagrams](#)
- [2.7: Common Models of Electric Field](#)
- [2.8: Motion of a Charged Particle in an Electric Field](#)
- [2.9: Conclusion](#)
- [2.10: The Electric Field \(Summary\)](#)
- [2.11: The Electric Field \(Exercises\)](#)
- [2.12: The Electric Field \(Answers\)](#)

3: The Electric Potential

- [3.1: Introduction](#)
- [3.2: Work and Energy](#)
- [3.3: Electric Potential Energy](#)
- [3.4: Electric Potential Energy of Point Charges](#)
- [3.5: Electric Potential](#)
- [3.6: Electric Potential of a Point Charge](#)
- [3.7: Common Models of Electric Potential](#)
- [3.8: Electric Potential \(Summary\)](#)
- [3.9: The Electric Potential \(Exercises\)](#)
- [3.10: The Electric Potential \(Answers\)](#)

4: Potential and Field Relationships

- [4.1: Electric Potential from Electric Field](#)
- [4.2: Electric Field from Electric Potential](#)
- [4.3: Equipotential Curves and Surfaces](#)

- 4.4: Conductors in Electrostatic Equilibrium
- 4.5: Applications of Electric Potential and Conductors in Electrostatic Equilibrium
- 4.6: Potential and Field Relationships (Summary)
- 4.7: Potential and Field Relationships (Exercises)
- 4.8: Potential and Field Relationships (Answers)

5: Electric Current and Resistance

- 5.1: Introduction
- 5.2: Electric Current
- 5.3: Basic Model of Conduction in Metals
- 5.4: Resistivity and Resistance
- 5.5: Ohm's Law
- 5.6: Electrical Energy and Power
- 5.7: Alternating Current versus Direct Current
- 5.8: Current and Resistance (Summary)
- 5.9: Current and Resistance (Exercises)
- 5.10: Current and Resistance (Answers)

6: Direct-Current (DC) Resistor Circuits

- 6.1: Introduction
- 6.2: Source Voltage
- 6.3: Kirchhoff's Rules
- 6.4: Resistors in Series and Parallel
- 6.5: Real Batteries
- 6.6: Multi-loop Circuits
- 6.7: Application - Electrical Meters
- 6.8: Application - Grounding and Electrical Safety
- 6.9: Direct-Current Circuits (Summary)
- 6.10: Direct-Current Circuits (Exercise)
- 6.11: Direct-Current Circuits (Answers)

7: Capacitance

- 7.1: Introduction
- 7.2: Capacitors and Capacitance
- 7.3: Capacitors in Series and in Parallel
- 7.4: Electrical Energy Stored in a Capacitor
- 7.5: Capacitor with a Dielectric
- 7.6: Molecular Model of a Dielectric
- 7.7: Application - RC Circuits
- 7.8: Application - RC Circuits with AC
- 7.9: Capacitance (Summary)
- 7.10: Capacitance (Exercises)
- 7.11: Capacitance (Answers)

8: The Magnetic Field

- 8.1: Introduction
- 8.2: Introduction to Magnetism
- 8.3: Magnetism and Its Historical Discoveries

- [8.4: The Biot-Savart Law](#)
- [8.5: Common Magnetic Field Models](#)
- [8.6: Magnetic Fields and Lines](#)
- [8.7: Motion of a Charged Particle in a Magnetic Field](#)
- [8.8: Magnetic Force on a Current-Carrying Conductor](#)
- [8.9: Magnetism in Matter](#)
- [8.10: The Magnetic Field \(Summary\)](#)
- [8.11: Magnetic Forces and Fields \(Exercise\)](#)
- [8.12: Sources of Magnetic Fields \(Exercise\)](#)
- [8.13: Magnetic Forces and Fields \(Answers\)](#)
- [8.14: Sources of Magnetic Fields \(Answers\)](#)

9: Electromagnetic Induction

- [9.1: Introduction](#)
- [9.2: Magnetic Flux](#)
- [9.3: Faraday's Law](#)
- [9.4: Lenz's Law](#)
- [9.5: Motional Source Voltage](#)
- [9.6: Induced Electric Fields](#)
- [9.7: Eddy Currents](#)
- [9.8: Electric Generators and Back Source Voltage](#)
- [9.9: Transformers](#)
- [9.10: Other Applications of Electromagnetic Induction](#)
- [9.11: Electromagnetic Induction \(Summary\)](#)
- [9.12: Electromagnetic Induction \(Exercises\)](#)
- [9.13: Electromagnetic Induction \(Answers\)](#)

10: Inductance

- [10.1: Inductance](#)
- [10.2: Mutual Inductance](#)
- [10.3: Self-Inductance and Inductors](#)
- [10.4: Energy in a Magnetic Field](#)
- [10.5: RL Circuits](#)
- [10.6: Oscillations in an LC Circuit](#)
- [10.7: RLC Series Circuits](#)
- [10.8: Basic Radio Circuits](#)
- [10.9: Application - RL Circuits with AC](#)
- [10.10: Application - RLC Series Circuits with AC](#)
- [10.11: Inductance \(Summary\)](#)
- [10.12: Inductance \(Exercise\)](#)
- [10.13: Inductance \(Answers\)](#)

11: Electromagnetic Waves

- [11.1: Introduction](#)
- [11.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed](#)
- [11.3: Energy Carried by Electromagnetic Waves](#)
- [11.4: The Electromagnetic Spectrum](#)
- [11.5: Polarization](#)

- [11.6: Electromagnetic Waves \(Summary\)](#)
- [11.7: Electromagnetic Waves \(Exercises\)](#)
- [11.8: Electromagnetic Waves \(Answer\)](#)

12: Antenna Systems

- [12.1: Introduction](#)
- [12.2: Production of Electromagnetic Waves](#)
- [12.3: Transmission Lines and Characteristic Impedance](#)
- [12.4: Finite-length Transmission Lines](#)
- [12.5: "Long" and "Short" Transmission Lines](#)
- [12.6: Standing Waves and Resonance](#)
- [12.7: Antenna Systems \(Summary\)](#)

13: Propagation of Electromagnetic Waves

- [13.1: Introduction](#)
- [13.2: Ray and Wave Models of Propagation](#)
- [13.3: Reflection of Rays](#)
- [13.4: Refraction of Rays](#)
- [13.5: Application- Line-of-Sight Transmission](#)
- [13.6: Diffraction of Waves](#)
- [13.7: Interference of Waves](#)
- [13.8: Double-Slit Interference](#)
- [13.9: Propagation of Electromagnetic Waves \(Summary\)](#)
- [13.10: Propagation of Electromagnetic Waves \(Exercises\)](#)
- [13.11: Propagation of Electromagnetic Waves \(Answers\)](#)

14: Introduction to Semiconductor Devices

- [14.1: Introduction](#)
- [14.2: Band Theory of Solids](#)
- [14.3: Semiconductors and Doping](#)
- [14.4: Introduction to Semiconductor Devices](#)
- [14.5: Junction Diodes](#)
- [14.6: Light Emitting Diode](#)
- [14.7: Solar Cells](#)
- [14.8: Bipolar Junction Transistors](#)
- [14.9: Junction Field-effect Transistors](#)

15: Part 2 - Detailed and/or Advanced Content

16: Direct Calculation of Electrical Quantities from Charge Distributions

- [16.1: Introduction](#)
- [16.2: Electric Dipoles](#)
- [16.3: Calculating Electric Fields of Charge Distributions](#)
- [16.4: Calculating Electric Potential of Charge Distributions](#)
- [16.5: Direct Calculation of Electrical Quantities from Charge Distributions \(Summary\)](#)
- [16.6: Direct Calculation of Electrical Quantities from Charge Distributions \(Exercises\)](#)
- [16.7: Direct Calculation of Electrical Quantities from Charge Distributions \(Answers\)](#)

17: Gauss's Law for Calculation of Electrical Field from Charge Distributions

- [17.1: Introduction to Gauss's Law](#)
- [17.2: Electric Flux](#)
- [17.3: Gauss's Law](#)
- [17.4: Calculating Electric Field Using Gauss's Law](#)
- [17.5: Conductors in Electrostatic Equilibrium via Gauss's Law](#)
- [17.6: Gauss's Law \(Summary\)](#)
- [17.7: Gauss's Law \(Exercises\)](#)
- [17.8: Gauss's Law \(Answers\)](#)

18: Calculation of Magnetic Quantities from Currents

- [18.1: Introduction](#)
- [18.2: Magnetic Field due to a Thin Straight Wire](#)
- [18.3: Magnetic Field of a Current Loop](#)
- [18.4: Magnetic Field using Ampère's Law](#)
- [18.5: Magnetic Field of Solenoids and Toroids](#)
- [18.6: Magnetic Force between Two Parallel Currents](#)
- [18.7: \(edit\) Magnetic Force and Torque on a Current Loop - Motors and Meters](#)
- [18.8: Magnetic Forces in a Conductor - The Hall Effect](#)
- [18.9: More Applications of Magnetism](#)
- [18.10: Superconductors](#)
- [18.11: Conclusion](#)
- [18.12: Magnetic Forces and Fields \(Summary\)](#)
- [18.13: Sources of Magnetic Fields \(Summary\)](#)
- [18.14: Current and Resistance \(Summary\)](#)
- [18.15: Magnetic Forces and Fields \(Exercise\)](#)
- [18.16: Sources of Magnetic Fields \(Exercise\)](#)
- [18.17: Magnetic Forces and Fields \(Answers\)](#)
- [18.18: Sources of Magnetic Fields \(Answers\)](#)

19: Alternating-Current (AC) Circuits

- [19.1: Introduction](#)
- [19.2: AC Sources](#)
- [19.3: Simple AC Circuits](#)
- [19.4: RLC Series Circuits with AC](#)
- [19.5: Power in an AC Circuit](#)
- [19.6: Resonance in an AC Circuit](#)
- [19.7: AC Safety - Grounding and Bonding](#)
- [19.8: Alternating-Current Circuits \(Summary\)](#)
- [19.9: Alternating-Current Circuits \(Exercise\)](#)
- [19.10: Alternating-Current Circuits \(Answers\)](#)

20: Maxwell's Equations

- [20.1: Introduction](#)
- [20.2: Electric Flux](#)
- [20.3: Gauss's Law](#)
- [20.4: Ampère's Law](#)

- 20.5: Maxwell's Equations and Electromagnetic Waves
- 20.6: Plane Electromagnetic Waves
- 20.7: Momentum and Radiation Pressure

21: Electrical Transmission Lines

- 21.1: Introduction
- 21.2: Phasors
- 21.3: Introduction to Transmission Lines
- 21.4: Types of Transmission Lines
- 21.5: Transmission Lines as Two-Port Devices
- 21.6: Lumped-Element Model
- 21.7: Telegrapher's Equations
- 21.8: Wave Equation for a Transmission Line
- 21.9: Characteristic Impedance of a Transmission Line
- 21.10: Wave Propagation on a Transmission Line
- 21.11: Lossless and Low-Loss Transmission Lines
- 21.12: Voltage Reflection Coefficient
- 21.13: Standing Waves
- 21.14: Standing Wave Ratio
- 21.15: Parallel Wire Transmission Line
- 21.16: Attenuation in Coaxial Cable
- 21.17: Power Handling Capability of Coaxial Cable
- 21.18: Why 50 Ohms?
- 21.19: Conclusion

22: Generation and Detection of Electromagnetic Waves

- 22.1: Introduction
- 22.2: Production of Electromagnetic Waves - The Antenna
- 22.3: Radiation from a Current Moment
- 22.4: Radiation from an Electrically-Short Dipole
- 22.5: Far-Field Radiation from a Half-Wave Dipole
- 22.6: Equivalent Circuit Model for Transmission; Radiation Efficiency
- 22.7: Equivalent Circuit Model for Reception
- 22.8: Potential Induced in a Dipole
- 22.9: Decibel Scale for Power Ratio
- 22.10: Antenna Radiation Patterns, Directivity, and Gain
- 22.11: Friis Transmission Equation

23: Signal Modulation

- 23.1: Introduction
- 23.2: Historical Context - The Origins of Radio Communication
- 23.3: Radio Signal Metrics
- 23.4: Modulation Overview
- 23.5: Analog Modulation
- 23.6: Digital Modulation
- 23.7: Frequency Shift Keying, FSK
- 23.8: Carrier Recovery
- 23.9: Phase Shift Keying Modulation

- [23.10: Quadrature Amplitude Modulation](#)
- [23.11: Digital Modulation Summary](#)
- [23.12: References](#)
- [23.13: Exercises](#)

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

[Detailed Licensing](#)

CHAPTER OVERVIEW

11: Electromagnetic Waves

- [11.1: Introduction](#)
- [11.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed](#)
- [11.3: Energy Carried by Electromagnetic Waves](#)
- [11.4: The Electromagnetic Spectrum](#)
- [11.5: Polarization](#)
- [11.6: Electromagnetic Waves \(Summary\)](#)
- [11.7: Electromagnetic Waves \(Exercises\)](#)
- [11.8: Electromagnetic Waves \(Answer\)](#)

11: Electromagnetic Waves is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

11.1: Introduction

Our view of objects in the sky at night, the warm radiance of sunshine, the sting of sunburn, our cell phone conversations, and the X-rays revealing a broken bone—all are brought to us by electromagnetic waves. It would be hard to overstate the practical importance of electromagnetic waves, through their role in vision, through countless technological applications, and through their ability to transport the energy from the Sun through space to sustain life and almost all of its activities on Earth.



Figure 11.1.16: The pressure from sunlight predicted by Maxwell's equations helped produce the tail of Comet McNaught. (credit: modification of work by Sebastian Deiries—ESO)

Theory predicted the general phenomenon of electromagnetic waves before anyone realized that light is a form of an electromagnetic wave. In the mid-nineteenth century, James Clerk Maxwell formulated a single theory combining all the electric and magnetic effects known at that time. Maxwell's equations, summarizing this theory, predicted the existence of electromagnetic waves that travel at the speed of light. His theory also predicted how these waves behave, and how they carry both energy and momentum. The tails of comets, such as Comet McNaught in Figure 16.1, provide a spectacular example. Energy carried by light from the Sun warms the comet to release dust and gas. The momentum carried by the light exerts a weak force that shapes the dust into a tail of the kind seen here. The flux of particles emitted by the Sun, called the solar wind, typically produces an additional, second tail, as described in detail in this chapter.

In this chapter, we explain Maxwell's theory and show how it leads to his prediction of electromagnetic waves. We use his theory to examine what electromagnetic waves are, how they are produced, and how they transport energy and momentum. We conclude by summarizing some of the many practical applications of electromagnetic waves.

This page titled [11.1: Introduction](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via source content that was edited to the style and standards of the LibreTexts platform.

- [16.1: Prelude to Electromagnetic Waves](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

11.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed

Learning Objectives

By the end of this section, you will be able to:

- Restate Maxwell's equations.

The Scotsman James Clerk Maxwell (1831–1879) is regarded as the greatest theoretical physicist of the 19th century. (See Figure 1.) Although he died young, Maxwell not only formulated a complete electromagnetic theory, represented by **Maxwell's equations**, he also developed the kinetic theory of gases and made significant contributions to the understanding of color vision and the nature of Saturn's rings.



Figure 11.2.1: James Clerk Maxwell, a 19th-century physicist, developed a theory that explained the relationship between electricity and magnetism and correctly predicted that visible light is caused by electromagnetic waves. (credit: G. J. Stodart)

Maxwell brought together all the work that had been done by brilliant physicists such as Oersted, Coulomb, Gauss, and Faraday, and added his own insights to develop the overarching theory of electromagnetism. Maxwell's equations are paraphrased here in words because their mathematical statement is beyond the level of this text. However, the equations illustrate how apparently simple mathematical statements can elegantly unite and express a multitude of concepts—why mathematics is the language of science.

MAXWELL'S EQUATIONS

1. **Electric field lines** originate on positive charges and terminate on negative charges. The electric field is defined as the force per unit charge on a test charge, and the strength of the force is related to the electric constant ϵ_0 , also known as the permittivity of free space. From Maxwell's first equation we obtain a special form of Coulomb's law known as Gauss's law for electricity.
2. **Magnetic field lines** are continuous, having no beginning or end. No magnetic monopoles are known to exist. The strength of the magnetic force is related to the magnetic constant μ_0 , also known as the permeability of free space. This second of Maxwell's equations is known as Gauss's law for magnetism.
3. A changing magnetic field induces an electromotive force (emf) and, hence, an electric field. The direction of the emf opposes the change. This third of Maxwell's equations is Faraday's law of induction, and includes Lenz's law.
4. Magnetic fields are generated by moving charges or by changing electric fields. This fourth of Maxwell's equations encompasses Ampere's law and adds another source of magnetism—changing electric fields.

Maxwell's equations encompass the major laws of electricity and magnetism. What is not so apparent is the symmetry that Maxwell introduced in his mathematical framework. Especially important is his addition of the hypothesis that changing electric fields create magnetic fields. This is exactly analogous (and symmetric) to Faraday's law of induction and had been suspected for some time, but fits beautifully into Maxwell's equations.

Symmetry is apparent in nature in a wide range of situations. In contemporary research, symmetry plays a major part in the search for sub-atomic particles using massive multinational particle accelerators such as the new Large Hadron Collider at CERN.

MAKING CONNECTIONS: UNIFICATION OF FORCES

Maxwell's complete and symmetric theory showed that electric and magnetic forces are not separate, but different manifestations of the same thing—the electromagnetic force. This classical unification of forces is one motivation for current attempts to unify the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces.

Since changing electric fields create relatively weak magnetic fields, they could not be easily detected at the time of Maxwell's hypothesis. Maxwell realized, however, that oscillating charges, like those in AC circuits, produce changing electric fields. He predicted that these changing fields would propagate from the source like waves generated on a lake by a jumping fish.

The waves predicted by Maxwell would consist of oscillating electric and magnetic fields—defined to be an electromagnetic wave (EM wave). Electromagnetic waves would be capable of exerting forces on charges great distances from their source, and they might thus be detectable. Maxwell calculated that electromagnetic waves would propagate at a speed given by the equation

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}. \quad (11.2.1)$$

When the values for μ_0 and ϵ_0 are entered into the equation for c , we find that

$$c = \frac{1}{\sqrt{(8.85 \times 10^{-12} \frac{C^2}{N \cdot m^2})(4\pi \times 10^{-7} \frac{T \cdot m}{A})}} = 3.00 \times 10^8 \text{ m/s}, \quad (11.2.2)$$

which is the speed of light. In fact, Maxwell concluded that light is an electromagnetic wave having such wavelengths that it can be detected by the eye.

Other wavelengths should exist—it remained to be seen if they did. If so, Maxwell's theory and remarkable predictions would be verified, the greatest triumph of physics since Newton. Experimental verification came within a few years, but not before Maxwell's death.

Hertz's Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves, but also verified that they travel at the speed of light.

Hertz used an AC ***RLC*** (resistor-inductor-capacitor) circuit that resonates at a known frequency $f_0 = \frac{1}{2\pi\sqrt{LC}}$ and connected it to a loop of wire as shown in Figure 2. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and that helped generate electromagnetic waves.

Across the laboratory, Hertz had another loop attached to another ***RLC*** circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could, thus, be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.

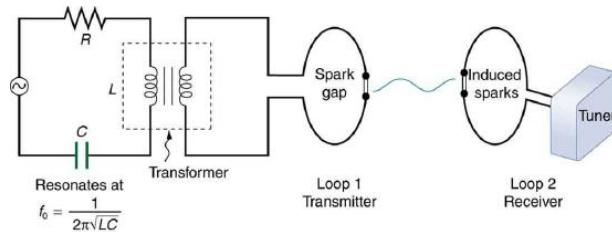


Figure 11.2.2: The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves. An ***RLC*** circuit connected to the first loop caused sparks across a gap in the wire loop and generated electromagnetic waves. Sparks across a gap in the second loop located across the laboratory gave evidence that the waves had been received.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, verifying their wave character. He was able to determine wavelength from the interference patterns, and knowing their frequency, he could calculate the propagation speed using the equation. Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, verifying their wave character. He was able to determine wavelength from the interference

patterns, and knowing their frequency, he could calculate the propagation speed using the equation $v = f\lambda$ (velocity—or speed—equals frequency times wavelength). Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz (**1Hz = 1cycle/sec**), is named in his honor.

Summary

- Electromagnetic waves consist of oscillating electric and magnetic fields and propagate at the speed of light c . They were predicted by Maxwell, who also showed that

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}, \quad (11.2.3)$$

where μ_0 is the permeability of free space and ϵ_0 is the permittivity of free space.

- Maxwell's prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell's equations.
- These four equations are paraphrased in this text, rather than presented numerically, and encompass the major laws of electricity and magnetism. First is Gauss's law for electricity, second is Gauss's law for magnetism, third is Faraday's law of induction, including Lenz's law, and fourth is Ampere's law in a symmetric formulation that adds another source of magnetism—changing electric fields.

Glossary

electromagnetic waves

radiation in the form of waves of electric and magnetic energy

Maxwell's equations

a set of four equations that comprise a complete, overarching theory of electromagnetism

RLC circuit

an electric circuit that includes a resistor, capacitor and inductor

hertz

an SI unit denoting the frequency of an electromagnetic wave, in cycles per second

speed of light

in a vacuum, such as space, the speed of light is a constant 3×10^8 m/s

electromotive force (emf)

energy produced per unit charge, drawn from a source that produces an electrical current

electric field lines

a pattern of imaginary lines that extend between an electric source and charged objects in the surrounding area, with arrows pointed away from positively charged objects and toward negatively charged objects. The more lines in the pattern, the stronger the electric field in that region

magnetic field lines

a pattern of continuous, imaginary lines that emerge from and enter into opposite magnetic poles. The density of the lines indicates the magnitude of the magnetic field

This page titled [11.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [24.1: Maxwell's Equations- Electromagnetic Waves Predicted and Observed](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/college-physics>.

11.3: Energy Carried by Electromagnetic Waves

Learning Objectives

By the end of this section, you will be able to:

- Express the time-averaged energy density of electromagnetic waves in terms of their electric and magnetic field amplitudes
- Calculate the Poynting vector and the energy intensity of electromagnetic waves
- Explain how the energy of an electromagnetic wave depends on its amplitude, whereas the energy of a photon is proportional to its frequency

Anyone who has used a microwave oven knows there is energy in electromagnetic waves. Sometimes this energy is obvious, such as in the warmth of the summer Sun. Other times, it is subtle, such as the unfelt energy of gamma rays, which can destroy living cells.

Electromagnetic waves bring energy into a system by virtue of their electric and magnetic fields. These fields can exert forces and move charges in the system and, thus, do work on them. However, there is energy in an electromagnetic wave itself, whether it is absorbed or not. Once created, the fields carry energy away from a source. If some energy is later absorbed, the field strengths are diminished and anything left travels on.

Clearly, the larger the strength of the electric and magnetic fields, the more work they can do and the greater the energy the electromagnetic wave carries. In electromagnetic waves, the amplitude is the maximum field strength of the electric and magnetic fields (Figure 11.3.1). The wave energy is determined by the wave amplitude.

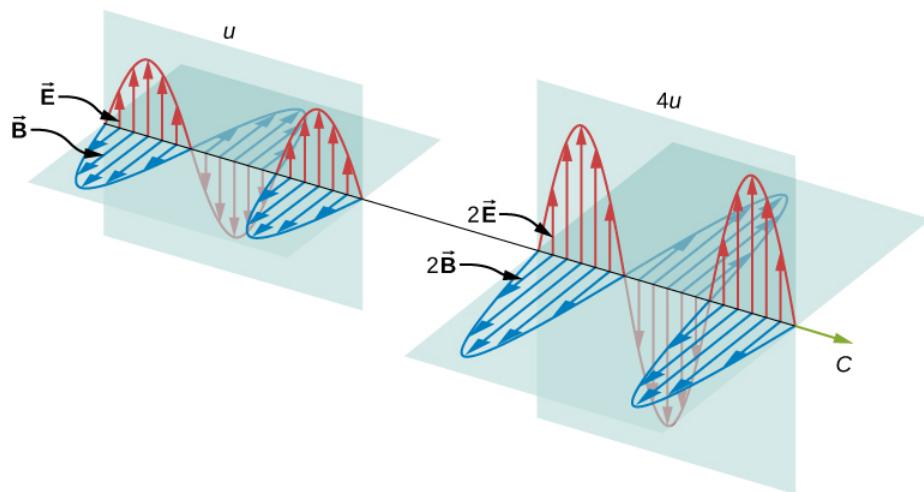


Figure 11.3.1: Energy carried by a wave depends on its amplitude. With electromagnetic waves, doubling the \mathbf{E} fields and \mathbf{B} fields quadruples the energy density \mathbf{u} and the energy flux \mathbf{uc} .

For a plane wave traveling in the direction of the positive x -axis with the phase of the wave chosen so that the wave maximum is at the origin at $t = 0$, the electric and magnetic fields obey the equations

$$E_y(x, t) = E_0 \cos(kx - \omega t)$$

$$B_x(x, t) = B_0 \cos(kx - \omega t).$$

The energy in any part of the electromagnetic wave is the sum of the energies of the electric and magnetic fields. This energy per unit volume, or energy density \mathbf{u} , is the sum of the energy density from the electric field and the energy density from the magnetic field. Expressions for both field energy densities were discussed earlier (u_E in [Capacitance](#) and u_B in [Inductance](#)). Combining these the contributions, we obtain

$$u(x, t) = u_E + u_B = \frac{1}{2} \epsilon_0 E^2 + \frac{1}{2\mu_0} B^2.$$

The expression $\mathbf{E} = c\mathbf{B} = \frac{1}{\sqrt{\epsilon_0\mu_0}}\mathbf{B}$ then shows that the magnetic energy density u_B and electric energy density u_E are equal, despite the fact that changing electric fields generally produce only small magnetic fields. The equality of the electric and magnetic energy densities leads to

$$u(x, t) = \epsilon_0 E^2 = \frac{B^2}{\mu_0}. \quad (11.3.1)$$

The energy density moves with the electric and magnetic fields in a similar manner to the waves themselves.

We can find the rate of transport of energy by considering a small time interval Δt . As shown in Figure 11.3.2, the energy contained in a cylinder of length $c\Delta t$ and cross-sectional area A passes through the cross-sectional plane in the interval Δt .

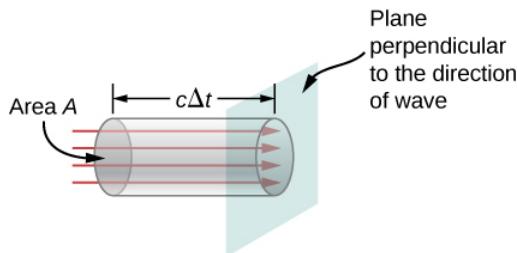


Figure 11.3.2: The energy $uAc\Delta t$ contained in the electric and magnetic fields of the electromagnetic wave in the volume $Ac\Delta t$ passes through the area A in time Δt .

The energy passing through area A in time Δt is

$$u \times \text{volume} = uAc\Delta t.$$

The energy per unit area per unit time passing through a plane perpendicular to the wave, called the **energy flux** and denoted by S , can be calculated by dividing the energy by the area A and the time interval Δt .

$$S = \frac{\text{Energy passing area } A \text{ in time } \Delta t}{A\Delta t} = uc = \epsilon_0 c E^2 = \frac{1}{\mu_0} EB.$$

More generally, the flux of energy through any surface also depends on the orientation of the surface. To take the direction into account, we introduce a vector \vec{S} , called the **Poynting vector**, with the following definition:

$$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B}.$$

The cross-product of \vec{E} and \vec{B} points in the direction perpendicular to both vectors. To confirm that the direction of \vec{S} is that of wave propagation, and not its negative, return to Figure 16.3.2. Note that Lenz's and Faraday's laws imply that when the magnetic field shown is increasing in time, the electric field is greater at x than at $x + \Delta x$. The electric field is decreasing with increasing x at the given time and location. The proportionality between electric and magnetic fields requires the electric field to increase in time along with the magnetic field. This is possible only if the wave is propagating to the right in the diagram, in which case, the relative orientations show that $\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B}$ is specifically in the direction of propagation of the electromagnetic wave.

The energy flux at any place also varies in time, as can be seen by substituting u from Equation 16.3.19 into Equation 11.3.1.

$$S(x, t) = c\epsilon_0 E_0^2 \cos^2(kx - \omega t) \quad (11.3.2)$$

Because the frequency of visible light is very high, of the order of 10^{14} Hz, the energy flux for visible light through any area is an extremely rapidly varying quantity. Most measuring devices, including our eyes, detect only an average over many cycles. The time average of the energy flux is the **intensity I** of the electromagnetic wave and is the power per unit area. It can be expressed by averaging the cosine function in Equation 11.3.2 over one complete cycle, which is the same as time-averaging over many cycles (here, T is one period):

$$I = S_{avg} = c\epsilon_0 E_0^2 \frac{1}{T} \int_0^T \cos^2 \left(2\pi \frac{t}{T} \right) dt. \quad (11.3.3)$$

We can either evaluate the integral, or else note that because the sine and cosine differ merely in phase, the average over a complete cycle for $\cos^2(\xi)$ is the same as for $\sin^2(\xi)$, to obtain

$$\langle \cos^2 \xi \rangle = \frac{1}{2} [\langle \cos^2 \xi \rangle + \langle \sin^2 \xi \rangle] = \frac{1}{2} \langle 1 \rangle = \frac{1}{2}.$$

where the angle brackets $\langle \dots \rangle$ stand for the time-averaging operation. The intensity of light moving at speed c in vacuum is then found to be

$$I = S_{avg} = \frac{1}{2} c \epsilon_0 E_0^2 \quad (11.3.4)$$

in terms of the maximum electric field strength E_0 , which is also the electric field amplitude. Algebraic manipulation produces the relationship

$$I = \frac{c B_0^2}{2 \mu_0} \quad (11.3.5)$$

where B_0 is the magnetic field amplitude, which is the same as the maximum magnetic field strength. One more expression for I_{avg} in terms of both electric and magnetic field strengths is useful. Substituting the fact that $cB_0 = E_0$, the previous expression becomes

$$I = \frac{E_0 B_0}{2 \mu_0}. \quad (11.3.6)$$

We can use whichever of the three preceding equations is most convenient, because the three equations are really just different versions of the same result: The energy in a wave is related to amplitude squared. Furthermore, because these equations are based on the assumption that the electromagnetic waves are sinusoidal, the peak intensity is twice the average intensity; that is, $I_0 = 2I$.

✓ Example 11.3.1: A Laser Beam

The beam from a small laboratory laser typically has an intensity of about $1.0 \times 10^{-3} W/m^2$. Assuming that the beam is composed of plane waves, calculate the amplitudes of the electric and magnetic fields in the beam.

Strategy

Use the equation expressing intensity in terms of electric field to calculate the electric field from the intensity.

Solution

From Equation 11.3.4, the intensity of the laser beam is

$$I = \frac{1}{2} c \epsilon_0 E_0^2.$$

The amplitude of the electric field is therefore

$$\begin{aligned} E_0 &= \sqrt{\frac{2}{c \epsilon_0} I} \\ &= \sqrt{\frac{2}{(3.00 \times 10^8 m/s)(8.85 \times 10^{-12} F/m)}} (1.0 \times 10^{-3} W/m^2) \\ &= 0.87 V/m. \end{aligned}$$

The amplitude of the magnetic field can be obtained from:

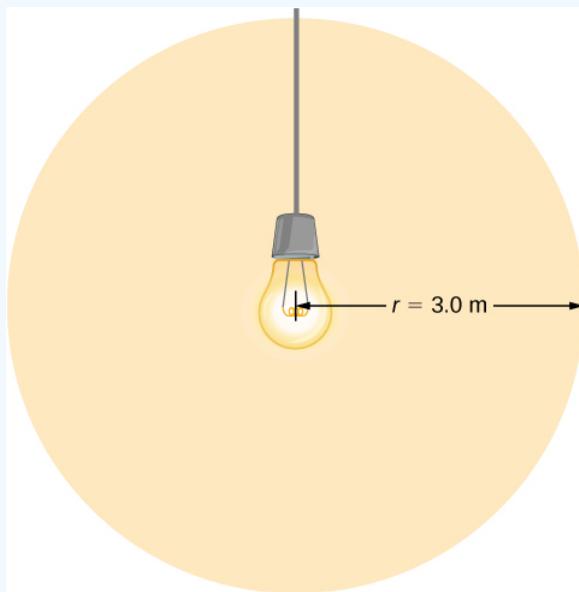
$$B_0 = \frac{E_0}{c} = 2.9 \times 10^{-9} T.$$

✓ Light Bulb Fields

A light bulb emits 5.00 W of power as visible light. What are the average electric and magnetic fields from the light at a distance of 3.0 m?

Strategy

Assume the bulb's power output P is distributed uniformly over a sphere of radius 3.0 m to calculate the intensity, and from it, the electric field.



Solution

The power radiated as visible light is then

$$I = \frac{P}{4\pi r^2} = \frac{\epsilon_0 E_0^2}{2},$$

$$E_0 = \sqrt{2 \frac{P}{4\pi r^2 \epsilon_0}} = \sqrt{2 \frac{5.00 \text{ W}}{4\pi (3.0 \text{ m})^2 (3.00 \times 10^8 \text{ m/s}) (8.85 \times 10^{-12} \text{ C}^2/\text{N}\cdot\text{m}^2)}} = 5.77 \text{ N/C},$$

$$B_0 = E_0/c = 1.92 \times 10^{-8} \text{ T}.$$

Significance

The intensity I falls off as the distance squared if the radiation is dispersed uniformly in all directions.

✓ Radio Range

A 60-kW radio transmitter on Earth sends its signal to a satellite 100 km away (Figure 11.3.3). At what distance in the same direction would the signal have the same maximum field strength if the transmitter's output power were increased to 90 kW?

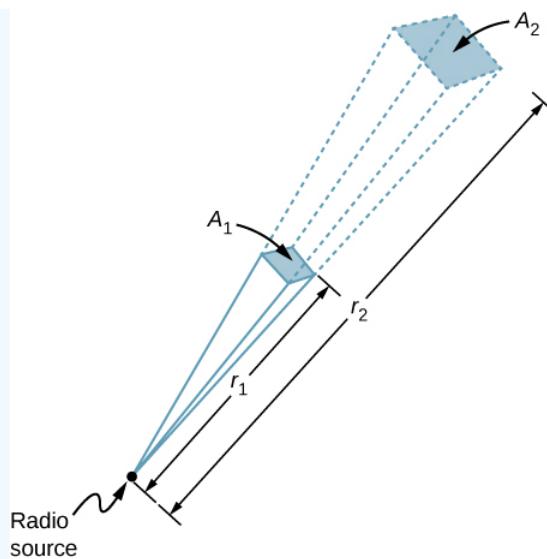


Figure 11.3.3: In three dimensions, a signal spreads over a solid angle as it travels outward from its source.

Strategy

The area over which the power in a particular direction is dispersed increases as distance squared, as illustrated in Figure 11.3.3. Change the power output \mathbf{P} by a factor of (90 kW/60 kW) and change the area by the same factor to keep $I = \frac{\mathbf{P}}{A} = \frac{\epsilon_0 E_0^2}{2}$ the same. Then use the proportionality of area \mathbf{A} in the diagram to distance squared to find the distance that produces the calculated change in area.

Solution

Using the proportionality of the areas to the squares of the distances, and solving, we obtain from the diagram

$$\begin{aligned}\frac{r_2^2}{r_1^2} &= \frac{A_2}{A_1} = \frac{90\text{ W}}{60\text{ W}}, \\ r_2 &= \sqrt{\frac{90}{60}}(100\text{ km}) \\ &= 122\text{ km}.\end{aligned}$$

Significance

The range of a radio signal is the maximum distance between the transmitter and receiver that allows for normal operation. In the absence of complications such as reflections from obstacles, the intensity follows an inverse square law, and doubling the range would require multiplying the power by four.

This page titled [11.3: Energy Carried by Electromagnetic Waves](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [16.4: Energy Carried by Electromagnetic Waves](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

11.4: The Electromagnetic Spectrum

Learning Objectives

By the end of this section, you will be able to:

- List three “rules of thumb” that apply to the different frequencies along the electromagnetic spectrum.
- Explain why the higher the frequency, the shorter the wavelength of an electromagnetic wave.
- Draw a simplified electromagnetic spectrum, indicating the relative positions, frequencies, and spacing of the different types of radiation bands.
- List and explain the different methods by which electromagnetic waves are produced across the spectrum.

In this module we examine how electromagnetic waves are classified into categories such as radio, infrared, ultraviolet, and so on, so that we can understand some of their similarities as well as some of their differences. We will also find that there are many connections with previously discussed topics, such as wavelength and resonance. A brief overview of the production and utilization of electromagnetic waves is found in Table 11.4.1.

Table 11.4.1: Electromagnetic Wave

Type of EM wave	Production	Applications	Life sciences aspect	Issues
Radio & TV	Acceleration charges	Communications Remote controls	MRI	Requires controls for band use
Microwaves	Accelerating charges & thermal agitation	Communication Ovens Radar	Deep heating	Cell phone use
Infrared	Thermal agitations & electronic transitions	Thermal imaging Heating	Absorbed by atmosphere	Greenhouse effect
Visible light	Thermal agitations & electronic transitions	All pervasive	Photosynthesis Human vision	
Ultraviolet	Thermal agitations & electronic transitions	Sterilization Cancer control	Vitamin D production	Ozone depletion Cancer causing
X-rays	Inner electronic transitions and fast collisions	Medical Security	Medical diagnosis Cancer therapy	Cancer causing
Gamma rays	Nuclear decay	Nuclear medicine Security	Medical diagnosis Cancer therapy	Cancer causing Radiation damage

Wave

There are many types of waves, such as water waves and even earthquakes. Among the many shared attributes of waves are propagation speed, frequency, and wavelength. These are always related by the expression $vw = f\lambda$. This module concentrates on EM waves, but other modules contain examples of all of these characteristics for sound waves and submicroscopic particles.

As noted before, an electromagnetic wave has a frequency and a wavelength associated with it and travels at the speed of light, or c . The relationship among these wave characteristics can be described by $vw = f\lambda$, where vw is the propagation speed of the wave, f is the frequency, and λ is the wavelength. Here $vw = c$, so that for all electromagnetic waves,

$$c = f\lambda. \quad (11.4.1)$$

Thus, for all electromagnetic waves, the greater the frequency, the smaller the wavelength.

Figure 11.4.1 shows how the various types of electromagnetic waves are categorized according to their wavelengths and frequencies -- that is, it shows the electromagnetic spectrum. Many of the characteristics of the various types of electromagnetic waves are related to their frequencies and wavelengths, as we shall see.

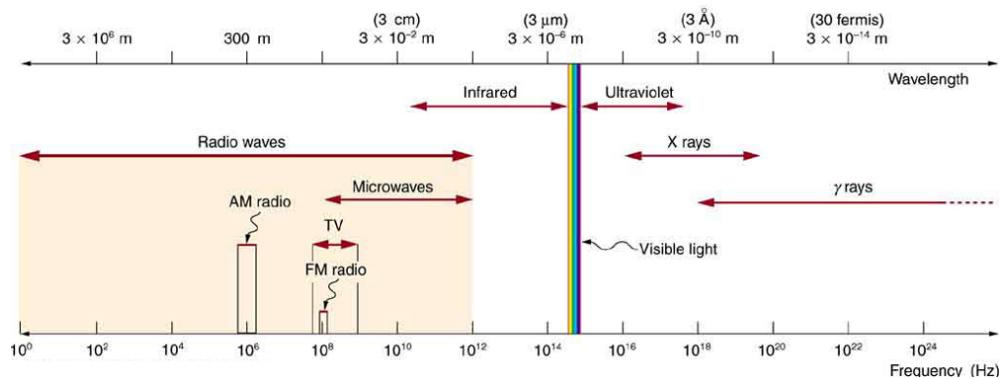


Figure 11.4.1: The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

ELECTROMAGNETIC SPECTRUM: RULES OF THUMB

Three rules that apply to electromagnetic waves in general are as follows:

- High-frequency electromagnetic waves are more energetic and are more able to penetrate than low-frequency waves.
- High-frequency electromagnetic waves can carry more information per unit time than low-frequency waves.
- The shorter the wavelength of any electromagnetic wave probing a material, the smaller the detail it is possible to resolve.

Note that there are exceptions to these rules of thumb.

Transmission, Reflection, and Absorption

What happens when an electromagnetic wave impinges on a material? If the material is transparent to the particular frequency, then the wave can largely be transmitted. If the material is opaque to the frequency, then the wave can be totally reflected. The wave can also be absorbed by the material, indicating that there is some interaction between the wave and the material, such as the thermal agitation of molecules.

Of course it is possible to have partial transmission, reflection, and absorption. We normally associate these properties with visible light, but they do apply to all electromagnetic waves. What is not obvious is that something that is transparent to light may be opaque at other frequencies. For example, ordinary glass is transparent to visible light but largely opaque to ultraviolet radiation. Human skin is opaque to visible light -- we cannot see through people -- but transparent to X-rays.

Radio and TV Waves

The broad category of **radio waves** is defined to contain any electromagnetic wave produced by currents in wires and circuits. Its name derives from their most common use as a carrier of audio information (i.e., radio). The name is applied to electromagnetic waves of similar frequencies regardless of source. Radio waves from outer space, for example, do not come from alien radio stations. They are created by many astronomical phenomena, and their study has revealed much about nature on the largest scales.

There are many uses for radio waves, and so the category is divided into many subcategories, including microwaves and those electromagnetic waves used for AM and FM radio, cellular telephones, and TV.

The lowest commonly encountered radio frequencies are produced by high-voltage AC power transmission lines at frequencies of 50 or 60 Hz. (Figure 11.4.2). These extremely long wavelength electromagnetic waves (about 6000 km!) are one means of energy loss in long-distance power transmission.



Figure 11.4.2: This high-voltage traction power line running to Eutingen Railway Substation in Germany radiates electromagnetic waves with very long wavelengths. (credit: Zonk43, Wikimedia Commons)

There is an ongoing controversy regarding potential health hazards associated with exposure to these electromagnetic fields (*E*-fields). Some people suspect that living near such transmission lines may cause a variety of illnesses, including cancer. But demographic data are either inconclusive or simply do not support the hazard theory. Recent reports that have looked at many European and American epidemiological studies have found no increase in risk for cancer due to exposure to *E*-fields.

Extremely low frequency (ELF) radio waves of about 1 kHz are used to communicate with submerged submarines. The ability of radio waves to penetrate salt water is related to their wavelength (much like ultrasound penetrating tissue) -- the longer the wavelength, the farther they penetrate. Since salt water is a good conductor, radio waves are strongly absorbed by it, and very long wavelengths are needed to reach a submarine under the surface. (Figure 11.4.3).

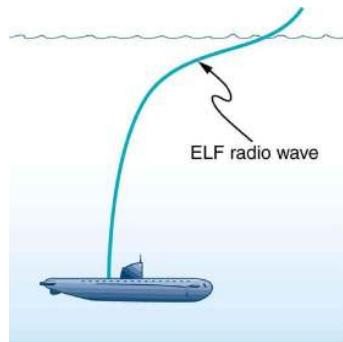


Figure 11.4.3: Very long wavelength radio waves are needed to reach this submarine, requiring extremely low frequency signals (ELF). Shorter wavelengths do not penetrate to any significant depth.

AM radio waves are used to carry commercial radio signals in the frequency range from 540 to 1600 kHz. The abbreviation AM stands for **amplitude modulation**, which is the method for placing information on these waves (Figure 11.4.4). A **carrier wave** having the basic frequency of the radio station, say 1530 kHz, is varied or modulated in amplitude by an audio signal. The resulting wave has a constant frequency, but a varying amplitude.

A radio receiver tuned to have the same resonant frequency as the carrier wave can pick up the signal, while rejecting the many other frequencies impinging on its antenna. The receiver's circuitry is designed to respond to variations in amplitude of the carrier wave to replicate the original audio signal. That audio signal is amplified to drive a speaker or perhaps to be recorded.

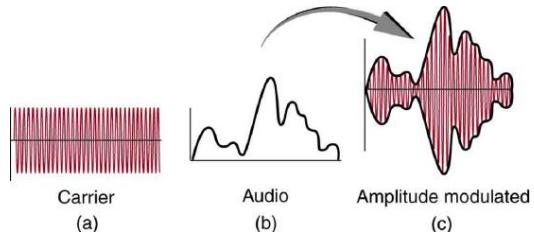


Figure 11.4.4: Amplitude modulation for AM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The amplitude of the carrier is modulated by the audio signal without changing its basic frequency.

FM Radio Waves

FM radio waves are also used for commercial radio transmission, but in the frequency range of 88 to 108 MHz. FM stands for **frequency modulation**, another method of carrying information (Figure 11.4.5). Here a carrier wave having the basic frequency of the radio station, perhaps 105.1 MHz, is modulated in frequency by the audio signal, producing a wave of constant amplitude but varying frequency.

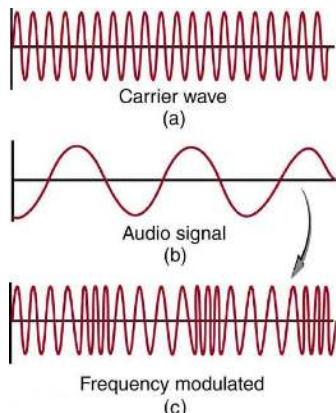


Figure 11.4.5: Frequency modulation for FM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The frequency of the carrier is modulated by the audio signal without changing its amplitude.

Since audible frequencies range up to 20 kHz (or 0.020 MHz) at most, the frequency of the FM radio wave can vary from the carrier by as much as 0.020 MHz. Thus the carrier frequencies of two different radio stations cannot be closer than 0.020 MHz. An FM receiver is tuned to resonate at the carrier frequency and has circuitry that responds to variations in frequency, reproducing the audio information.

FM radio is inherently less subject to noise from stray radio sources than AM radio. The reason is that amplitudes of waves add. So an AM receiver would interpret noise added onto the amplitude of its carrier wave as part of the information. An FM receiver can be made to reject amplitudes other than that of the basic carrier wave and only look for variations in frequency. It is thus easier to reject noise from FM, since noise produces a variation in amplitude.

Television is also broadcast on electromagnetic waves. Since the waves must carry a great deal of visual as well as audio information, each channel requires a larger range of frequencies than simple radio transmission. TV channels utilize frequencies in the range of 54 to 88 MHz and 174 to 222 MHz. (The entire FM radio band lies between channels 88 MHz and 174 MHz.) These TV channels are called VHF (for **very high frequency**). Other channels called UHF (for **ultra high frequency**) utilize an even higher frequency range of 470 to 1000 MHz.

The TV video signal is AM, while the TV audio is FM. Note that these frequencies are those of free transmission with the user utilizing an old-fashioned roof antenna. Satellite dishes and cable transmission of TV occurs at significantly higher frequencies and is rapidly evolving with the use of the high-definition or HD format.

Example 11.4.1: Calculating Wavelengths of Radio Waves:

Calculate the wavelengths of a 1530-kHz AM radio signal, a 105.1-MHz FM radio signal, and a 1.90-GHz cell phone signal.

Strategy

The relationship between wavelength and frequency is $c = f\lambda$, where $c = 3.00 \times 10^8 \text{ m/s}$ is the speed of light (the speed of light is only very slightly smaller in air than it is in a vacuum). We can rearrange this equation to find the wavelength for all three frequencies.

Solution

Rearranging gives

$$\lambda = \frac{c}{f}. \quad (11.4.2)$$

a. For the $f = 1530\text{kHz}$ AM radio signal, then,

$$\lambda = \frac{3.00 \times 10^8 \text{m/s}}{1530 \times 10^3 \text{cycles/s}} \quad (11.4.3)$$

$$= 196\text{m.} \quad (11.4.4)$$

b. For the $f = 105.1\text{MHz}$ AM radio signal, then,

$$\lambda = \frac{3.00 \times 10^8 \text{m/s}}{105.1 \times 10^6 \text{cycles/s}} \quad (11.4.5)$$

$$= 2.85\text{m.} \quad (11.4.6)$$

c. For the $f = 1.90\text{GHz}$ AM radio signal, then,

$$\lambda = \frac{3.00 \times 10^8 \text{m/s}}{1.90 \times 10^9 \text{cycles/s}} \quad (11.4.7)$$

$$= 0.158\text{m.} \quad (11.4.8)$$

Discussion

These wavelengths are consistent with the spectrum in Figure 11.4.1. The wavelengths are also related to other properties of these electromagnetic waves, as we shall see.

The wavelengths found in the preceding example are representative of AM, FM, and cell phones, and account for some of the differences in how they are broadcast and how well they travel. The most efficient length for a linear antenna, such as discussed in 24.3, is $\lambda/2$, half the wavelength of the electromagnetic wave. Thus a very large antenna is needed to efficiently broadcast typical AM radio with its carrier wavelengths on the order of hundreds of meters.

One benefit to these long AM wavelengths is that they can go over and around rather large obstacles (like buildings and hills), just as ocean waves can go around large rocks. FM and TV are best received when there is a line of sight between the broadcast antenna and receiver, and they are often sent from very tall structures. FM, TV, and mobile phone antennas themselves are much smaller than those used for AM, but they are elevated to achieve an unobstructed line of sight (Figure 11.4.6).

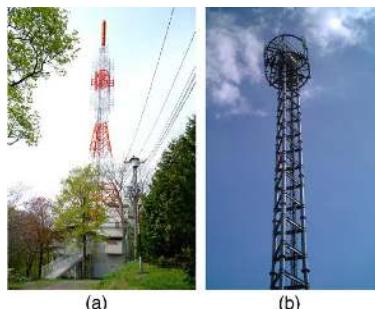


Figure 11.4.6: (a) A large tower is used to broadcast TV signals. The actual antennas are small structures on top of the tower -- they are placed at great heights to have a clear line of sight over a large broadcast area. (credit: Ozizo, Wikimedia Commons) (b) The NTT Dokomo mobile phone tower at Tokorozawa City, Japan. (credit: tokoroten, Wikimedia Commons)

Radio Wave Interference

Astronomers and astrophysicists collect signals from outer space using electromagnetic waves. A common problem for astrophysicists is the “pollution” from electromagnetic radiation pervading our surroundings from communication systems in general. Even everyday gadgets like our car keys having the facility to lock car doors remotely and being able to turn TVs on and off using remotes involve radio-wave frequencies. In order to prevent interference between all these electromagnetic signals, strict regulations are drawn up for different organizations to utilize different radio frequency bands.

One reason why we are sometimes asked to switch off our mobile phones (operating in the range of 1.9 GHz) on airplanes and in hospitals is that important communications or medical equipment often uses similar radio frequencies and their operation can be

affected by frequencies used in the communication devices.

For example, radio waves used in magnetic resonance imaging (MRI) have frequencies on the order of 100 MHz, although this varies significantly depending on the strength of the magnetic field used and the nuclear type being scanned. MRI is an important medical imaging and research tool, producing highly detailed two- and three-dimensional images. Radio waves are broadcast, absorbed, and reemitted in a resonance process that is sensitive to the density of nuclei (usually protons or hydrogen nuclei).

The wavelength of 100-MHz radio waves is 3 m, yet using the sensitivity of the resonant frequency to the magnetic field strength, details smaller than a millimeter can be imaged. This is a good example of an exception to a rule of thumb (in this case, the rubric that details much smaller than the probe's wavelength cannot be detected). The intensity of the radio waves used in MRI presents little or no hazard to human health.

Microwaves

Microwaves are the highest-frequency electromagnetic waves that can be produced by currents in macroscopic circuits and devices. Microwave frequencies range from about 10^9 Hz to the highest practical *LC* resonance at nearly 10^{12} Hz . Since they have high frequencies, their wavelengths are short compared with those of other radio waves -- hence the name "microwave."

Microwaves can also be produced by atoms and molecules. They are, for example, a component of electromagnetic radiation generated by **thermal agitation**. The thermal motion of atoms and molecules in any object at a temperature above absolute zero causes them to emit and absorb radiation.

Since it is possible to carry more information per unit time on high frequencies, microwaves are quite suitable for communications. Most satellite-transmitted information is carried on microwaves, as are land-based long-distance transmissions. A clear line of sight between transmitter and receiver is needed because of the short wavelengths involved.

Radar is a common application of microwaves that was first developed in World War II. By detecting and timing microwave echoes, radar systems can determine the distance to objects as diverse as clouds and aircraft. A Doppler shift in the radar echo can be used to determine the speed of a car or the intensity of a rainstorm. Sophisticated radar systems are used to map the Earth and other planets, with a resolution limited by wavelength. (Figure 11.4.7). The shorter the wavelength of any probe, the smaller the detail it is possible to observe.



Figure 11.4.7: An image of [Sif Mons](#) with lava flows on Venus, based on Magellan synthetic aperture radar data combined with radar altimetry to produce a three-dimensional map of the surface. The Venusian atmosphere is opaque to visible light, but not to the microwaves that were used to create this image. (credit: NSSDC, NASA/JPL)

Heating with Microwaves

How does the ubiquitous microwave oven produce microwaves electronically, and why does food absorb them preferentially? Microwaves at a frequency of 2.45 GHz are produced by accelerating electrons. The microwaves are then used to induce an alternating electric field in the oven.

Water and some other constituents of food have a slightly negative charge at one end and a slightly positive charge at one end (called polar molecules). The range of microwave frequencies is specially selected so that the polar molecules, in trying to keep orienting themselves with the electric field, absorb these energies and increase their temperatures -- called dielectric heating.

The energy thereby absorbed results in thermal agitation heating food and not the plate, which does not contain water. Hot spots in the food are related to constructive and destructive interference patterns. Rotating antennas and food turntables help spread out the hot spots.

Another use of microwaves for heating is within the human body. Microwaves will penetrate more than shorter wavelengths into tissue and so can accomplish “deep heating” (called microwave diathermy). This is used for treating muscular pains, spasms, tendonitis, and rheumatoid arthritis.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT - MICROWAVE OVENS

1. Look at the door of a microwave oven. Describe the structure of the door. Why is there a metal grid on the door? How does the size of the holes in the grid compare with the wavelengths of microwaves used in microwave ovens? What is this wavelength?
2. Place a glass of water (about 250 ml) in the microwave and heat it for 30 seconds. Measure the temperature gain (the ΔT). Assuming that the power output of the oven is 1000 W, calculate the efficiency of the heat-transfer process.
3. Remove the rotating turntable or moving plate and place a cup of water in several places along a line parallel with the opening. Heat for 30 seconds and measure the ΔT for each position. Do you see cases of destructive interference?

Microwaves generated by atoms and molecules far away in time and space can be received and detected by electronic circuits. Deep space acts like a blackbody with a 2.7 K temperature, radiating most of its energy in the microwave frequency range. In 1964, Penzias and Wilson detected this radiation and eventually recognized that it was the radiation of the Big Bang’s cooled remnants.

Infrared Radiation

The microwave and infrared regions of the electromagnetic spectrum overlap (Figure 11.4.1). **Infrared radiation** is generally produced by thermal motion and the vibration and rotation of atoms and molecules. Electronic transitions in atoms and molecules can also produce infrared radiation.

The range of infrared frequencies extends up to the lower limit of visible light, just below red. In fact, infrared means “below red.” Frequencies at its upper limit are too high to be produced by accelerating electrons in circuits, but small systems, such as atoms and molecules, can vibrate fast enough to produce these waves.

Water molecules rotate and vibrate particularly well at infrared frequencies, emitting and absorbing them so efficiently that the emissivity for skin is $e = 0.97$ in the infrared. Night-vision scopes can detect the infrared emitted by various warm objects, including humans, and convert it to visible light.

We can examine radiant heat transfer from a house by using a camera capable of detecting infrared radiation. Reconnaissance satellites can detect buildings, vehicles, and even individual humans by their infrared emissions, whose power radiation is proportional to the fourth power of the absolute temperature. More mundanely, we use infrared lamps, some of which are called quartz heaters, to preferentially warm us because we absorb infrared better than our surroundings.

The Sun radiates like a nearly perfect blackbody (that is, it has $e = 1$), with a 6000 K surface temperature. About half of the solar energy arriving at the Earth is in the infrared region, with most of the rest in the visible part of the spectrum, and a relatively small amount in the ultraviolet. On average, 50 percent of the incident solar energy is absorbed by the Earth.

The relatively constant temperature of the Earth is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by CO_2 and H_2O in the atmosphere and then radiated back to Earth or into outer space. This radiation back to Earth is known as the greenhouse effect, and it maintains the surface temperature of the Earth about $40^\circ C$ higher than it would be if there is no absorption. Some scientists think that the increased concentration of CO_2 and other greenhouse gases in the atmosphere, resulting from increases in fossil fuel burning, has increased global average temperatures.

Visible Light

Visible light is the narrow segment of the electromagnetic spectrum to which the normal human eye responds. Visible light is produced by vibrations and rotations of atoms and molecules, as well as by electronic transitions within atoms and molecules. The receivers or detectors of light largely utilize electronic transitions. We say the atoms and molecules are excited when they absorb and relax when they emit through electronic transitions.

Figure 11.4.8 shows this part of the spectrum, together with the colors associated with particular pure wavelengths. We usually refer to visible light as having wavelengths of between 400 nm and 750 nm. (The retina of the eye actually responds to the lowest

ultraviolet frequencies, but these do not normally reach the retina because they are absorbed by the cornea and lens of the eye.)

Red light has the lowest frequencies and longest wavelengths, while violet has the highest frequencies and shortest wavelengths. Blackbody radiation from the Sun peaks in the visible part of the spectrum but is more intense in the red than in the violet, making the Sun yellowish in appearance.

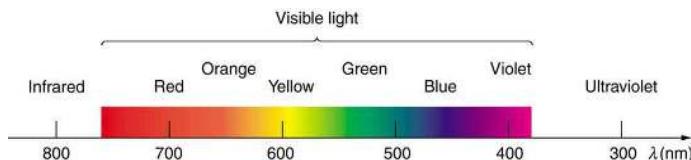


Figure 11.4.8: A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly distinct, nor are those between the seven rainbow colors.

Living things--plants and animals-- have evolved to utilize and respond to parts of the electromagnetic spectrum they are embedded in. Visible light is the most predominant and we enjoy the beauty of nature through visible light. Plants are more selective. Photosynthesis makes use of parts of the visible spectrum to make sugars.

Example 11.4.2: Integrated Concept Problem: Correcting Vision with Lasers

During laser vision correction, a brief burst of 193-nm ultraviolet light is projected onto the cornea of a patient. It makes a spot 0.80 mm in diameter and evaporates a layer of cornea $0.30\mu\text{m}$ thick. Calculate the energy absorbed, assuming the corneal tissue has the same properties as water; it is initially at 34°C . Assume the evaporated tissue leaves at a temperature of 100°C .

Strategy

The energy from the laser light goes toward raising the temperature of the tissue and also toward evaporating it. Thus we have two amounts of heat to add together. Also, we need to find the mass of corneal tissue involved.

Solution

To figure out the heat required to raise the temperature of the tissue to 100°C , we can apply concepts of thermal energy. We know that

$$Q = mc\Delta T, \quad (11.4.9)$$

where Q is the heat required to raise the temperature, ΔT is the desired change in temperature, m is the mass of tissue to be heated, and c is the specific heat of water equal to 4186 J/kg/K .

Without knowing the mass m at this point, we have

$$\begin{aligned} Q &= m(4186 \text{ J/kg/K})(100^\circ\text{C} - 34^\circ\text{C}) \\ &= m(276,276 \text{ J/kg}) \\ &= m(276 \text{ J/kg}). \end{aligned}$$

The latent heat of vaporization of water is 2256 kJ/kg , so that the energy needed to evaporate mass m is

$$Q_v = mL_v = m(2256 \text{ kJ/kg}).$$

To find the mass m , we use the equation $\rho = m/V$, where ρ is the density of the tissue and V is its volume. For this case,

$$\begin{aligned} m &= \rho V \\ &= (1000 \text{ kg/m}^3)(\text{area} \times \text{thickness} (\text{m}^3)) \\ &= (1000 \text{ kg/m}^3)(\pi(0.80 \times 10^{-3} \text{ m})^2 / 4)(0.30 \times 10^{-6} \text{ m}) \\ &= 0.151 \times 10^{-9} \text{ kg}. \end{aligned}$$

Therefore, the total energy absorbed by the tissue in the eye is the sum of Q and Q_v :

$$\begin{aligned}Q_{\text{tot}} &= m(c\Delta T + L_v) \\&= (0.151 \times 10^{-9} \text{ kg})(276 \text{ kJ/kg} + 2256 \text{ kJ/kg}) \\&= 382 \times 10^{-9} \text{ kJ.}\end{aligned}$$

Discussion

The lasers used for this eye surgery are excimer lasers, whose light is well absorbed by biological tissue. They evaporate rather than burn the tissue, and can be used for precision work. Most lasers used for this type of eye surgery have an average power rating of about one watt. For our example, if we assume that each laser burst from this pulsed laser lasts for 10 ns, and there are 400 bursts per second, then the average power is

$$Q_{\text{tot}} \times 400 = 150 \text{ mW}$$

Optics is the study of the behavior of visible light and other forms of electromagnetic waves. Optics falls into two distinct categories. When electromagnetic radiation, such as visible light, interacts with objects that are large compared with its wavelength, its motion can be represented by straight lines like rays. Ray optics is the study of such situations and includes lenses and mirrors.

When electromagnetic radiation interacts with objects about the same size as the wavelength or smaller, its wave nature becomes apparent. For example, observable detail is limited by the wavelength, and so visible light can never detect individual atoms, because they are so much smaller than its wavelength. Physical or wave optics is the study of such situations and includes all wave characteristics.

TAKE-HOME EXPERIMENT: COLORS THAT MATCH

When you light a match you see largely orange light; when you light a gas stove you see blue light. Why are the colors different? What other colors are present in these?

Ultraviolet Radiation

Ultraviolet means “above violet.” The electromagnetic frequencies of **ultraviolet radiation (UV)** extend upward from violet, the highest-frequency visible light. Ultraviolet is also produced by atomic and molecular motions and electronic transitions. The wavelengths of ultraviolet extend from 400 nm down to about 10 nm at its highest frequencies, which overlap with the lowest X-ray frequencies. It was recognized as early as 1801 by Johann Ritter that the solar spectrum had an invisible component beyond the violet range.

Solar UV radiation is broadly subdivided into three regions: UV-A (320–400 nm), UV-B (290–320 nm), and UV-C (220–290 nm), ranked from long to shorter wavelengths (from smaller to larger energies). Most UV-B and all UV-C is absorbed by ozone (O_3) molecules in the upper atmosphere. Consequently, 99% of the solar UV radiation reaching the Earth’s surface is UV-A.

Human Exposure to UV Radiation

It is largely exposure to UV-B that causes skin cancer. It is estimated that as many as 20% of adults will develop skin cancer over the course of their lifetime. Again, treatment is often successful if caught early. Despite very little UV-B reaching the Earth’s surface, there are substantial increases in skin-cancer rates in countries such as Australia, indicating how important it is that UV-B and UV-C continue to be absorbed by the upper atmosphere.

All UV radiation can damage collagen fibers, resulting in an acceleration of the aging process of skin and the formation of wrinkles. Because there is so little UV-B and UV-C reaching the Earth’s surface, sunburn is caused by large exposures, and skin cancer from repeated exposure. Some studies indicate a link between overexposure to the Sun when young and melanoma later in life.

The tanning response is a defense mechanism in which the body produces pigments to absorb future exposures in inert skin layers above living cells. Basically UV-B radiation excites DNA molecules, distorting the DNA helix, leading to mutations and the possible formation of cancerous cells.

Repeated exposure to UV-B may also lead to the formation of cataracts in the eyes -- a cause of blindness among people living in the equatorial belt where medical treatment is limited. Cataracts, clouding in the eye’s lens and a loss of vision, are age related;

60% of those between the ages of 65 and 74 will develop cataracts. However, treatment is easy and successful, as one replaces the lens of the eye with a plastic lens. Prevention is important. Eye protection from UV is more effective with plastic sunglasses than those made of glass.

A major acute effect of extreme UV exposure is the suppression of the immune system, both locally and throughout the body.

Low-intensity ultraviolet is used to sterilize haircutting implements, implying that the energy associated with ultraviolet is deposited in a manner different from lower-frequency electromagnetic waves. (Actually this is true for all electromagnetic waves with frequencies greater than visible light.)

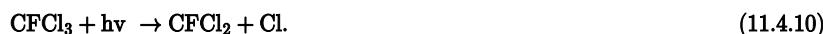
Flash photography is generally not allowed of precious artworks and colored prints because the UV radiation from the flash can cause photo-degradation in the artworks. Often artworks will have an extra-thick layer of glass in front of them, which is especially designed to absorb UV radiation.

UV Light and the Ozone Layer

If all of the Sun's ultraviolet radiation reached the Earth's surface, there would be extremely grave effects on the biosphere from the severe cell damage it causes. However, the layer of ozone (O_3) in our upper atmosphere (10 to 50 km above the Earth) protects life by absorbing most of the dangerous UV radiation.

Unfortunately, today we are observing a depletion in ozone concentrations in the upper atmosphere. This depletion has led to the formation of an "ozone hole" in the upper atmosphere. The hole is more centered over the southern hemisphere, and changes with the seasons, being largest in the spring. This depletion is attributed to the breakdown of ozone molecules by refrigerant gases called chlorofluorocarbons (CFCs).

The UV radiation helps dissociate the CFC's, releasing highly reactive chlorine (Cl) atoms, which catalyze the destruction of the ozone layer. For example, the reaction of $CFCl_3$ with a photon of light ($\hbar\nu$) can be written as:



The Cl atom then catalyzes the breakdown of ozone as follows:



and



A single chlorine atom could destroy ozone molecules for up to two years before being transported down to the surface. The CFCs are relatively stable and will contribute to ozone depletion for years to come. CFCs are found in refrigerants, air conditioning systems, foams, and aerosols.

International concern over this problem led to the establishment of the "Montreal Protocol" agreement (1987) to phase out CFC production in most countries. However, developing-country participation is needed if worldwide production and elimination of CFCs is to be achieved. Probably the largest contributor to CFC emissions today is India. But the protocol seems to be working, as there are signs of an ozone recovery (Figure 11.4.9).

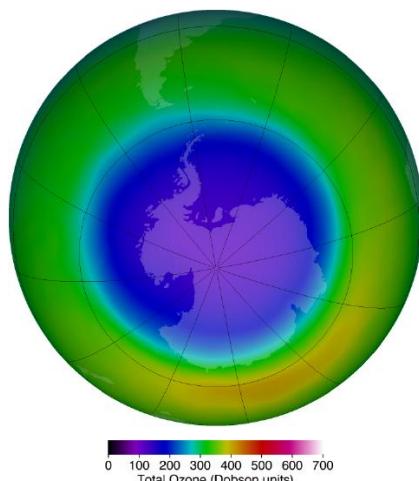


Figure 11.4.9: This map of ozone concentration over Antarctica in October 2011 shows severe depletion suspected to be caused by CFCs. Less dramatic but more general depletion has been observed over northern latitudes, suggesting the effect is global. With less ozone, more ultraviolet radiation from the Sun reaches the surface, causing more damage. (credit: NASA Ozone Watch)

Benefits of UV Light

Besides the adverse effects of ultraviolet radiation, there are also benefits of exposure in nature and uses in technology. Vitamin D production in the skin (epidermis) results from exposure to UVB radiation, generally from sunlight. A number of studies indicate lack of vitamin D can result in the development of a range of cancers (prostate, breast, colon), so a certain amount of UV exposure is helpful. Lack of vitamin D is also linked to osteoporosis. Exposures (with no sunscreen) of 10 minutes a day to arms, face, and legs might be sufficient to provide the accepted dietary level. However, in the winter time north of about 37° latitude, most UVB gets blocked by the atmosphere.

UV radiation is used in the treatment of infantile jaundice and in some skin conditions. It is also used in sterilizing workspaces and tools, and killing germs in a wide range of applications. It is also used as an analytical tool to identify substances.

When exposed to ultraviolet, some substances, such as minerals, glow in characteristic visible wavelengths, a process called fluorescence. So-called black lights emit ultraviolet to cause posters and clothing to fluoresce in the visible. Ultraviolet is also used in special microscopes to detect details smaller than those observable with longer-wavelength visible-light microscopes.

THINGS GREAT AND SMALL: A SUBMICROSCOPIC VIEW OF X-RAY PRODUCTION

X-rays can be created in a high-voltage discharge. They are emitted in the material struck by electrons in the discharge current. There are two mechanisms by which the electrons create X-rays.

The first method is illustrated in the Figure 11.4.10. An electron is accelerated in an evacuated tube by a high positive voltage. The electron strikes a metal plate (e.g., copper) and produces X-rays. Since this is a high-voltage discharge, the electron gains sufficient energy to ionize the atom.

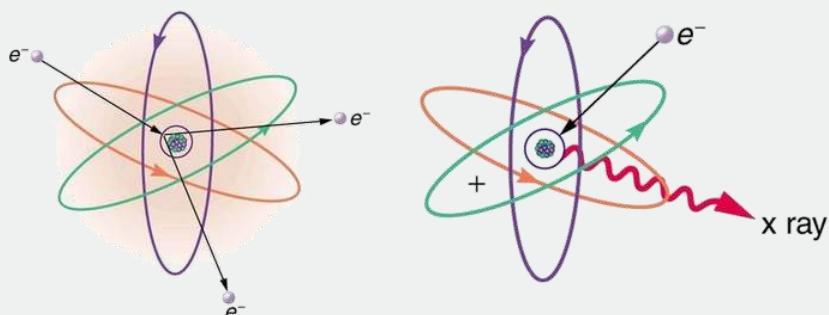


Figure 11.4.10: Artist's conception of an electron ionizing an atom followed by the recapture of an electron and emission of an X-ray. An energetic electron strikes an atom and knocks an electron out of one of the orbits closest to the nucleus. Later, the atom captures another electron, and the energy released by its fall into a low orbit generates a high-energy EM wave called an X-ray.

In the case shown, an inner-shell electron (one in an orbit relatively close to and tightly bound to the nucleus) is ejected. A short time later, another electron is captured and falls into the orbit in a single great plunge. The energy released by this fall is given to an EM wave known as an X-ray. Since the orbits of the atom are unique to the type of atom, the energy of the X-ray is characteristic of the atom, hence the name characteristic X-ray.

The second method by which an energetic electron creates an X-ray when it strikes a material is illustrated in the figure below. The electron interacts with charges in the material as it penetrates. These collisions transfer kinetic energy from the electron to the electrons and atoms in the material.

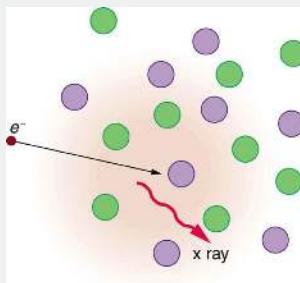


Figure 11.4.11: Artist's conception of an electron being slowed by collisions in a material and emitting X-ray radiation. This energetic electron makes numerous collisions with electrons and atoms in a material it penetrates. An accelerated charge radiates EM waves, a second method by which X-rays are created.

A loss of kinetic energy implies an acceleration, in this case decreasing the electron's velocity. Whenever a charge is accelerated, it radiates EM waves. Given the high energy of the electron, these EM waves can have high energy. We call them X-rays. Since the process is random, a broad spectrum of X-ray energy is emitted that is more characteristic of the electron energy than the type of material the electron encounters. Such EM radiation is called "bremsstrahlung" (German for "braking radiation").

X-Rays

In the 1850s, scientists (such as Faraday) began experimenting with high-voltage electrical discharges in tubes filled with rarefied gases. It was later found that these discharges created an invisible, penetrating form of very high frequency electromagnetic radiation. This radiation was called an **X-ray**, because its identity and nature were unknown.

As described in "Things Great and Small," there are two methods by which X-rays are created --both are submicroscopic processes and can be caused by high-voltage discharges. While the low-frequency end of the X-ray range overlaps with the ultraviolet, X-rays extend to much higher frequencies (and energies).

X-rays have adverse effects on living cells similar to those of ultraviolet radiation, and they have the additional liability of being more penetrating, affecting more than the surface layers of cells. Cancer and genetic defects can be induced by exposure to X-rays. Because of their effect on rapidly dividing cells, X-rays can also be used to treat and even cure cancer.

The widest use of X-rays is for imaging objects that are opaque to visible light, such as the human body or aircraft parts. In humans, the risk of cell damage is weighed carefully against the benefit of the diagnostic information obtained. However, questions have risen in recent years as to accidental overexposure of some people during CT scans --a mistake at least in part due to poor monitoring of radiation dose.

The ability of X-rays to penetrate matter depends on density, and so an X-ray image can reveal very detailed density information. Figure 11.4.12 shows an example of the simplest type of X-ray image, an X-ray shadow on film. The amount of information in a simple X-ray image is impressive, but more sophisticated techniques, such as CT scans, can reveal three-dimensional information with details smaller than a millimeter.



Figure 11.4.12: This shadow X-ray image shows many interesting features, such as artificial heart valves, a pacemaker, and the wires used to close the sternum. (credit: P. P. Urone)

The use of X-ray technology in medicine is called radiology -- an established and relatively cheap tool in comparison to more sophisticated technologies. Consequently, X-rays are widely available and used extensively in medical diagnostics. During World War I, mobile X-ray units, advocated by Madame Marie Curie, were used to diagnose soldiers.

Because they can have wavelengths less than 0.01 nm, X-rays can be scattered (a process called X-ray diffraction) to detect the shape of molecules and the structure of crystals. X-ray diffraction was crucial to Crick, Watson, and Wilkins in the determination of the shape of the double-helix DNA molecule.

X-rays are also used as a precise tool for trace-metal analysis in X-ray induced fluorescence, in which the energy of the X-ray emissions are related to the specific types of elements and amounts of materials present.

Gamma Rays

Soon after nuclear radioactivity was first detected in 1896, it was found that at least three distinct types of radiation were being emitted. The most penetrating nuclear radiation was called a **gamma ray** (γ ray) (again a name given because its identity and character were unknown), and it was later found to be an extremely high frequency electromagnetic wave.

In fact, γ rays are any electromagnetic radiation emitted by a nucleus. This can be from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ -ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation.

Gamma rays have characteristics identical to X-rays of the same frequency -- they differ only in source. At higher frequencies, γ rays are more penetrating and more damaging to living tissue. They have many of the same uses as X-rays, including cancer therapy. Gamma radiation from radioactive materials is used in nuclear medicine.

Figure 13 shows a medical image based on γ rays. Food spoilage can be greatly inhibited by exposing it to large doses of γ radiation, thereby obliterating responsible microorganisms. Damage to food cells through irradiation occurs as well, and the long-term hazards of consuming radiation-preserved food are unknown and controversial for some groups. Both X-ray and γ -ray technologies are also used in scanning luggage at airports.

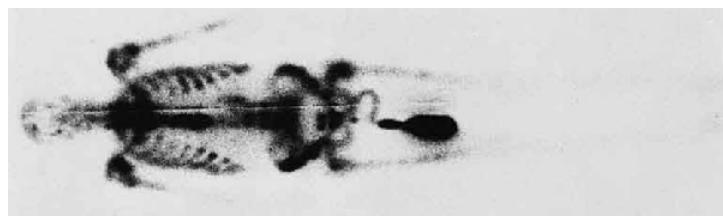


Figure 11.4.13: This is an image of the γ rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Bone cancer is evidenced by nonuniform concentration in similar structures. For example, some ribs are darker than others. (credit: P. P. Urone)

Detecting Electromagnetic Waves from Space

A final note on star gazing. The entire electromagnetic spectrum is used by researchers for investigating stars, space, and time. As noted earlier, Penzias and Wilson detected microwaves to identify the background radiation originating from the Big Bang. Radio

telescopes such as the Arecibo Radio Telescope in Puerto Rico and Parkes Observatory in Australia were designed to detect radio waves.

Infrared telescopes need to have their detectors cooled by liquid nitrogen to be able to gather useful signals. Since infrared radiation is predominantly from thermal agitation, if the detectors were not cooled, the vibrations of the molecules in the antenna would be stronger than the signal being collected.

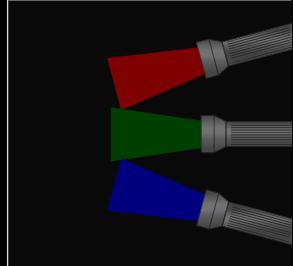
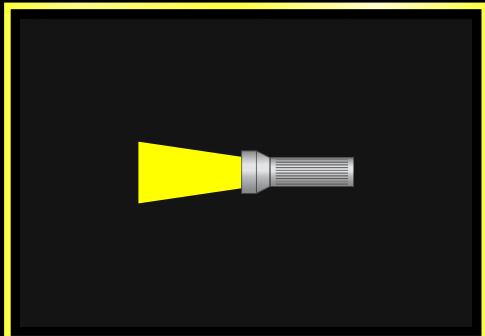
The most famous of these infrared sensitive telescopes is the James Clerk Maxwell Telescope in Hawaii. The earliest telescopes, developed in the seventeenth century, were optical telescopes, collecting visible light. Telescopes in the ultraviolet, X-ray, and γ -ray regions are placed outside the atmosphere on satellites orbiting the Earth.

The Hubble Space Telescope (launched in 1990) gathers ultraviolet radiation as well as visible light. In the X-ray region, there is the Chandra X-ray Observatory (launched in 1999), and in the γ -ray region, there is the new Fermi Gamma-ray Space Telescope (launched in 2008—taking the place of the Compton Gamma Ray Observatory, 1991–2000.).

PHET EXPLORATIONS: COLOR VISION

Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.

Color Vision



RGB Bulbs

Single Bulb

Summary

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v_w = f\lambda$, so that for electromagnetic waves,

$$c = f\lambda,$$

where f is the frequency, λ is the wavelength, and c is the speed of light.

- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.
- Any electromagnetic wave produced by currents in wires is classified as a radio wave, the lowest frequency electromagnetic waves. Radio waves are divided into many types, depending on their applications, ranging up to microwaves at their highest frequencies.
- Infrared radiation lies below visible light in frequency and is produced by thermal motion and the vibration and rotation of atoms and molecules. Infrared's lower frequencies overlap with the highest-frequency microwaves.

- Visible light is largely produced by electronic transitions in atoms and molecules, and is defined as being detectable by the human eye. Its colors vary with frequency, from red at the lowest to violet at the highest.
- Ultraviolet radiation starts with frequencies just above violet in the visible range and is produced primarily by electronic transitions in atoms and molecules.
- X-rays are created in high-voltage discharges and by electron bombardment of metal targets. Their lowest frequencies overlap the ultraviolet range but extend to much higher values, overlapping at the high end with gamma rays.
- Gamma rays are nuclear in origin and are defined to include the highest-frequency electromagnetic radiation of any type.

Glossary

electromagnetic spectrum

the full range of wavelengths or frequencies of electromagnetic radiation

radio waves

electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves

electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

thermal agitation

the thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation

radar

a common application of microwaves. Radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a rainstorm

infrared radiation (IR)

a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from **0.74 μm** to **300 μm**

ultraviolet radiation (UV)

electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light

the narrow segment of the electromagnetic spectrum to which the normal human eye responds

amplitude modulation (AM)

a method for placing information on electromagnetic waves by modulating the amplitude of a carrier wave with an audio signal, resulting in a wave with constant frequency but varying amplitude

extremely low frequency (ELF)

electromagnetic radiation with wavelengths usually in the range of 0 to 300 Hz, but also about 1kHz

carrier wave

an electromagnetic wave that carries a signal by modulation of its amplitude or frequency

frequency modulation (FM)

a method of placing information on electromagnetic waves by modulating the frequency of a carrier wave with an audio signal, producing a wave of constant amplitude but varying frequency

TV

video and audio signals broadcast on electromagnetic waves

very high frequency (VHF)

TV channels utilizing frequencies in the two ranges of 54 to 88 MHz and 174 to 222 MHz

ultra-high frequency (UHF)

TV channels in an even higher frequency range than VHF, of 470 to 1000 MHz

X-ray

invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ -ray range

gamma ray

(γ ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ -ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

This page titled [11.4: The Electromagnetic Spectrum](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [24.3: The Electromagnetic Spectrum](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/college-physics>.

11.5: Polarization

Learning Objectives

By the end of this section, you will be able

- Discuss the meaning of polarization.
- Discuss the property of optical activity of certain materials.

Polaroid sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (Figure 11.5.1). Polaroids have this ability because of a wave characteristic of electromagnetic waves called polarization. What is polarization? How is it produced? What are some of its uses?



Figure 11.5.1: These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit: Amithshs, Wikimedia Commons)

As noted earlier, electromagnetic waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (Figure 11.5.2). There are specific directions for the oscillations of the electric and magnetic fields. **Polarization** is the attribute that a wave's oscillations have a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be **polarized**. For an EM wave, we define the **direction of polarization** to be the direction parallel to the electric field. Thus we can think of the electric field arrows as showing the direction of polarization, as in Figure 11.5.2.

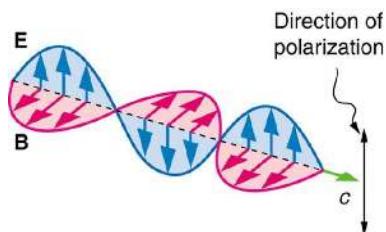


Figure 11.5.2: An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

To examine this further, consider the transverse waves in the ropes shown in Figure 11.5.3. The oscillations in one rope are in a vertical plane and are said to be **vertically polarized**. Those in the other rope are in a horizontal plane and are **horizontally polarized**. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.

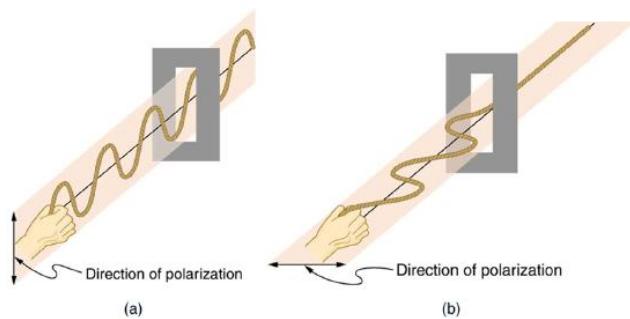


Figure 11.5.3: The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

For the remainder of this section, we will use light as an example of an electromagnetic wave, but many of the characteristics of the polarization of light also apply to other kinds of electromagnetic waves.

As an example, the Sun and many other light sources produce electromagnetic waves that are randomly polarized (Figure 11.5.4). Such light is said to be **unpolarized** because it is composed of many waves with all possible directions of polarization.

Random polarization

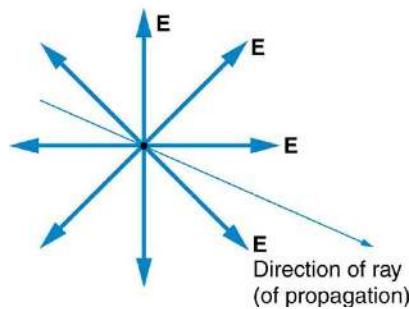


Figure 11.5.4: The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.

Polaroid materials, invented by the founder of Polaroid Corporation, Edwin Land, act as a polarizing slit for light, allowing only polarization in one direction to pass through. Polarizing filters are composed of long molecules aligned in one direction. Thinking of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The axis of a polarizing filter is the direction along which the filter passes the electric field of an EM wave (Figure 11.5.5).

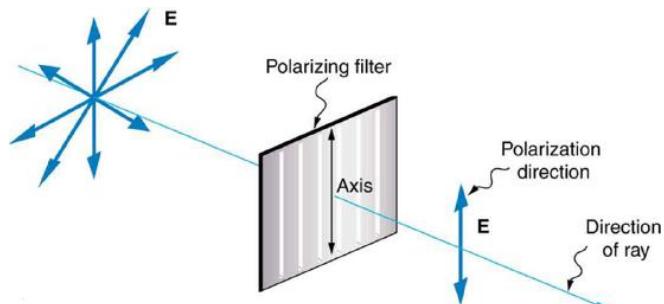


Figure 11.5.5: A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

Figure 11.5.6 shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second. If the second polarizing filter is rotated, only the component of the light parallel to the second filter's axis is passed. When the axes are perpendicular, no light is passed by the second.

Only the component of the EM wave parallel to the axis of a filter is passed. Let us call the angle between the direction of polarization and the axis of a filter θ . If the electric field has an amplitude E , then the transmitted part of the wave has an amplitude $E \cos \theta$ (Figure 11.5.7). Since the intensity of a wave is proportional to its amplitude squared, the intensity I of the transmitted wave is related to the incident wave by

$$I = I_0 \cos^2 \theta, \quad (11.5.1)$$

where I_0 is the intensity of the polarized wave before passing through the filter. Equation 11.5.1 is known as Malus's law.

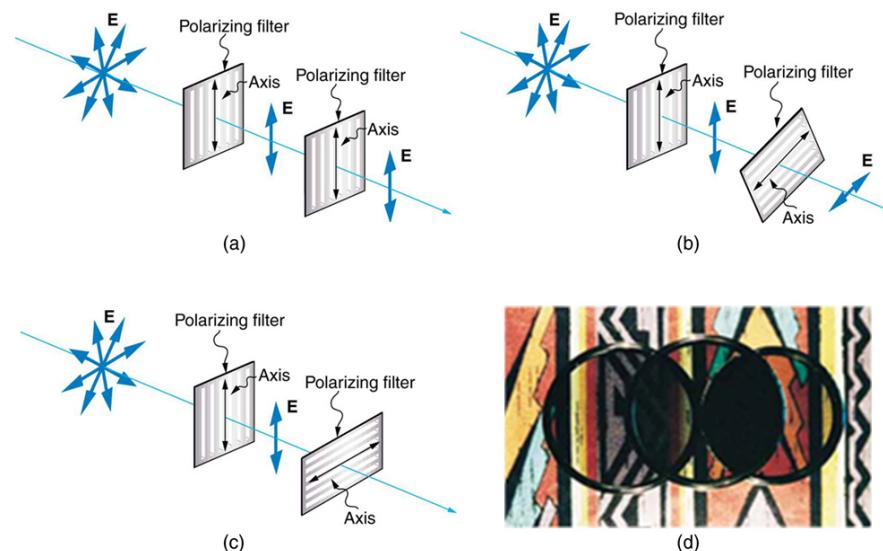


Figure 11.5.6: The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit: P.P. Urone)

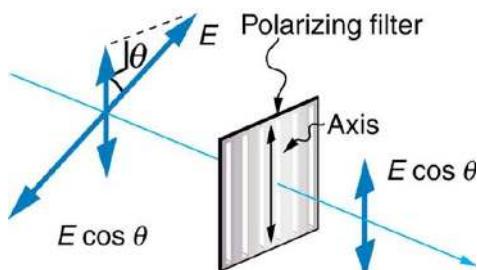


Figure 11.5.7: A polarizing filter transmits only the component of the wave parallel to its axis, $E \cos \theta$, reducing the intensity of any light not polarized parallel to its axis.

✓ Example 11.5.1: Calculating Intensity Reduction by a Polarizing Filter

What angle is needed between the direction of polarized light and the axis of a polarizing filter to reduce its intensity by 90.0%?

Strategy:

When the intensity is reduced by 90.0%, it is 10.0% or 0.100 times its original value. That is, $I = 0.100I_0$. Using this information, the equation $I = I_0 \cos^2 \theta$ can be used to solve for the needed angle.

Solution

Solving the equation $I = I_0 \cos^2 \theta$ for $\cos \theta$ and substituting with the relationship between I and I_0 gives

$$\cos \theta = \sqrt{\frac{I}{I_0}} = \sqrt{\frac{0.100I_0}{I_0}} = 0.3162. \quad (11.5.2)$$

Solving for θ yields

$$\theta = \cos^{-1} 0.3162 = 71.6^\circ \quad (11.5.3)$$

Discussion:

A fairly large angle between the direction of polarization and the filter axis is needed to reduce the intensity to 10.0% of its original value. This seems reasonable based on experimenting with polarizing films. It is interesting that, at an angle of 45°, the intensity is reduced to 50% of its original value (as you will show in this section's Problems & Exercises). Note that 71.6° is 18.4° from reducing the intensity to zero, and that at an angle of 18.4° the intensity is reduced to 90.0% of its original value (as you will also show in Problems & Exercises), giving evidence of symmetry.

Polarization by Reflection

By now you can probably guess that Polaroid sunglasses cut the glare in reflected light because that light is polarized. You can check this for yourself by holding Polaroid sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

Figure 8 illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so that *the reflected light is left more horizontally polarized*. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization would be like an arrow perpendicular to the surface and would be more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and would be more likely to be reflected. Sunglasses with vertical axes would then block more reflected light than unpolarized light from other sources.

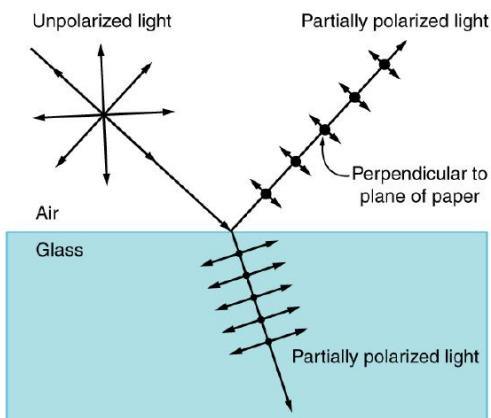


Figure 11.5.8: Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides bouncing off, whereas arrows striking on their tips go into the surface.

Since the part of the light that is not reflected is refracted, the amount of polarization depends on the indices of refraction of the media involved. It can be shown that **reflected light is completely polarized** at an angle of reflection θ_b , given by

$$\tan \theta_b = \frac{n_2}{n_1}, \quad (11.5.4)$$

where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light. This equation is known as **Brewster's law**, and θ_b is known as **Brewster's angle**, named after the 19th-century Scottish physicist who discovered them.

THINGS GREAT AND SMALL: ATOMIC EXPLANATION OF POLARIZING FILTERS:

Polarizing filters have a polarization axis that acts as a slit. This slit passes electromagnetic waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis as shown in

Figure 9.

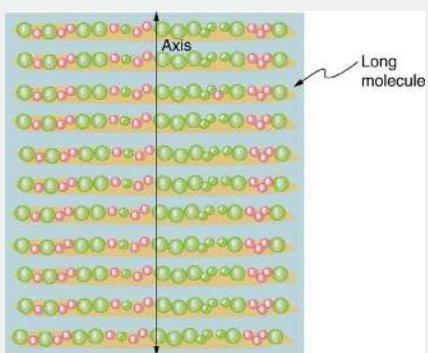


Figure 11.5.9: Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

Figure 10 illustrates how the component of the electric field parallel to the long molecules is absorbed. An electromagnetic wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons in the molecules, since electron masses are small. If the electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the fields in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and will allow those fields to pass. Thus the axis of the polarizing filter is perpendicular to the length of the molecule.

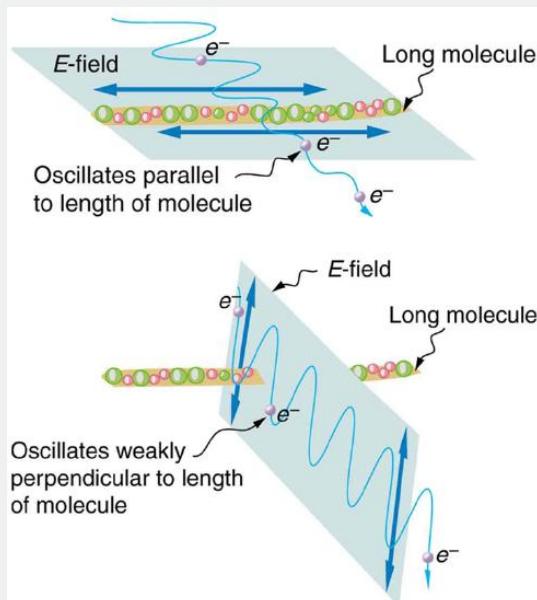


Figure 11.5.10: Artist's conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

✓ Example 11.5.2: Calculating Polarization by Reflection

- At what angle will light traveling in air be completely polarized horizontally when reflected from water?
- From glass?

Strategy:

All we need to solve these problems are the indices of refraction. Air has $n_1 = 1.00$, water has $n_2 = 1.333$, and crown glass has $n'_2 = 1.520$. The equation $\tan \theta_b = \frac{n_2}{n_1}$ can be directly applied to find θ_b in each case.

Solution (a):

Putting the known quantities into the equation

Label '27.9.4' multiply defined

gives

$$\tan \theta_b = \frac{n_2}{n_1} = \frac{1.333}{1.00} = 1.333. \quad (11.5.5)$$

Solving for the angle θ_b yields

$$\theta_b = \tan^{-1} 1.333 = 53.1^\circ. \quad (11.5.6)$$

Solution (b):

Similarly, for crown glass and air,

$$\tan \theta'_b = \frac{n'_2}{n_1} = \frac{1.520}{1.00} = 1.52. \quad (11.5.7)$$

Thus,

$$\theta'_b = \tan^{-1} 1.52 = 56.7^\circ. \quad (11.5.8)$$

Discussion:

Light reflected at these angles could be completely blocked by a good polarizing filter held with its *axis vertical*. Brewster's angle for water and air are similar to those for glass and air, so that sunglasses are equally effective for light reflected from either water or glass under similar circumstances. Light not reflected is refracted into these media. So at an incident angle equal to Brewster's angle, the refracted light will be slightly polarized vertically. It will not be completely polarized vertically, because only a small fraction of the incident light is reflected, and so a significant amount of horizontally polarized light is refracted.

Polarization by Scattering

If you hold your Polaroid sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. Figure 11.5.11 helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in Figure 11, there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light will only be partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.

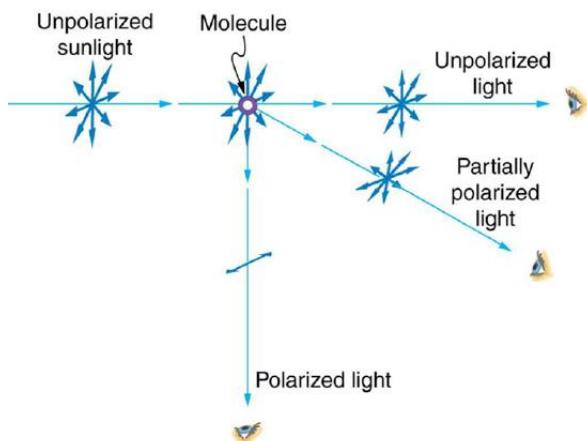


Figure 11.5.11: Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

Photographs of the sky can be darkened by polarizing filters, a trick used by many photographers to make clouds brighter by contrast. Scattering from other particles, such as smoke or dust, can also polarize light. Detecting polarization in scattered EM waves can be a useful analytical tool in determining the scattering source.

There is a range of optical effects used in sunglasses. Besides being Polaroid, other sunglasses have colored pigments embedded in them, while others use non-reflective or even reflective coatings. A recent development is photochromic lenses, which darken in the sunlight and become clear indoors. Photochromic lenses are embedded with organic microcrystalline molecules that change their properties when exposed to UV in sunlight, but become clear in artificial lighting with no UV.

TAKE-HOME EXPERIMENT: POLARIZATION

Find Polaroid sunglasses and rotate one while holding the other still and look at different surfaces and objects. Explain your observations. What is the difference in angle from when you see a maximum intensity to when you see a minimum intensity? Find a reflective glass surface and do the same. At what angle does the glass need to be oriented to give minimum glare?

Liquid Crystals and Other Polarization Effects in Materials

While you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and other myriad places, you may not be aware that they are based on polarization. Liquid crystals are so named because their molecules can be aligned even though they are in a liquid. Liquid crystals have the property that they can rotate the polarization of light passing through them by 90° . Furthermore, this property can be turned off by the application of a voltage, as illustrated in Figure 11.5.12. It is possible to manipulate this characteristic quickly and in small well-defined regions to create the contrast patterns we see in so many LCD devices.

In flat screen LCD televisions, there is a large light at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in Figure 11.5.12a and 11.5.12b. Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. One can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.

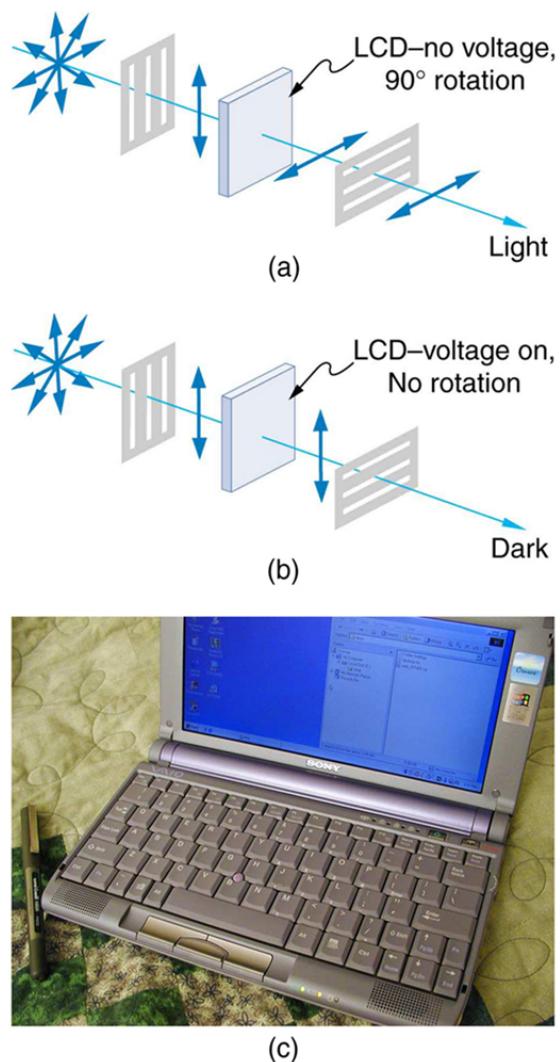


Figure 11.5.12: (a) Polarized light is rotated 90° by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making the region dark in comparison with its surroundings. (c) LCDs can be made color specific, small, and fast enough to use in laptop computers and TVs. (credit: Jon Sullivan)

Many crystals and solutions rotate the plane of polarization of light passing through them. Such substances are said to be **optically active**. Examples include sugar water, insulin, and collagen (Figure 11.5.13). In addition to depending on the type of substance, the amount and direction of rotation depends on a number of factors. Among these is the concentration of the substance, the distance the light travels through it, and the wavelength of light. Optical activity is due to the asymmetric shape of molecules in the substance, such as being helical. Measurements of the rotation of polarized light passing through substances can thus be used to measure concentrations, a standard technique for sugars. It can also give information on the shapes of molecules, such as proteins, and factors that affect their shapes, such as temperature and pH.

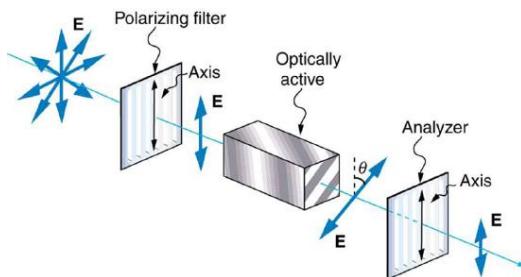


Figure 11.5.13: Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

Glass and plastic become optically active when stressed; the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in Figure 14. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.

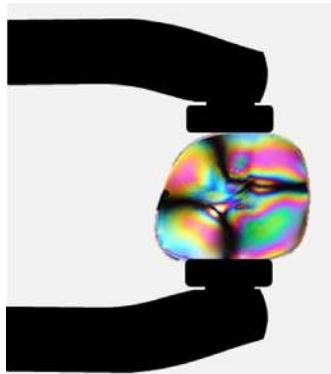


Figure 11.5.14: Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two. Such crystals are said to be **birefringent** (see Figure 15). Each of the separated rays has a specific polarization. One behaves normally and is called the ordinary ray, whereas the other does not obey Snell's law and is called the extraordinary ray. Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work. The interested reader is invited to further pursue the numerous properties of materials related to polarization.

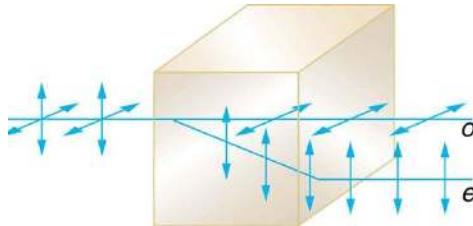


Figure 11.5.15: Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell's law.

Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave.
- EM waves are transverse waves that may be polarized.
- The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.
- Light can be polarized by passing it through a polarizing filter or other polarizing material. The intensity I of polarized light after passing through a polarizing filter is $I = I_0 \cos \theta^2$, where I_0 is the original intensity and θ is the angle between the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Brewster's law states that reflected light will be completely polarized at the angle of reflection θ_b , known as Brewster's angle, given by a statement known as Brewster's law: $\tan \theta_b = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light.
- Polarization can also be produced by scattering.
- There are a number of types of optically active substances that rotate the direction of polarization of light passing through them.

Glossary

axis of a polarizing filter

the direction along which the filter passes the electric field of an EM wave

birefringent

crystals that split an unpolarized beam of light into two beams

Brewster's angle

$\theta_b = \tan\left(\frac{n_2}{n_1}\right)^{-1}$, where n_2 is the index of refraction of the medium from which the light is reflected and n_1 is the index of refraction of the medium in which the reflected light travels

Brewster's law

$\tan\theta_b = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light

direction of polarization

the direction parallel to the electric field for EM waves

horizontally polarized

the oscillations are in a horizontal plane

optically active

substances that rotate the plane of polarization of light passing through them

polarization

the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized

waves having the electric and magnetic field oscillations in a definite direction

reflected light that is completely polarized

light reflected at the angle of reflection θ_b , known as Brewster's angle

unpolarized

waves that are randomly polarized

vertically polarized

the oscillations are in a vertical plane

This page titled [11.5: Polarization](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via source content that was edited to the style and standards of the LibreTexts platform.

- [27.8: Polarization](#) by [OpenStax](#) is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/college-physics>.

11.6: Electromagnetic Waves (Summary)

Key Terms

direction of polarization	direction parallel to the electric field for EM waves
gamma ray (γ ray)	extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons; the lower end of the γ -ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation
horizontally polarized	electric field oscillations are in a horizontal plane
infrared radiation	region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from $0.74\mu\text{m}$ to $300\mu\text{m}$
Malus's law	$I = I_0 \cos^2 \theta$ where I_0 is the intensity of the polarized wave before passing through the filter and θ is the tilt angle of the filter
Maxwell's equations	set of four equations that comprise a complete, overarching theory of electromagnetism
microwaves	electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices
optically active	substances that rotate the plane of polarization of light passing through them
polarized	refers to waves having the electric and magnetic field oscillations in a definite direction
Poynting vector	vector equal to the cross product of the electric-and magnetic fields, that describes the flow of electromagnetic energy through a surface
radar	common application of microwaves; radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a rainstorm
radio waves	electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena
thermal agitation	thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation
ultraviolet radiation	electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm
unpolarized	refers to waves that are randomly polarized
vertically polarized	oscillations are in a vertical plane
visible light	narrow segment of the electromagnetic spectrum to which the normal human eye responds, from about 400 to 750 nm
x-ray	invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ -ray

Key Equations

Speed of EM waves	$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$
Ratio of E field to B field in electromagnetic wave	$c = \frac{E}{B}$
Energy flux (Poynting) vector	$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B}$
Average intensity of an electromagnetic wave	$I = S_{avg} = \frac{c\epsilon_0 E_0^2}{2} = \frac{cB_0^2}{2\mu_0} = \frac{E_0 B_0}{2\mu_0}$
Malus's law	$I = I_0 \cos^2 \theta$

Summary

Maxwell's Equations and Electromagnetic Waves

James Clerk Maxwell (1831–1879) was one of the major contributors to physics in the nineteenth century. Although he died young, he made major contributions to the development of the kinetic theory of gases, to the understanding of color vision, and to the nature of Saturn's rings. He is best known for having combined existing knowledge of the laws of electricity and of magnetism with insights of his own into a complete overarching electromagnetic theory, represented by Maxwell's equations.

- Maxwell's prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell's equations.
- The four Maxwell's equations together with the Lorentz force law encompass the major laws of electricity and magnetism. The first of these is Gauss's law for electricity; the second is Gauss's law for magnetism; the third is Faraday's law of induction (including Lenz's law); and the fourth is Ampère's law in a symmetric formulation that adds another source of magnetism, namely changing electric fields.
- The symmetry introduced between electric and magnetic fields through Maxwell's displacement current explains the mechanism of electromagnetic wave propagation, in which changing magnetic fields produce changing electric fields and vice versa.
- Although light was already known to be a wave, the nature of the wave was not understood before Maxwell. Maxwell's equations also predicted electromagnetic waves with wavelengths and frequencies outside the range of light. These theoretical predictions were first confirmed experimentally by Heinrich Hertz.

Energy Carried by Electromagnetic Waves

- The energy carried by any wave is proportional to its amplitude squared. For electromagnetic waves, this means intensity can be expressed as

$$I = \frac{c\epsilon_0 E_0^2}{2}$$

where I is the average intensity in W/m^2 and E_0 is the maximum electric field strength of a continuous sinusoidal wave. This can also be expressed in terms of the maximum magnetic field strength B_0 as

$$I = \frac{cB_0^2}{2\mu_0}$$

and in terms of both electric and magnetic fields as

$$I = \frac{E_0 B_0}{2\mu_0}$$

The three expressions for I_{avg} are all equivalent.

The Electromagnetic Spectrum

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v = f\lambda$, so that for electromagnetic waves, $c = f\lambda$, where f is the frequency, λ is the wavelength, and c is the speed of light.

- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.

Polarization

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave. The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.
- Unpolarized light can be polarized by passing it through a polarizing filter or other polarizing material. The process of polarizing light decreases its intensity by a factor of 2.
- The intensity, I , of polarized light after passing through a polarizing filter is $I = I_0 \cos^2 \theta$, where I_0 is the incident intensity and θ is the angle between the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Brewster's law states that reflected light is completely polarized at the angle of reflection θ_b , known as Brewster's angle.
- Polarization can also be produced by scattering.
- Several types of optically active substances rotate the direction of polarization of light passing through them.

Contributors and Attributions

Samuel J. Ling (Truman State University), Jeff Sanny (Loyola Marymount University), and Bill Moebs with many contributing authors. This work is licensed by OpenStax University Physics under a [Creative Commons Attribution License \(by 4.0\)](#).

11.6: Electromagnetic Waves (Summary) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax.

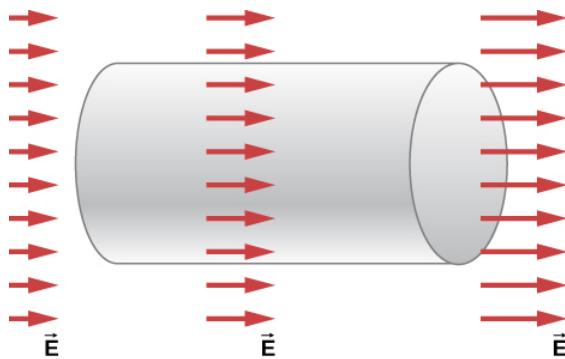
- [16.7: Electromagnetic Waves \(Summary\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.
- [1.S: The Nature of Light \(Summary\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-3>.

11.7: Electromagnetic Waves (Exercises)

Conceptual Questions

16.2 Maxwell's Equations and Electromagnetic Waves

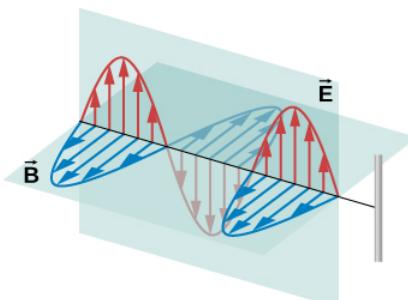
1. Explain how the displacement current maintains the continuity of current in a circuit containing a capacitor.
2. Describe the field lines of the induced magnetic field along the edge of the imaginary horizontal cylinder shown below if the cylinder is in a spatially uniform electric field that is horizontal, pointing to the right, and increasing in magnitude.



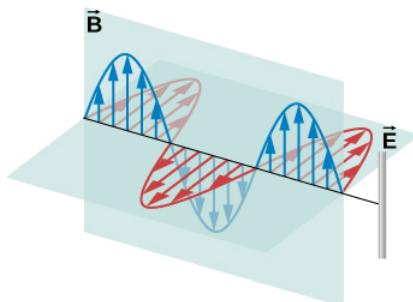
3. Why is it much easier to demonstrate in a student lab that a changing magnetic field induces an electric field than it is to demonstrate that a changing electric field produces a magnetic field?

16.3 Plane Electromagnetic Waves

4. If the electric field of an electromagnetic wave is oscillating along the z-axis and the magnetic field is oscillating along the x-axis, in what possible direction is the wave traveling?
5. In which situation shown below will the electromagnetic wave be more successful in inducing a current in the wire? Explain.

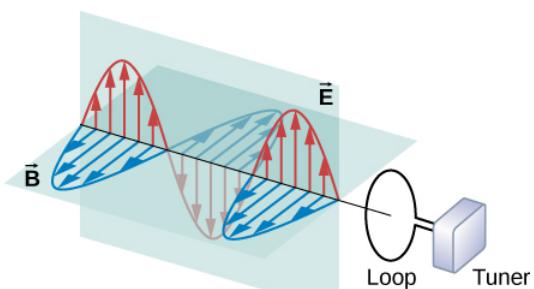


(a)

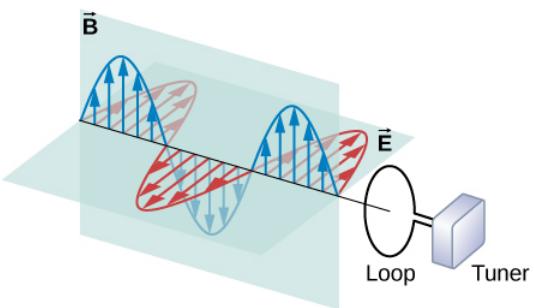


(b)

6. In which situation shown below will the electromagnetic wave be more successful in inducing a current in the loop? Explain.

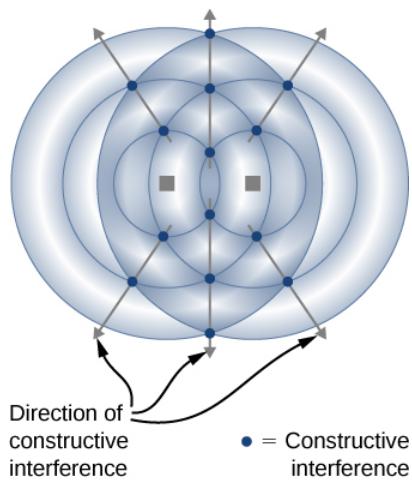


(a)



(b)

7. Under what conditions might wires in a circuit where the current flows in only one direction emit electromagnetic waves?
8. Shown below is the interference pattern of two radio antennas broadcasting the same signal. Explain how this is analogous to the interference pattern for sound produced by two speakers. Could this be used to make a directional antenna system that broadcasts preferentially in certain directions? Explain.



16.4 Energy Carried by Electromagnetic Waves

9. When you stand outdoors in the sunlight, why can you feel the energy that the sunlight carries, but not the momentum it carries?
10. How does the intensity of an electromagnetic wave depend on its electric field? How does it depend on its magnetic field?
11. What is the physical significance of the Poynting vector?

12. A 2.0-mW helium-neon laser transmits a continuous beam of red light of cross-sectional area 0.25cm^2 . If the beam does not diverge appreciably, how would its rms electric field vary with distance from the laser? Explain.

16.5 Momentum and Radiation Pressure

13. Why is the radiation pressure of an electromagnetic wave on a perfectly reflecting surface twice as large as the pressure on a perfectly absorbing surface?

14. Why did the early Hubble Telescope photos of Comet Ison approaching Earth show it to have merely a fuzzy coma around it, and not the pronounced double tail that developed later (see below)?



(credit: ESA, Hubble)

15. (a) If the electric field and magnetic field in a sinusoidal plane wave were interchanged, in which direction relative to before would the energy propagate?

- (b) What if the electric and the magnetic fields were both changed to their negatives?

16.6 The Electromagnetic Spectrum

16. Compare the speed, wavelength, and frequency of radio waves and X-rays traveling in a vacuum.

17. Accelerating electric charge emits electromagnetic radiation. How does this apply in each case: (a) radio waves, (b) infrared radiation.

18. Compare and contrast the meaning of the prefix “micro” in the names of SI units in the term microwaves.

19. Part of the light passing through the air is scattered in all directions by the molecules comprising the atmosphere. The wavelengths of visible light are larger than molecular sizes, and the scattering is strongest for wavelengths of light closest to sizes of molecules.

- (a) Which of the main colors of light is scattered the most?

- (b) Explain why this would give the sky its familiar background color at midday.

20. When a bowl of soup is removed from a microwave oven, the soup is found to be steaming hot, whereas the bowl is only warm to the touch. Discuss the temperature changes that have occurred in terms of energy transfer.

21. Certain orientations of a broadcast television antenna give better reception than others for a particular station. Explain.

22. What property of light corresponds to loudness in sound?

23. Is the visible region a major portion of the electromagnetic spectrum?

24. Can the human body detect electromagnetic radiation that is outside the visible region of the spectrum?

25. Radio waves normally have their **E** and **B** fields in specific directions, whereas visible light usually has its **E** and **B** fields in random and rapidly changing directions that are perpendicular to each other and to the propagation direction. Can you explain why?

26. Give an example of resonance in the reception of electromagnetic waves.

27. Illustrate that the size of details of an object that can be detected with electromagnetic waves is related to their wavelength, by comparing details observable with two different types (for example, radar and visible light).

28. In which part of the electromagnetic spectrum are each of these waves:

- (a) $f = 10.0 \text{ kHz}$,
- (b) $f = \lambda = 750 \text{ nm}$,
- (c) $f = 1.25 \times 10^8 \text{ Hz}$,
- (d) 0.30 nm

29. In what range of electromagnetic radiation are the electromagnetic waves emitted by power lines in a country that uses 50-Hz ac current?

30. If a microwave oven could be modified to merely tune the waves generated to be in the infrared range instead of using microwaves, how would this affect the uneven heating of the oven?

31. A leaky microwave oven in a home can sometimes cause interference with the homeowner's WiFi system. Why?

32. When a television news anchor in a studio speaks to a reporter in a distant country, there is sometimes a noticeable lag between when the anchor speaks in the studio and when the remote reporter hears it and replies. Explain what causes this delay.

Problems

16.2 Maxwell's Equations and Electromagnetic Waves

33. Show that the magnetic field at a distance \mathbf{r} from the axis of two circular parallel plates, produced by placing charge $\mathbf{Q}(t)$ on the plates is

$$\mathbf{B}_{\text{ind}} = \frac{\mu_0}{2\pi r} \frac{dQ(t)}{dt}$$

34. Express the displacement current in a capacitor in terms of the capacitance and the rate of change of the voltage across the capacitor.

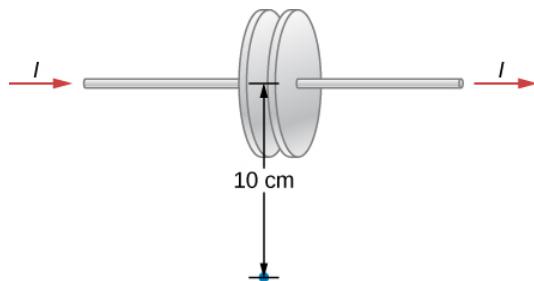
35. A potential difference $V(t) = V_0 \sin \omega t$ is maintained across a parallel-plate capacitor with capacitance \mathbf{C} consisting of two circular parallel plates. A thin wire with resistance \mathbf{R} connects the centers of the two plates, allowing charge to leak between plates while they are charging.

(a) Obtain expressions for the leakage current $I_{\text{res}}(t)$ in the thin wire. Use these results to obtain an expression for the current $I_{\text{real}}(t)$ in the wires connected to the capacitor.

(b) Find the displacement current in the space between the plates from the changing electric field between the plates.

(c) Compare $I_{\text{real}}(t)$ with the sum of the displacement current $I_d(t)$ and resistor current $I_{\text{res}}(t)$ between the plates, and explain why the relationship you observe would be expected.

36. Suppose the parallel-plate capacitor shown below is accumulating charge at a rate of 0.010 C/s. What is the induced magnetic field at a distance of 10 cm from the capacitor?



37. The potential difference $V(t)$ between parallel plates shown above is instantaneously increasing at a rate of $10^7 V/s$. What is the displacement current between the plates if the separation of the plates is 1.00 cm and they have an area of $0.200 m^2$?

38. A parallel-plate capacitor has a plate area of $A = 0.250 m^2$ and a separation of 0.0100 m. What must be the angular frequency ω for a voltage $V(t) = V_0 \sin \omega t$ with $V_0 = 100 V$ to produce a maximum displacement induced current of 1.00 A between the plates?

39. The voltage across a parallel-plate capacitor with area $A = 800 cm^2$ and separation $d = 2 mm$ varies sinusoidally as $V = (15 mV) \cos(150t)$, where t is in seconds. Find the displacement current between the plates.

40. The voltage across a parallel-plate capacitor with area A and separation d varies with time t as $V = at^2$, where a is a constant. Find the displacement current between the plates.

16.3 Plane Electromagnetic Waves

41. If the Sun suddenly turned off, we would not know it until its light stopped coming. How long would that be, given that the Sun is $1.496 \times 10^{11} m$ away?

42. What is the maximum electric field strength in an electromagnetic wave that has a maximum magnetic field strength of $5.00 \times 10^{-4} T$ (about 10 times Earth's magnetic field)?

43. An electromagnetic wave has a frequency of 12 MHz. What is its wavelength in vacuum?

44. If electric and magnetic field strengths vary sinusoidally in time at frequency 1.00 GHz, being zero at $t = 0$, then $E = E_0 \sin 2\pi ft$ and $B = B_0 \sin 2\pi ft$.

(a) When are the field strengths next equal to zero?

(b) When do they reach their most negative value? (c) How much time is needed for them to complete one cycle?

45. The electric field of an electromagnetic wave traveling in vacuum is described by the following wave function:

$$\vec{E} = (5.00 V/m) \cos[kx - (6.00 \times 10^9 s^{-1})t + 0.40] \hat{j}$$

where k is the wavenumber in rad/m, x is in m, t is in s.

Find the following quantities:

(a) amplitude

(b) frequency

(c) wavelength

(d) the direction of the travel of the wave

(e) the associated magnetic field wave

46. A plane electromagnetic wave of frequency 20 GHz moves in the positive y-axis direction such that its electric field is pointed along the z-axis. The amplitude of the electric field is 10 V/m. The start of time is chosen so that at $t = 0$, the electric field has a value 10 V/m at the origin.

(a) Write the wave function that will describe the electric field wave.

(b) Find the wave function that will describe the associated magnetic field wave.

47. The following represents an electromagnetic wave traveling in the direction of the positive y-axis:

$$E_x = 0; E_y = E_0 \cos(kx - \omega t); E_z = 0$$

$$B_x = 0; B_y = 0; B_z = B_0 \cos(kx - \omega t)$$

The wave is passing through a wide tube of circular cross-section of radius R whose axis is along the y-axis. Find the expression for the displacement current through the tube.

16.4 Energy Carried by Electromagnetic Waves

48. While outdoors on a sunny day, a student holds a large convex lens of radius 4.0 cm above a sheet of paper to produce a bright spot on the paper that is 1.0 cm in radius, rather than a sharp focus. By what factor is the electric field in the bright spot of light related to the electric field of sunlight leaving the side of the lens facing the paper?

49. A plane electromagnetic wave travels northward. At one instant, its electric field has a magnitude of 6.0 V/m and points eastward. What are the magnitude and direction of the magnetic field at this instant?

50. The electric field of an electromagnetic wave is given by

$$\mathbf{E} = (6.0 \times 10^{-3} \text{ V/m}) \sin[2\pi(\frac{x}{18m} - \frac{t}{6.0 \times 10^{-8} \text{ s}})] \hat{j}$$

Write the equations for the associated magnetic field and Poynting vector.

51. A radio station broadcasts at a frequency of 760 kHz. At a receiver some distance from the antenna, the maximum magnetic field of the electromagnetic wave detected is $2.15 \times 10^{-11} \text{ T}$.

(a) What is the maximum electric field?

(b) What is the wavelength of the electromagnetic wave?

52. The filament in a clear incandescent light bulb radiates visible light at a power of 5.00 W. Model the glass part of the bulb as a sphere of radius $r_0 = 3.00 \text{ cm}$ and calculate the amount of electromagnetic energy from visible light inside the bulb.

53. At what distance does a 100-W lightbulb produce the same intensity of light as a 75-W lightbulb produces 10 m away? (Assume both have the same efficiency for converting electrical energy in the circuit into emitted electromagnetic energy.)

54. An incandescent light bulb emits only 2.6 W of its power as visible light. What is the rms electric field of the emitted light at a distance of 3.0 m from the bulb?

55. A 150-W lightbulb emits 5% of its energy as electromagnetic radiation. What is the magnitude of the average Poynting vector 10 m from the bulb?

56. A small helium-neon laser has a power output of 2.5 mW. What is the electromagnetic energy in a 1.0-m length of the beam?

57. At the top of Earth's atmosphere, the time-averaged Poynting vector associated with sunlight has a magnitude of about 1.4 kW/m^2 .

(a) What are the maximum values of the electric and magnetic fields for a wave of this intensity?

(b) What is the total power radiated by the sun? Assume that the Earth is $1.5 \times 10^{11} \text{ m}$ from the Sun and that sunlight is composed of electromagnetic plane waves.

58. The magnetic field of a plane electromagnetic wave moving along the z axis is given by

$$\vec{B} = B_0 (\cos kz + \omega t) \hat{j}, \text{ where } B_0 = 5.00 \times 10^{-10} \text{ T} \text{ and } k = 3.14 \times 10^{-2} \text{ m}^{-1}$$

(a) Write an expression for the electric field associated with the wave.

(b) What are the frequency and the wavelength of the wave?

(c) What is its average Poynting vector?

59. What is the intensity of an electromagnetic wave with a peak electric field strength of 125 V/m?

60. Assume the helium-neon lasers commonly used in student physics laboratories have power outputs of 0.500 mW.

- (a) If such a laser beam is projected onto a circular spot 1.00 mm in diameter, what is its intensity?
- (b) Find the peak magnetic field strength.
- (c) Find the peak electric field strength.

61. An AM radio transmitter broadcasts 50.0 kW of power uniformly in all directions. (a) Assuming all of the radio waves that strike the ground are completely absorbed, and that there is no absorption by the atmosphere or other objects, what is the intensity 30.0 km away? (**Hint:** Half the power will be spread over the area of a hemisphere.) (b) What is the maximum electric field strength at this distance?

62. Suppose the maximum safe intensity of microwaves for human exposure is taken to be 1.00 W/m^2 .

- (a) If a radar unit leaks 10.0 W of microwaves (other than those sent by its antenna) uniformly in all directions, how far away must you be to be exposed to an intensity considered to be safe? Assume that the power spreads uniformly over the area of a sphere with no complications from absorption or reflection.
- (b) What is the maximum electric field strength at the safe intensity? (Note that early radar units leaked more than modern ones do. This caused identifiable health problems, such as cataracts, for people who worked near them.)

63. A 2.50-m-diameter university communications satellite dish receives TV signals that have a maximum electric field strength (for one channel) of $7.50 \mu\text{V/m}$ (see below). (a) What is the intensity of this wave? (b) What is the power received by the antenna? (c) If the orbiting satellite broadcasts uniformly over an area of $1.50 \times 10^{13} \text{ m}^2$ (a large fraction of North America), how much power does it radiate?



64. Lasers can be constructed that produce an extremely high intensity electromagnetic wave for a brief time—called pulsed lasers. They are used to initiate nuclear fusion, for example. Such a laser may produce an electromagnetic wave with a maximum electric field strength of $1.00 \times 10^{11} \text{ V/m}$ for a time of 1.00 ns.

- (a) What is the maximum magnetic field strength in the wave?
- (b) What is the intensity of the beam?
- (c) What energy does it deliver on an $1.00 - \text{mm}^2$ area?

16.5 Momentum and Radiation Pressure

65. A 150-W lightbulb emits 5% of its energy as electromagnetic radiation. What is the radiation pressure on an absorbing sphere of radius 10 m that surrounds the bulb?

- 66.** What pressure does light emitted uniformly in all directions from a 100-W incandescent light bulb exert on a mirror at a distance of 3.0 m, if 2.6 W of the power is emitted as visible light?
- 67.** A microscopic spherical dust particle of radius $2\mu\text{m}$ and mass $10\mu\text{g}$ is moving in outer space at a constant speed of 30 cm/sec. A wave of light strikes it from the opposite direction of its motion and gets absorbed. Assuming the particle decelerates uniformly to zero speed in one second, what is the average electric field amplitude in the light?
- 68.** A Styrofoam spherical ball of radius 2 mm and mass $20\mu\text{g}$ is to be suspended by the radiation pressure in a vacuum tube in a lab. How much intensity will be required if the light is completely absorbed by the ball?
- 69.** Suppose that \vec{S}_{avg} for sunlight at a point on the surface of Earth is 900W/m^2 .
- If sunlight falls perpendicularly on a kite with a reflecting surface of area 0.75m^2 , what is the average force on the kite due to radiation pressure?
 - How is your answer affected if the kite material is black and absorbs all sunlight?
- 70.** Sunlight reaches the ground with an intensity of about 1.0kW/m^2 . A sunbather has a body surface area of 0.8m^2 facing the sun while reclining on a beach chair on a clear day.
- how much energy from direct sunlight reaches the sunbather's skin per second?
 - What pressure does the sunlight exert if it is absorbed?
- 71.** Suppose a spherical particle of mass m and radius R in space absorbs light of intensity I for time t .
- How much work does the radiation pressure do to accelerate the particle from rest in the given time it absorbs the light?
 - How much energy carried by the electromagnetic waves is absorbed by the particle over this time based on the radiant energy incident on the particle?

16.6 The Electromagnetic Spectrum

- 72.** How many helium atoms, each with a radius of about 31 pm, must be placed end to end to have a length equal to one wavelength of 470 nm blue light?
- 73.** If you wish to detect details of the size of atoms (about 0.2 nm) with electromagnetic radiation, it must have a wavelength of about this size.
- What is its frequency?
 - What type of electromagnetic radiation might this be?
- 74.** Find the frequency range of visible light, given that it encompasses wavelengths from 380 to 760 nm.
- 75.** (a) Calculate the wavelength range for AM radio given its frequency range is 540 to 1600 kHz.
- Do the same for the FM frequency range of 88.0 to 108 MHz.
- 76.** Radio station WWVB, operated by the National Institute of Standards and Technology (NIST) from Fort Collins, Colorado, at a low frequency of 60 kHz, broadcasts a time synchronization signal whose range covers the entire continental US. The timing of the synchronization signal is controlled by a set of atomic clocks to an accuracy of $1 \times 10^{-12}\text{s}$, and repeats every 1 minute. The signal is used for devices, such as radio-controlled watches, that automatically synchronize with it at preset local times. WWVB's long wavelength signal tends to propagate close to the ground.
- Calculate the wavelength of the radio waves from WWVB.
 - Estimate the error that the travel time of the signal causes in synchronizing a radio controlled watch in Norfolk, Virginia, which is 1570 mi (2527 km) from Fort Collins, Colorado.
- 77.** An outdoor WiFi unit for a picnic area has a 100-mW output and a range of about 30 m. What output power would reduce its range to 12 m for use with the same devices as before? Assume there are no obstacles in the way and that microwaves into the ground are simply absorbed.

- 78.** The prefix “mega” (M) and “kilo” (k), when referring to amounts of computer data, refer to factors of 1024 or 210210 rather than 1000 for the prefix **kilo**, and $1024^2 = 2^{20}$ rather than 1,000,000 for the prefix **Mega** (M). If a wireless (WiFi) router transfers 150 Mbps of data, how many bits per second is that in decimal arithmetic?
- 79.** A computer user finds that his wireless router transmits data at a rate of 75 Mbps (megabits per second). Compare the average time to transmit one bit of data with the time difference between the wifi signal reaching an observer’s cell phone directly and by bouncing back to the observer from a wall 8.00 m past the observer.
- 80.** (a) The ideal size (most efficient) for a broadcast antenna with one end on the ground is one-fourth the wavelength ($\lambda/4$) of the electromagnetic radiation being sent out. If a new radio station has such an antenna that is 50.0 m high, what frequency does it broadcast most efficiently? Is this in the AM or FM band?
- (b) Discuss the analogy of the fundamental resonant mode of an air column closed at one end to the resonance of currents on an antenna that is one-fourth their wavelength.
- 81.** What are the wavelengths of (a) X-rays of frequency $2.0 \times 10^{17} \text{ Hz}$?
 (b) Yellow light of frequency $5.1 \times 10^{14} \text{ Hz}$?
 (c) Gamma rays of frequency $1.0 \times 10^{23} \text{ Hz}$?
- 82.** For red light of $\lambda=660\text{nm}$, what are f, ω and k ?
- 83.** A radio transmitter broadcasts plane electromagnetic waves whose maximum electric field at a particular location is $1.55 \times 10^{-3} \text{ V/m}$. What is the maximum magnitude of the oscillating magnetic field at that location? How does it compare with Earth’s magnetic field?
- 84.** (a) Two microwave frequencies authorized for use in microwave ovens are: 915 and 2450 MHz. Calculate the wavelength of each.
 (b) Which frequency would produce smaller hot spots in foods due to interference effects?
- 85.** During normal beating, the heart creates a maximum 4.00-mV potential across 0.300 m of a person’s chest, creating a 1.00-Hz electromagnetic wave.
 (a) What is the maximum electric field strength created?
 (b) What is the corresponding maximum magnetic field strength in the electromagnetic wave?
 (c) What is the wavelength of the electromagnetic wave?
- 86.** Distances in space are often quoted in units of light-years, the distance light travels in 1 year.
 (a) How many meters is a light-year?
 (b) How many meters is it to Andromeda, the nearest large galaxy, given that it is 2.54×10^6 ly away?
 (c) The most distant galaxy yet discovered is 13.4×10^9 ly away. How far is this in meters?
- 87.** A certain 60.0-Hz ac power line radiates an electromagnetic wave having a maximum electric field strength of 13.0 kV/m.
 (a) What is the wavelength of this very-low-frequency electromagnetic wave?
 (b) What type of electromagnetic radiation is this wave
 (c) What is its maximum magnetic field strength?
- 88.** (a) What is the frequency of the 193-nm ultraviolet radiation used in laser eye surgery? (b) Assuming the accuracy with which this electromagnetic radiation can ablate (reshape) the cornea is directly proportional to wavelength, how much more accurate can this UV radiation be than the shortest visible wavelength of light?

Additional Problems

- 89.** In a region of space, the electric field is pointed along the x-axis, but its magnitude changes as described by

$$E_x = (10N/C)\sin(20x - 500t)$$

$$E_y = E_z = 0$$

where t is in nanoseconds and x is in cm. Find the displacement current through a circle of radius 3 cm in the $x = 0$ plane at $t = 0$.

90. A microwave oven uses electromagnetic waves of frequency $f = 2.45 \times 10^9 \text{ Hz}$ to heat foods. The waves reflect from the inside walls of the oven to produce an interference pattern of standing waves whose antinodes are hot spots that can leave observable pit marks in some foods. The pit marks are measured to be 6.0 cm apart. Use the method employed by Heinrich Hertz to calculate the speed of electromagnetic waves this implies.

Use the Appendix D for the next two exercises

91. Galileo proposed measuring the speed of light by uncovering a lantern and having an assistant a known distance away uncover his lantern when he saw the light from Galileo's lantern, and timing the delay. How far away must the assistant be for the delay to equal the human reaction time of about 0.25 s?

92. Show that the wave equation in one dimension

$$\frac{\partial^2 f}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 f}{\partial t^2}$$

is satisfied by any doubly differentiable function of either the form $f(x - vt)$ or $f(x + vt)$.

93. On its highest power setting, a microwave oven increases the temperature of 0.400 kg of spaghetti by 45.0°C in 120 s.

(a) What was the rate of energy absorption by the spaghetti, given that its specific heat is $3.76 \times 10^3 \text{ J/kg}\cdot^\circ\text{C}$? Assume the spaghetti is perfectly absorbing.

(b) Find the average intensity of the microwaves, given that they are absorbed over a circular area 20.0 cm in diameter.

(c) What is the peak electric field strength of the microwave?

(d) What is its peak magnetic field strength?

94. A certain microwave oven projects 1.00 kW of microwaves onto a 30-cm-by-40-cm area.

(a) What is its intensity in W/m^2 ?

(b) Calculate the maximum electric field strength E_0 in these waves.

(c) What is the maximum magnetic field strength B_0 ?

95. Electromagnetic radiation from a 5.00-mW laser is concentrated on a $1.00 - \text{mm}^2$ area.

(a) What is the intensity in W/m^2 ?

(b) Suppose a 2.00-nC electric charge is in the beam. What is the maximum electric force it experiences?

(c) If the electric charge moves at 400 m/s, what maximum magnetic force can it feel?

96. A 200-turn flat coil of wire 30.0 cm in diameter acts as an antenna for FM radio at a frequency of 100 MHz. The magnetic field of the incoming electromagnetic wave is perpendicular to the coil and has a maximum strength of $1.00 \times 10^{-12} \text{ T}$.

(a) What power is incident on the coil?

(b) What average emf is induced in the coil over one-fourth of a cycle?

(c) If the radio receiver has an inductance of $2.50 \mu\text{H}$, what capacitance must it have to resonate at 100 MHz?

97. Suppose a source of electromagnetic waves radiates uniformly in all directions in empty space where there are no absorption or interference effects.

(a) Show that the intensity is inversely proportional to r^2 , the distance from the source squared.

(b) Show that the magnitudes of the electric and magnetic fields are inversely proportional to r .

98. A radio station broadcasts its radio waves with a power of 50,000 W. What would be the intensity of this signal if it is received on a planet orbiting Proxima Centauri, the closest star to our Sun, at 4.243 ly away?

99. The Poynting vector describes a flow of energy whenever electric and magnetic fields are present. Consider a long cylindrical wire of radius r with a current \mathbf{I} in the wire, with resistance \mathbf{R} and voltage \mathbf{V} . From the expressions for the electric field along the wire and the magnetic field around the wire, obtain the magnitude and direction of the Poynting vector at the surface. Show that it accounts for an energy flow into the wire from the fields around it that accounts for the Ohmic heating of the wire.

100. The Sun's energy strikes Earth at an intensity of 1.37kW/m^2 . Assume as a model approximation that all of the light is absorbed. (Actually, about 30% of the light intensity is reflected out into space.)

(a) Calculate the total force that the Sun's radiation exerts on Earth.

(b) Compare this to the force of gravity between the Sun and Earth.

Note: Earth's mass is $5.972 \times 10^{24}\text{kg}$.

101. If a **Lightsail** spacecraft were sent on a Mars mission, by what fraction would its propulsion force be reduced when it reached Mars?

102. Lunar astronauts placed a reflector on the Moon's surface, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time.

(a) To what accuracy in meters can the distance to the Moon be determined, if this time can be measured to 0.100 ns?

(b) What percent accuracy is this, given the average distance to the Moon is 384,400 km?

103. Radar is used to determine distances to various objects by measuring the round-trip time for an echo from the object.

(a) How far away is the planet Venus if the echo time is 1000 s?

(b) What is the echo time for a car 75.0 m from a highway police radar unit?

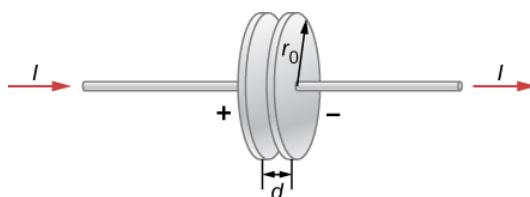
(c) How accurately (in nanoseconds) must you be able to measure the echo time to an airplane 12.0 km away to determine its distance within 10.0 m?

104. Calculate the ratio of the highest to lowest frequencies of electromagnetic waves the eye can see, given the wavelength range of visible light is from 380 to 760 nm. (Note that the ratio of highest to lowest frequencies the ear can hear is 1000.)

105. How does the wavelength of radio waves for an AM radio station broadcasting at 1030 KHz compare with the wavelength of the lowest audible sound waves (of 20 Hz). The speed of sound in air at 20°C is about 343 m/s.

Challenge Problems

106. A parallel-plate capacitor with plate separation d is connected to a source of emf that places a time-dependent voltage $\mathbf{V(t)}$ across its circular plates of radius r_0 and area $A = \pi r_0^2$ (see below).



(a) Write an expression for the time rate of change of energy inside the capacitor in terms of $\mathbf{V(t)}$ and $d\mathbf{V(t)}/dt$.

(b) Assuming that $\mathbf{V(t)}$ is increasing with time, identify the directions of the electric field lines inside the capacitor and of the magnetic field lines at the edge of the region between the plates, and then the direction of the Poynting vector \vec{S} at this location.

(c) Obtain expressions for the time dependence of $\mathbf{E(t)}$, for $\mathbf{B(t)}$ from the displacement current, and for the magnitude of the Poynting vector at the edge of the region between the plates.

(d) From \vec{S} , obtain an expression in terms of $\mathbf{V}(t)$ and $d\mathbf{V}(t)/dt$ for the rate at which electromagnetic field energy enters the region between the plates.

(e) Compare the results of parts (a) and (d) and explain the relationship between them.

107. A particle of cosmic dust has a density $\rho = 2.0 \text{ g/cm}^3$.

(a) Assuming the dust particles are spherical and light absorbing, and are at the same distance as Earth from the Sun, determine the particle size for which radiation pressure from sunlight is equal to the Sun's force of gravity on the dust particle.

(b) Explain how the forces compare if the particle radius is smaller.

(c) Explain what this implies about the sizes of dust particle likely to be present in the inner solar system compared with outside the Oort cloud.

This page titled [11.7: Electromagnetic Waves \(Exercises\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [16.8: Electromagnetic Waves \(Exercises\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

11.8: Electromagnetic Waves (Answer)

Check Your Understanding

16.1. It is greatest immediately after the current is switched on. The displacement current and the magnetic field from it are proportional to the rate of change of electric field between the plates, which is greatest when the plates first begin to charge.

16.2. No. The changing electric field according to the modified version of Ampère's law would necessarily induce a changing magnetic field.

16.3. (1) Faraday's law, (2) the Ampère-Maxwell law

16.4. a. The directions of wave propagation, of the **E** field, and of **B** field are all mutually perpendicular.

b. The speed of the electromagnetic wave is the speed of light $c = 1/\sqrt{\epsilon_0\mu_0}$ independent of frequency.

c. The ratio of electric and magnetic field amplitudes is $E/B = c$.

16.5. Its acceleration would decrease because the radiation force is proportional to the intensity of light from the Sun, which decreases with distance. Its speed, however, would not change except for the effects of gravity from the Sun and planets.

16.6. They fall into different ranges of wavelength, and therefore also different corresponding ranges of frequency.

Conceptual Questions

1. The current into the capacitor to change the electric field between the plates is equal to the displacement current between the plates.

3. The first demonstration requires simply observing the current produced in a wire that experiences a changing magnetic field. The second demonstration requires moving electric charge from one location to another, and therefore involves electric currents that generate a changing electric field. The magnetic fields from these currents are not easily separated from the magnetic field that the displacement current produces.

5. in (a), because the electric field is parallel to the wire, accelerating the electrons

7. A steady current in a dc circuit will not produce electromagnetic waves. If the magnitude of the current varies while remaining in the same direction, the wires will emit electromagnetic waves, for example, if the current is turned on or off.

9. The amount of energy (about $100W/m^2$) is can quickly produce a considerable change in temperature, but the light pressure (about $3.00 \times 10^{-7}N/m^2$) is much too small to notice.

11. It has the magnitude of the energy flux and points in the direction of wave propagation. It gives the direction of energy flow and the amount of energy per area transported per second.

13. The force on a surface acting over time Δt is the momentum that the force would impart to the object. The momentum change of the light is doubled if the light is reflected back compared with when it is absorbed, so the force acting on the object is twice as great.

15. a. According to the right hand rule, the direction of energy propagation would reverse.

b. This would leave the vector \vec{S} , and therefore the propagation direction, the same.

17. a. Radio waves are generally produced by alternating current in a wire or an oscillating electric field between two plates;

b. Infrared radiation is commonly produced by heated bodies whose atoms and the charges in them vibrate at about the right frequency.

19. a. blue;

b. Light of longer wavelengths than blue passes through the air with less scattering, whereas more of the blue light is scattered in different directions in the sky to give it its blue color.

21. A typical antenna has a stronger response when the wires forming it are orientated parallel to the electric field of the radio wave.

23. No, it is very narrow and just a small portion of the overall electromagnetic spectrum.
25. Visible light is typically produced by changes of energies of electrons in randomly oriented atoms and molecules. Radio waves are typically emitted by an ac current flowing along a wire, that has fixed orientation and produces electric fields pointed in particular directions.
27. Radar can observe objects the size of an airplane and uses radio waves of about 0.5 cm in wavelength. Visible light can be used to view single biological cells and has wavelengths of about 10^{-7} m .
29. ELF radio waves
31. The frequency of 2.45 GHz of a microwave oven is close to the specific frequencies in the 2.4 GHz band used for WiFi.

Problems

33. $B_{ind} = \frac{\mu_0}{P2\pi r} I_{ind} = \frac{\mu_0}{2\pi r} \epsilon_0 \frac{\partial \Phi_E}{\partial t} = \frac{\mu_0}{2\pi r} \epsilon_0 (A \frac{\partial E}{\partial t}) = \frac{\mu_0}{2\pi r} \epsilon_0 A (\frac{1}{d} \frac{dV(t)}{dt}) = \frac{\mu_0}{2\pi r} [\frac{\epsilon_0 A}{d}] [\frac{1}{C} \frac{dQ(t)}{dt}] = \frac{\mu_0}{2\pi r} \frac{dQ(t)}{dt}$ because C

$$= \frac{\epsilon_0 A}{d}$$

35. a. $I_{res} = \frac{V_0 \sin \omega t}{R}$;

b. $I_d = CV_0 \omega \cos \omega t$;

c. $I_{real} = I_{res} + \frac{dQ}{dt} = \frac{V_0 \sin \omega t}{R} + CV_0 \frac{d}{dt} \sin \omega t = \frac{V_0 \sin \omega t}{R} + CV_0 \omega \cos \omega t$; which is the sum of I_{res} and I_{real} , consistent with how the displacement current maintaining the continuity of current.

37. $1.77 \times 10^{-3} \text{ A}$

39. $I_d = (7.97 \times 10^{-10} \text{ A}) \sin(150t)$

41. 499 s

43. 25 m

45. a. 5.00 V/m;

b. $9.55 \times 10^8 \text{ Hz}$;

c. 31.4 cm;

d. toward the +x-axis;

e. $B = (1.67 \times 10^{-8} T) \cos[kx - (6 \times 10^9 \text{ s}^{-1})t + 0.40] \hat{k}$

47. $I_d = \pi \epsilon_0 \omega R^2 E_0 \sin(kx - \omega t)$

49. The magnetic field is downward, and it has magnitude $2.00 \times 10^{-8} \text{ T}$.

51. a. $6.45 \times 10^{-3} \text{ V/m}$;

b. 394 m

53. 11.5 m

55. $5.97 \times 10^{-3} \text{ W/m}^2$

57. a. $E_0 = 1027 \text{ V/m}$, $B_0 = 3.42 \times 10^{-6} \text{ T}$;

b. $3.96 \times 10^{26} \text{ W}$

59. 20.8 W/m^2

61. a. $4.42 \times 10^{-6} \text{ W/m}^2$;

b. $5.77 \times 10^{-2} \text{ V/m}$

63. a. $7.47 \times 10^{-14} \text{ W/m}^2$;

b. $3.66 \times 10^{-13} \text{ W}$;

c. 1.12 W

65. $1.99 \times 10^{-11} N/m^2$

67. $F = ma = (p)(\pi r^2), p = \frac{ma}{\pi r^2} = \frac{\epsilon_0}{2E_0^2}$

$$E_0 = \sqrt{\frac{2ma}{\epsilon_0 \pi r^2}} = \sqrt{\frac{2(10^{-8} kg)(0.30 m/s^2)}{(8.854 \times 10^{-12} C^2/N \cdot m^2)(\pi)(2 \times 10^{-6} m)^2}}$$

$$E_0 = 7.34 \times 10^6 V/m$$

69. a. $4.50 \times 10^{-6} N$;

b. it is reduced to half the pressure, $2.25 \times 10^{-6} N$

71. a. $W = \frac{1}{2} \frac{\pi^2 r^4}{mc^2} I^2 t^2$;

b. $E = \pi r^2 It$

73. a. $1.5 \times 10^{18} Hz$;

b. X-rays

75. a. The wavelength range is 187 m to 556 m.

b. The wavelength range is 2.78 m to 3.41 m.

77. $P' = (\frac{12m}{30m})^2 (100mW) = 16mW$

79. time for 1 bit = 1.27×10^{-8} s, difference in travel time is 5.34×10^{-8} s

81. a. $1.5 \times 10^{-9} m$;

b. $5.9 \times 10^{-7} m$;

c. $3.0 \times 10^{-15} m$

83. $5.17 \times 10^{-12} T$, the non-oscillating geomagnetic field of 25–65 μT is much larger

85. a. $1.33 \times 10^{-2} V/m$;

b. $4.34 \times 10^{-11} T$;

c. $3.00 \times 10^8 m$

87. a. $5.00 \times 10^6 m$;

b. radio wave;

c. $4.33 \times 10^{-5} T$

Additional Problems

89. $I_d = (10N/C)(8.845 \times 10^{-12} C^2/N \cdot m^2)\pi(0.03m)^2(5000) = 1.25 \times 10^{-5} mA$

91. $3.75 \times 10^7 km$, which is much greater than Earth's circumference

93. a. 564 W;

b. $1.80 \times 10^4 W/m^2$;

c. $3.68 \times 10^3 V/m$;

d. $1.23 \times 10^{-5} T$

95. a. $5.00 \times 10^3 W/m^2$;

b. $3.88 \times 10^{-6} N$;

c. $5.18 \times 10^{-12} N$

— —

97. a. $I = \frac{P}{A} = \frac{P}{4\pi r^2} \propto \frac{1}{r^2}$;

b. $I \propto E_0^2, B_0^2 \Rightarrow E_0^2, B_0^2 \propto \frac{1}{r^2} \Rightarrow E_0, B_0 \propto \frac{1}{r}$

99. Power into the wire = $\int \vec{S} \cdot d\vec{A} = (\frac{1}{\mu_0} EB)(2\pi r L) = \frac{1}{\mu_0} (\frac{V}{L}) (\frac{\mu_0 i}{2\pi r})(2\pi r L) = iV = i^2 R$

101. 0.431

103. a. $1.5 \times 10^{11} m$;

b. $5.0 \times 10^{-7} s$;

c. 33 ns

105. sound : $\lambda_{sound} = \frac{v_s}{f} = \frac{343 m/s}{20.0 Hz} = 17.2 m$

radio : $\lambda_{radio} = \frac{c}{f} = \frac{3.00 \times 10^8 m/s}{1030 \times 10^3 Hz} = 291 m; or 17.1 \lambda_{sound}$

Challenge Problems

107. a. $0.29 \mu m$;

b. The radiation pressure is greater than the Sun's gravity if the particle size is smaller, because the gravitational force varies as the radius cubed while the radiation pressure varies as the radius squared.

c. The radiation force outward implies that particles smaller than this are less likely to be near the Sun than outside the range of the Sun's radiation pressure.

This page titled [11.8: Electromagnetic Waves \(Answer\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [16.9: Electromagnetic Waves \(Answer\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

CHAPTER OVERVIEW

12: Antenna Systems

- [12.1: Introduction](#)
- [12.2: Production of Electromagnetic Waves](#)
- [12.3: Transmission Lines and Characteristic Impedance](#)
- [12.4: Finite-length Transmission Lines](#)
- [12.5: “Long” and “Short” Transmission Lines](#)
- [12.6: Standing Waves and Resonance](#)
- [12.7: Antenna Systems \(Summary\)](#)

12: Antenna Systems is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

12.1: Introduction

We have seen that Maxwell's equations predict that electromagnetic waves are possible. We would like to understand how electromagnetic waves are created. We will explore a basic model of an **antenna**. However, in practice, antennas not usually connected to radios directly, but instead are connected with a cable called a **transmission line** or **feed line**. We will explore a simple model for how a transmission line can transfer electromagnetic energy through it from one end to the other.

12.1: Introduction is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Ronald Kumon.

12.2: Production of Electromagnetic Waves

Learning Objectives

By the end of this section, you will be able to:

- Describe the electric and magnetic waves as they move out from a source, such as an AC generator.
- Explain the mathematical relationship between the magnetic field strength and the electrical field strength.
- Calculate the maximum strength of the magnetic field in an electromagnetic wave, given the maximum electric field strength.

We can get a good understanding of **electromagnetic waves** (EM) by considering how they are produced. Whenever a current varies, associated electric and magnetic fields vary, moving out from the source like waves. Perhaps the easiest situation to visualize is a varying current in a long straight wire, produced by an AC generator at its center, as illustrated in Figure 12.2.1.

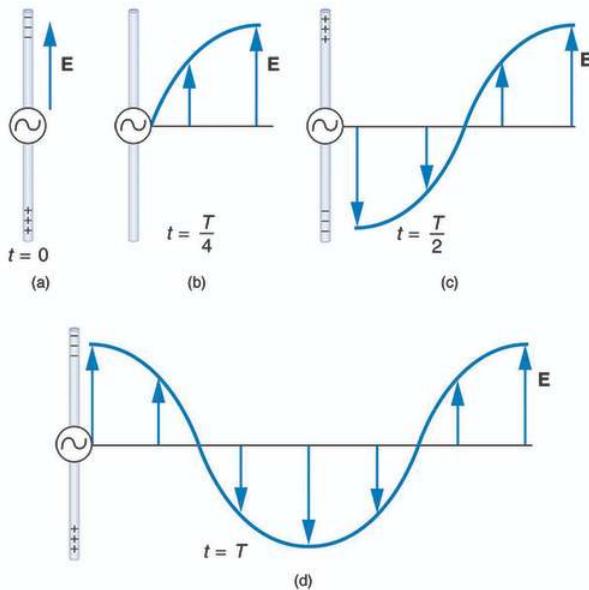


Figure 12.2.1: This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field (**E**) propagates away from the antenna at the speed of light, forming part of an electromagnetic wave.

The **electric field (E)** shown surrounding the wire is produced by the charge distribution on the wire. Both the **E** and the charge distribution vary as the current changes. The changing field propagates outward at the speed of light.

There is an associated **magnetic field (B)** which propagates outward as well (Figure 12.2.2). The electric and magnetic fields are closely related and propagate as an electromagnetic wave. This is what happens in broadcast antennae such as those in radio and TV stations.

Closer examination of the one complete cycle shown in Figure 12.2.1 reveals the periodic nature of the generator-driven charges oscillating up and down in the antenna and the electric field produced. At time $t = 0$, there is the maximum separation of charge, with negative charges at the top and positive charges at the bottom, producing the maximum magnitude of the electric field (or **E**-field) in the upward direction. One-fourth of a cycle later, there is no charge separation and the field next to the antenna is zero, while the maximum **E**-field has moved away at speed c .

As the process continues, the charge separation reverses and the field reaches its maximum downward value, returns to zero, and rises to its maximum upward value at the end of one complete cycle. The outgoing wave has an **amplitude** proportional to the maximum separation of charge. Its **wavelength (λ)** is proportional to the period of the oscillation and, hence, is smaller for short periods or high frequencies. (As usual, wavelength and **frequency (f)** are inversely proportional.)

Electric and Magnetic Waves: Moving Together

Following Ampere's law, current in the antenna produces a magnetic field, as shown in Figure 12.2.2. The relationship between \mathbf{E} and \mathbf{B} is shown at one instant in Figure 2a. As the current varies, the magnetic field varies in magnitude and direction.

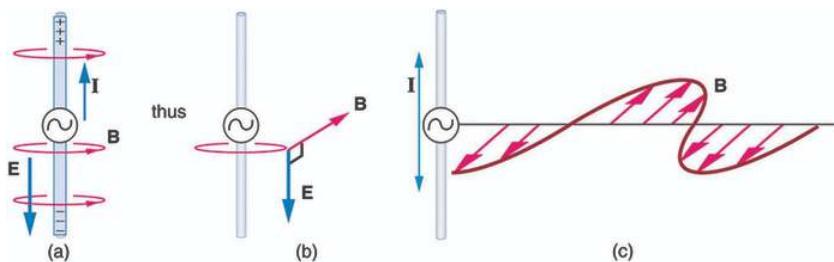


Figure 12.2.2: (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields (E and B) near the wire are perpendicular; they are shown here for one point in space. (c) The magnetic field varies with current and propagates away from the antenna at the speed of light.

The magnetic field lines also propagate away from the antenna at the speed of light, forming the other part of the electromagnetic wave, as seen in Figure 12.2.2b. The magnetic part of the wave has the same period and wavelength as the electric part, since they are both produced by the same movement and separation of charges in the antenna.

The electric and magnetic waves are shown together at one instant in time in Figure 12.2.3. The electric and magnetic fields produced by a long straight wire antenna are exactly in phase. Note that they are perpendicular to one another and to the direction of propagation, making this a **transverse wave**.

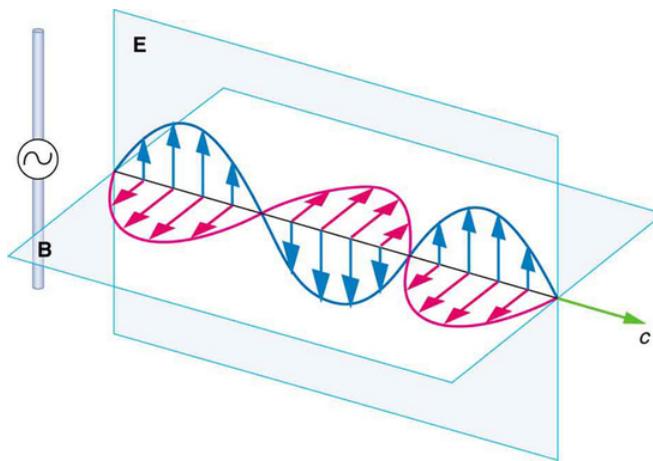


Figure 12.2.3: A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields (E and B) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one direction, but they propagate out in other directions too.

Electromagnetic waves generally propagate out from a source in all directions, sometimes forming a complex radiation pattern. A linear antenna like this one will not radiate parallel to its length, for example. The wave is shown in one direction from the antenna in Figure 12.2.3 to illustrate its basic characteristics.

Instead of the AC generator, the antenna can also be driven by an AC circuit. In fact, charges radiate whenever they are accelerated. But while a current in a circuit needs a complete path, an antenna has a varying charge distribution forming a **standing wave**, driven by the AC. The dimensions of the antenna are critical for determining the frequency of the radiated electromagnetic waves. This is a **resonant** phenomenon and when we tune radios or TV, we vary electrical properties to achieve appropriate resonant conditions in the antenna.

Receiving Electromagnetic Waves

Electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving EM signals works in reverse. And like antennas that produce EM waves, receiver

antennas are specially designed to resonate at particular frequencies.

An incoming electromagnetic wave accelerates electrons in the antenna, setting up a standing wave. If the radio or TV is switched on, electrical components pick up and amplify the signal formed by the accelerating electrons. The signal is then converted to audio and/or video format. Sometimes big receiver dishes are used to focus the signal onto an antenna.

In fact, charges radiate whenever they are accelerated. When designing circuits, we often assume that energy does not quickly escape AC circuits, and mostly this is true. A broadcast antenna is specially designed to enhance the rate of electromagnetic radiation, and shielding is necessary to keep the radiation close to zero. Some familiar phenomena are based on the production of electromagnetic waves by varying currents. Your microwave oven, for example, sends electromagnetic waves, called microwaves, from a concealed antenna that has an oscillating current imposed on it.

Relating E -Field and B -Field Strengths

There is a relationship between the E - and B - field strengths in an electromagnetic wave. This can be understood by again considering the antenna just described. The stronger the E -field created by a separation of charge, the greater the current and, hence, the greater the B -field created.

Since current is directly proportional to voltage (Ohm's law) and voltage is directly proportional to E -field strength, the two should be directly proportional. It can be shown that the magnitudes of the fields do have a constant ratio, equal to the speed of light. That is,

$$\frac{E}{B} = c \quad (12.2.1)$$

is the ratio of E -field strength to B -field strength in any electromagnetic wave. This is true at all times and at all locations in space. A simple and elegant result.

✓ Example 12.2.1: Calculating B -Field Strength in an Electromagnetic Wave

What is the maximum strength of the B -field in an electromagnetic wave that has a maximum E -field strength of $1000V/m$?

Strategy:

To find the B -field strength, we rearrange the Equation 12.2.1 to solve for B , yielding

$$B = \frac{E}{c}. \quad (12.2.2)$$

Solution:

We are given E , and c is the speed of light. Entering these into the expression for B yields

$$B = \frac{1000V/m}{3.00 \times 10^8 m/s} = 3.33 \times 10^{-6} T,$$

Where T stands for Tesla, a measure of magnetic field strength.

Discussion:

The B -field strength is less than a tenth of the Earth's admittedly weak magnetic field. This means that a relatively strong electric field of 1000 V/m is accompanied by a relatively weak magnetic field. Note that as this wave spreads out, say with distance from an antenna, its field strengths become progressively weaker.

The result of this example is consistent with the statement made in the module 24.2 that changing electric fields create relatively weak magnetic fields. They can be detected in electromagnetic waves, however, by taking advantage of the phenomenon of resonance, as Hertz did. A system with the same natural frequency as the electromagnetic wave can be made to oscillate. All radio and TV receivers use this principle to pick up and then amplify weak electromagnetic waves, while rejecting all others not at their resonant frequency.

💡 TAKE-HOME EXPERIMENT: ANTENNAS

For your TV or radio at home, identify the antenna, and sketch its shape. If you don't have cable, you might have an outdoor or indoor TV antenna. Estimate its size. If the TV signal is between 60 and 216 MHz for basic channels, then what is the wavelength of those EM waves?

Try tuning the radio and note the small range of frequencies at which a reasonable signal for that station is received. (This is easier with digital readout.) If you have a car with a radio and extendable antenna, note the quality of reception as the length of the antenna is changed.

💡 PHET EXPLORATIONS: RADIO WAVES AND ELECTROMAGNETIC FIELDS

Broadcast radio waves from [KPhET](#). Wiggle the transmitter electron manually or have it oscillate automatically. Display the field as a curve or vectors. The strip chart shows the electron positions at the transmitter and at the receiver.

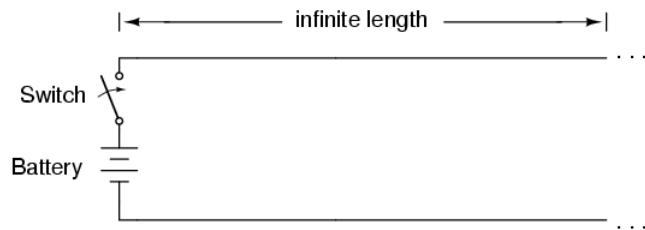
[12.2: Production of Electromagnetic Waves](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by LibreTexts.

- [24.2: Production of Electromagnetic Waves](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/college-physics>.

12.3: Transmission Lines and Characteristic Impedance

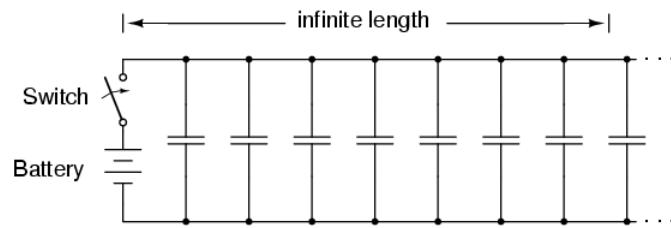
The Parallel Wires of Infinite Length

Suppose, though, that we had a set of parallel wires of *infinite* length, with no lamp at the end. What would happen when we close the switch? Being that there is no longer a load at the end of the wires, this circuit is open. Would there be no current at all? (Figure below)



Driving an infinite transmission line.

Despite being able to avoid wire resistance through the use of superconductors in this “thought experiment,” we cannot eliminate capacitance along the wires’ lengths. Any pair of conductors separated by an insulating medium creates capacitance between those conductors: (Figure below)

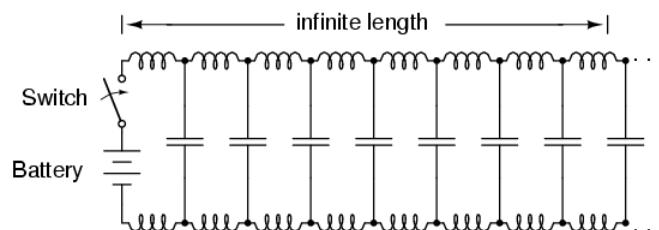


Equivalent circuit showing stray capacitance between conductors.

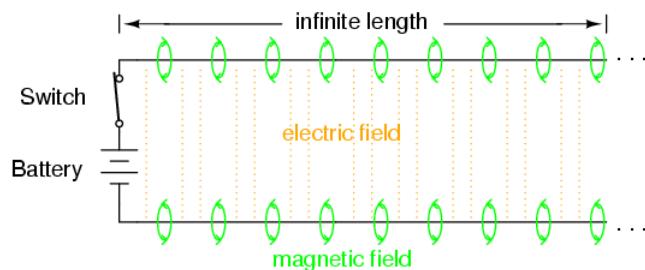
Voltage applied between two conductors creates an electric field between those conductors. Energy is stored in this electric field, and this storage of energy results in an opposition to change in voltage. The reaction of a capacitance against changes in voltage is described by the equation $i = C(de/dt)$, which tells us that current will be drawn proportional to the voltage’s rate of change over time. Thus, when the switch is closed, the capacitance between conductors will react against the sudden voltage increase by charging up and drawing current from the source. According to the equation, an instant rise in applied voltage (as produced by perfect switch closure) gives rise to an infinite charging current.

Capacitance and Inductance

However, the current drawn by a pair of parallel wires will not be infinite, because there exists series impedance along the wires due to inductance. (Figure below) Remember that current through *any* conductor develops a magnetic field of proportional magnitude. Energy is stored in this magnetic field, (Figure below) and this storage of energy results in an opposition to change in current. Each wire develops a magnetic field as it carries charging current for the capacitance between the wires, and in so doing drops voltage according to the inductance equation $e = L(di/dt)$. This voltage drop limits the voltage rate-of-change across the distributed capacitance, preventing the current from ever reaching an infinite magnitude:

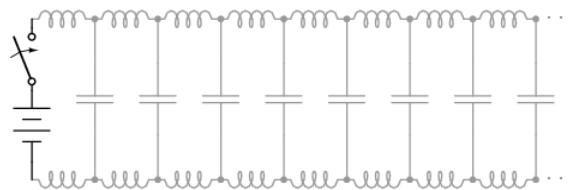


Equivalent circuit showing stray capacitance and inductance.



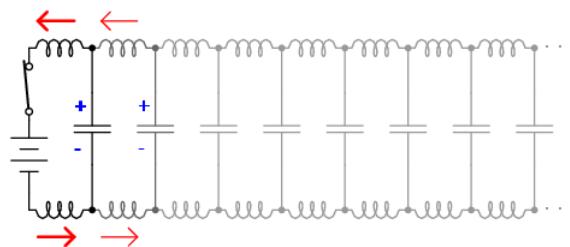
Voltage charges capacitance, current charges inductance.

Because the electrons in the two wires transfer motion to and from each other at nearly the speed of light, the “wave front” of voltage and current change will propagate down the length of the wires at that same velocity, resulting in the distributed capacitance and inductance progressively charging to full voltage and current, respectively, like this: (Figures below, below, below, below)

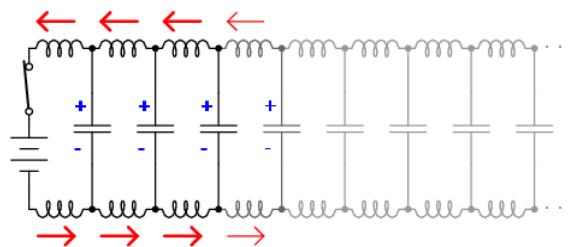


Uncharged transmission line.

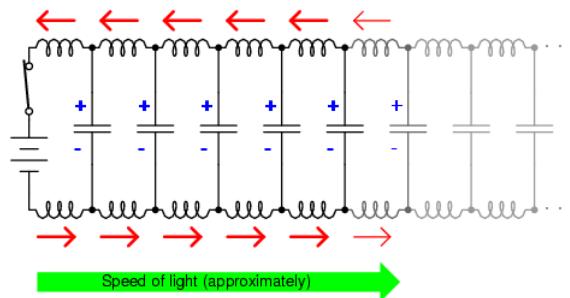
Switch closes!



Begin wave propagation.



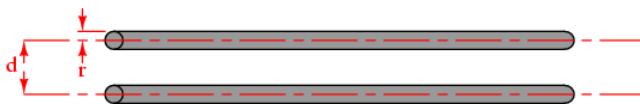
Continue wave propagation.



The Transmission Line

The end result of these interactions is a constant current of limited magnitude through the battery source. Since the wires are infinitely long, their distributed capacitance will never fully charge to the source voltage, and their distributed inductance will never allow unlimited charging current. In other words, this pair of wires will draw current from the source so long as the switch is closed, behaving as a constant load. No longer are the wires merely conductors of electrical current and carriers of voltage, but now constitute a circuit component in themselves, with unique characteristics. No longer are the two wires merely *a pair of conductors*, but rather a *transmission line*.

As a constant load, the transmission line's response to applied voltage is resistive rather than reactive, despite being comprised purely of inductance and capacitance (assuming superconducting wires with zero resistance). We can say this because there is no difference from the battery's perspective between a resistor eternally dissipating energy and an infinite transmission line eternally absorbing energy. The impedance (resistance) of this line in ohms is called the *characteristic impedance*, and it is fixed by the geometry of the two conductors. For a parallel-wire line with air insulation, the characteristic impedance may be calculated as such:



$$Z_0 = \frac{276}{\sqrt{k}} \log \frac{d}{r}$$

Where,

Z_0 = Characteristic impedance of line

d = Distance between conductor centers

r = Conductor radius

k = Relative permittivity of insulation between conductors

If the transmission line is coaxial in construction, the characteristic impedance follows a different equation:



$$Z_0 = \frac{138}{\sqrt{k}} \cdot \log \frac{d_1}{d_2}$$

Where,

Z_0 = Characteristic impedance of line

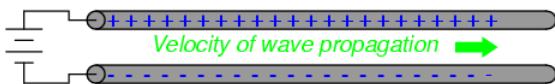
d_1 = Inside diameter of outer conductor

d_2 = Outside diameter of inner conductor

k = Relative permittivity of insulation between conductors

In both equations, identical units of measurement must be used in both terms of the fraction. If the insulating material is other than air (or a vacuum), both the characteristic impedance and the propagation velocity will be affected. The ratio of a transmission line's true propagation velocity and the speed of light in a vacuum is called the *velocity factor* of that line.

Velocity factor is purely a factor of the insulating material's relative permittivity (otherwise known as its *dielectric constant*), defined as the ratio of a material's electric field permittivity to that of a pure vacuum. The velocity factor of any cable type—coaxial or otherwise—may be calculated quite simply by the following formula:



$$\text{Velocity factor} = \frac{v}{c} = \frac{1}{\sqrt{k}}$$

Where,

v = Velocity of wave propagation

c = Velocity of light in a vacuum

k = Relative permittivity of insulation between conductors

The Natural Impedance

Characteristic impedance is also known as *natural impedance*, and it refers to the equivalent resistance of a transmission line if it were infinitely long, owing to distributed capacitance and inductance as the voltage and current “waves” propagate along its length at a propagation velocity equal to some large fraction of light speed.

It can be seen in either of the first two equations that a transmission line’s characteristic impedance (Z_0) increases as the conductor spacing increases. If the conductors are moved away from each other, the distributed capacitance will decrease (greater spacing between capacitor “plates”), and the distributed inductance will increase (less cancellation of the two opposing magnetic fields). Less parallel capacitance and more series inductance results in a smaller current drawn by the line for any given amount of applied voltage, which by definition is a greater impedance. Conversely, bringing the two conductors closer together increases the parallel capacitance and decreases the series inductance. Both changes result in a larger current drawn for a given applied voltage, equating to a lesser impedance.

Barring any dissipative effects such as dielectric “leakage” and conductor resistance, the characteristic impedance of a transmission line is equal to the square root of the ratio of the line’s inductance per unit length divided by the line’s capacitance per unit length:

$$Z_0 = \sqrt{\frac{L}{C}}$$

Where,

Z_0 = Characteristic impedance of line

L = Inductance per unit length of line

C = Capacitance per unit length of line

Review

- A *transmission line* is a pair of parallel conductors exhibiting certain characteristics due to distributed capacitance and inductance along its length.
- When a voltage is suddenly applied to one end of a transmission line, both a voltage “wave” and a current “wave” propagate along the line at nearly light speed.
- If a DC voltage is applied to one end of an infinitely long transmission line, the line will draw current from the DC source as though it were a constant resistance.
- The *characteristic impedance* (Z_0) of a transmission line is the resistance it would exhibit if it were infinite in length. This is entirely different from leakage resistance of the dielectric separating the two conductors, and the metallic resistance of the wires themselves. Characteristic impedance is purely a function of the capacitance and inductance distributed along the line’s length, and would exist even if the dielectric were perfect (infinite parallel resistance) and the wires superconducting (zero series resistance).
- *Velocity factor* is a fractional value relating a transmission line’s propagation speed to the speed of light in a vacuum. Values range between 0.66 and 0.80 for typical two-wire lines and coaxial cables. For any cable type, it is equal to the reciprocal (1/x) of the square root of the relative permittivity of the cable’s insulation.

12.3: Transmission Lines and Characteristic Impedance is shared under a [GNU Free Documentation License 1.3](#) license and was authored, remixed, and/or curated by Tony R Kuphaldt.

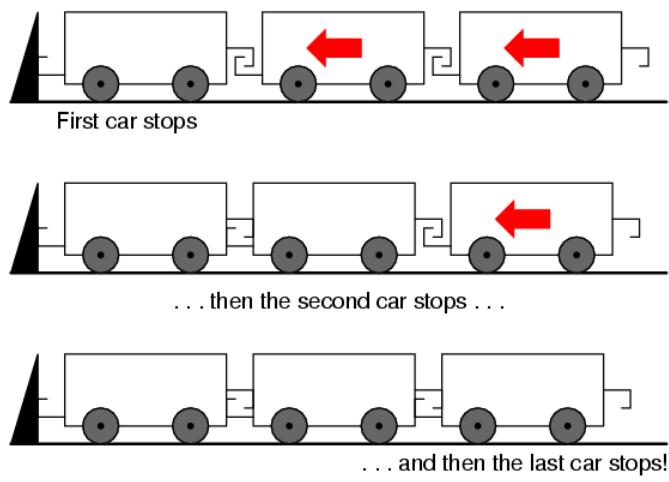
- **14.3: Characteristic Impedance** by Tony R. Kuphaldt is licensed [notset](#). Original source: <https://www.allaboutcircuits.com/textbook/alternating-current>.

12.4: Finite-length Transmission Lines

A transmission line of infinite length is an interesting abstraction, but physically impossible. All transmission lines have some finite length, and as such do not behave precisely the same as an infinite line. If that piece of $50\ \Omega$ “RG-58/U” cable I measured with an ohmmeter years ago had been infinitely long, I actually would have been able to measure $50\ \Omega$ worth of resistance between the inner and outer conductors. But it was not infinite in length, and so it measured as “open” (infinite resistance).

Nonetheless, the characteristic impedance rating of a transmission line is important even when dealing with limited lengths. An older term for characteristic impedance, which I like for its descriptive value, is *surge impedance*. If a transient voltage (a “surge”) is applied to the end of a transmission line, the line will draw a current proportional to the surge voltage magnitude divided by the line’s surge impedance ($I=E/Z$). This simple, Ohm’s Law relationship between current and voltage will hold true for a limited period of time, but not indefinitely.

If the end of a transmission line is open-circuited—that is, left unconnected—the current “wave” propagating down the line’s length will have to stop at the end, since electrons cannot flow where there is no continuing path. This abrupt cessation of current at the line’s end causes a “pile-up” to occur along the length of the transmission line, as the electrons successively find no place to go. Imagine a train traveling down the track with slack between the rail car couplings: if the lead car suddenly crashes into an immovable barricade, it will come to a stop, causing the one behind it to come to a stop as soon as the first coupling slack is taken up, which causes the next rail car to stop as soon as the next coupling’s slack is taken up, and so on until the last rail car stops. The train does not come to a halt together, but rather in sequence from first car to last: (Figure below)



Reflected wave.

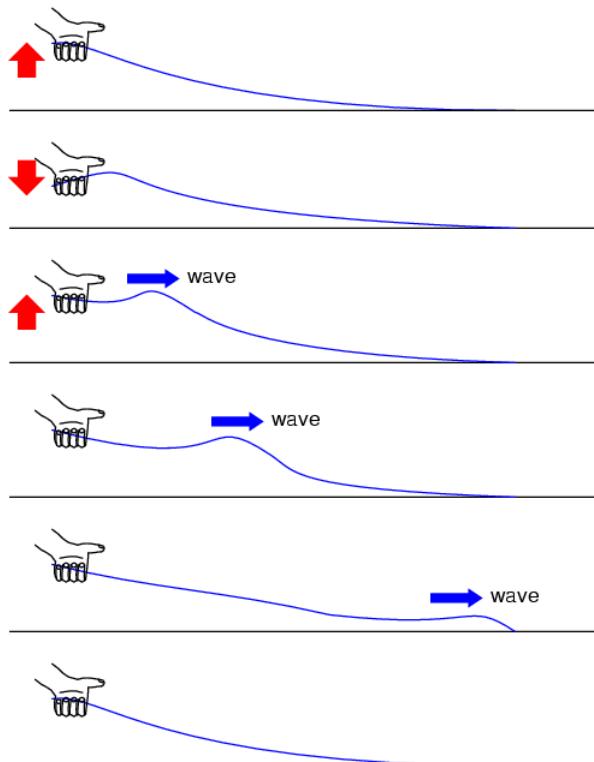
A signal propagating from the source-end of a transmission line to the load-end is called an *incident wave*. The propagation of a signal from load-end to source-end (such as what happened in this example with current encountering the end of an open-circuited transmission line) is called a *reflected wave*.

When this electron “pile-up” propagates back to the battery, current at the battery ceases, and the line acts as a simple open circuit. All this happens very quickly for transmission lines of reasonable length, and so an ohmmeter measurement of the line never reveals the brief time period where the line actually behaves as a resistor. For a mile-long cable with a velocity factor of 0.66 (signal propagation velocity is 66% of light speed, or 122,760 miles per second), it takes only $1/122,760$ of a second (8.146 microseconds) for a signal to travel from one end to the other. For the current signal to reach the line’s end and “reflect” back to the source, the round-trip time is twice this figure, or 16.292 μs .

High-speed measurement instruments are able to detect this transit time from source to line-end and back to source again, and may be used for the purpose of determining a cable’s length. This technique may also be used for determining the presence *and* location of a break in one or both of the cable’s conductors, since a current will “reflect” off the wire break just as it will off the end of an open-circuited cable. Instruments designed for such purposes are called *time-domain reflectometers* (TDRs). The basic principle is identical to that of sonar range-finding: generating a sound pulse and measuring the time it takes for the echo to return.

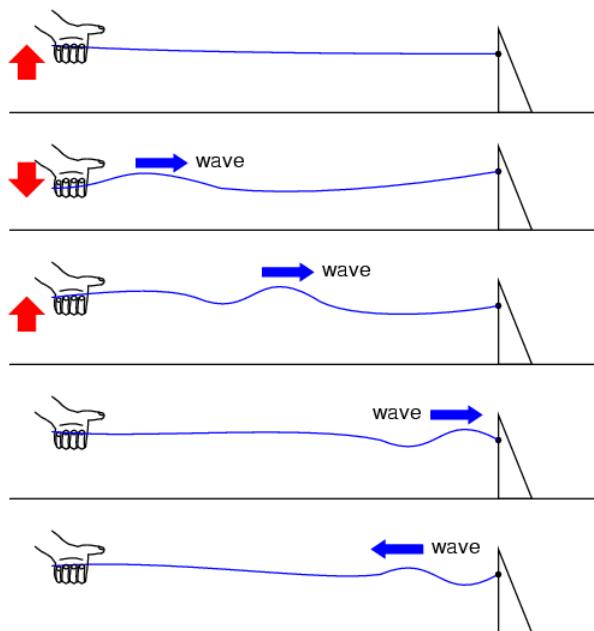
A similar phenomenon takes place if the end of a transmission line is short-circuited: when the voltage wave-front reaches the end of the line, it is reflected back to the source, because voltage cannot exist between two electrically common points. When this reflected wave reaches the source, the source sees the entire transmission line as a short-circuit. Again, this happens as quickly as the signal can propagate round-trip down and up the transmission line at whatever velocity allowed by the dielectric material between the line's conductors.

A simple experiment illustrates the phenomenon of wave reflection in transmission lines. Take a length of rope by one end and “whip” it with a rapid up-and-down motion of the wrist. A wave may be seen traveling down the rope’s length until it dissipates entirely due to friction: (Figure below)



Lossy transmission line.

This is analogous to a long transmission line with internal loss: the signal steadily grows weaker as it propagates down the line’s length, never reflecting back to the source. However, if the far end of the rope is secured to a solid object at a point prior to the incident wave’s total dissipation, a second wave will be reflected back to your hand: (Figure below)



Reflected wave.

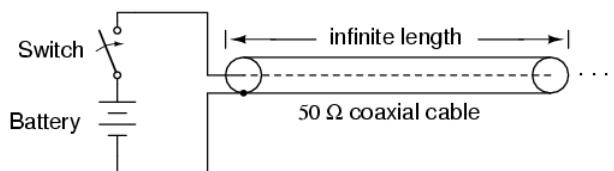
Usually, the purpose of a transmission line is to convey electrical energy from one point to another. Even if the signals are intended for information only, and not to power some significant load device, the ideal situation would be for all of the original signal energy to travel from the source to the load, and then be completely absorbed or dissipated by the load for maximum signal-to-noise ratio. Thus, “loss” along the length of a transmission line is undesirable, as are reflected waves, since reflected energy is energy not delivered to the end device.

Reflections may be eliminated from the transmission line if the load’s impedance exactly equals the characteristic (“surge”) impedance of the line. For example, a $50\ \Omega$ coaxial cable that is either open-circuited or short-circuited will reflect all of the incident energy back to the source. However, if a $50\ \Omega$ resistor is connected at the end of the cable, there will be no reflected energy, all signal energy being dissipated by the resistor.

This makes perfect sense if we return to our hypothetical, infinite-length transmission line example. A transmission line of $50\ \Omega$ characteristic impedance and infinite length behaves exactly like a $50\ \Omega$ resistance as measured from one end. (Figure below)

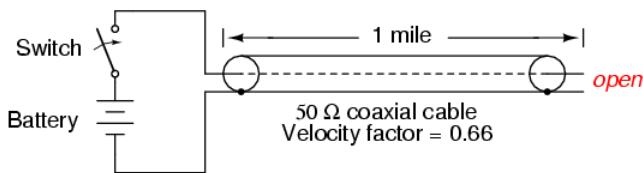
If we cut this line to some finite length, it will behave as a $50\ \Omega$ resistor to a constant source of DC voltage for a brief time, but then behave like an open- or a short-circuit, depending on what condition we leave the cut end of the line: open (Figure below) or shorted. (Figure below)

However, if we *terminate* the line with a $50\ \Omega$ resistor, the line will once again behave as a $50\ \Omega$ resistor, indefinitely: the same as if it were of infinite length again: (Figure below)



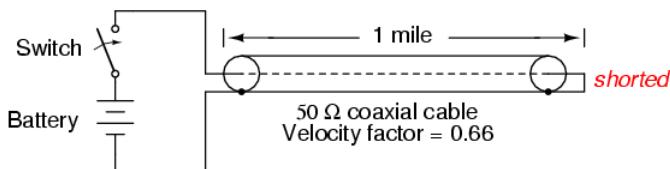
Cable's behavior from perspective of battery:
Exactly like a $50\ \Omega$ resistor

Infinite transmission line looks like resistor.


Cable's behavior from perspective of battery:

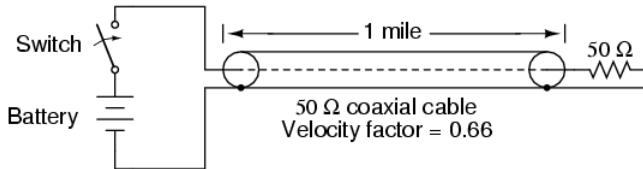
Like a 50Ω resistor for $16.292 \mu\text{s}$,
then like an open (infinite resistance)

One mile transmission.


Cable's behavior from perspective of battery:

Like a 50Ω resistor for $16.292 \mu\text{s}$,
then like a short (zero resistance)

Shorted transmission line.


Cable's behavior from perspective of battery:

Exactly like a 50Ω resistor

Line terminated in characteristic impedance.

In essence, a terminating resistor matching the natural impedance of the transmission line makes the line “appear” infinitely long from the perspective of the source, because a resistor has the ability to eternally dissipate energy in the same way a transmission line of infinite length is able to eternally absorb energy.

Reflected waves will also manifest if the terminating resistance isn’t precisely equal to the characteristic impedance of the transmission line, not just if the line is left unconnected (open) or jumpered (shorted). Though the energy reflection will not be total with a terminating impedance of slight mismatch, it will be partial. This happens whether or not the terminating resistance is *greater or less* than the line’s characteristic impedance.

Re-reflections of a reflected wave may also occur at the *source end* of a transmission line, if the source’s internal impedance (Thevenin equivalent impedance) is not exactly equal to the line’s characteristic impedance. A reflected wave returning back to the source will be dissipated entirely if the source impedance matches the line’s, but will be reflected back toward the line end like another incident wave, at least partially, if the source impedance does not match the line. This type of reflection may be particularly troublesome, as it makes it appear that the source has transmitted another pulse.

Review

- Characteristic impedance is also known as *surge impedance*, due to the temporarily resistive behavior of any length transmission line.
- A finite-length transmission line will appear to a DC voltage source as a constant resistance for some short time, then as whatever impedance the line is terminated with. Therefore, an open-ended cable simply reads “open” when measured with an ohmmeter, and “shorted” when its end is short-circuited.
- A transient (“surge”) signal applied to one end of an open-ended or short-circuited transmission line will “reflect” off the far end of the line as a secondary wave. A signal traveling on a transmission line from source to load is called an *incident wave*; a signal “bounced” off the end of a transmission line, traveling from load to source, is called a *reflected wave*.
- Reflected waves will also appear in transmission lines terminated by resistors not precisely matching the characteristic impedance.

- A finite-length transmission line may be made to appear infinite in length if terminated by a resistor of equal value to the line's characteristic impedance. This eliminates all signal reflections.
- A reflected wave may become re-reflected off the source-end of a transmission line if the source's internal impedance does not match the line's characteristic impedance. This re-reflected wave will appear, of course, like another pulse signal transmitted from the source.

12.4: Finite-length Transmission Lines is shared under a [GNU Free Documentation License 1.3](#) license and was authored, remixed, and/or curated by Tony R Kuphaldt.

- **14.4: Finite-length Transmission Lines** by [Tony R. Kuphaldt](#) is licensed [notset](#). Original source:
<https://www.allaboutcircuits.com/textbook/alternating-current>.

12.5: “Long” and “Short” Transmission Lines

In DC and low-frequency AC circuits, the characteristic impedance of parallel wires is usually ignored. This includes the use of coaxial cables in instrument circuits, often employed to protect weak voltage signals from being corrupted by induced “noise” caused by stray electric and magnetic fields. This is due to the relatively short timespans in which reflections take place in the line, as compared to the period of the waveforms or pulses of the significant signals in the circuit. As we saw in the last section, if a transmission line is connected to a DC voltage source, it will behave as a resistor equal in value to the line’s characteristic impedance only for as long as it takes the incident pulse to reach the end of the line and return as a reflected pulse, back to the source. After that time (a brief 16.292 μs for the mile-long coaxial cable of the last example), the source “sees” only the terminating impedance, whatever that may be.

If the circuit in question handles low-frequency AC power, such short time delays introduced by a transmission line between when the AC source outputs a voltage peak and when the source “sees” that peak loaded by the terminating impedance (round-trip time for the incident wave to reach the line’s end and reflect back to the source) are of little consequence. Even though we know that signal magnitudes along the line’s length are not equal at any given time due to signal propagation at (nearly) the speed of light, the actual phase difference between start-of-line and end-of-line signals is negligible, because line-length propagations occur within a very small fraction of the AC waveform’s period. For all practical purposes, we can say that voltage along all respective points on a low-frequency, two-conductor line are equal and in-phase with each other at any given point in time.

In these cases, we can say that the transmission lines in question are *electrically short*, because their propagation effects are much quicker than the periods of the conducted signals. By contrast, an *electrically long* line is one where the propagation time is a large fraction or even a multiple of the signal period. A “long” line is generally considered to be one where the source’s signal waveform completes at least a quarter-cycle (90° of “rotation”) before the incident signal reaches line’s end. Up until this chapter in the *Lessons In Electric Circuits* book series, all connecting lines were assumed to be electrically short.

To put this into perspective, we need to express the distance traveled by a voltage or current signal along a transmission line in relation to its source frequency. An AC waveform with a frequency of 60 Hz completes one cycle in 16.66 ms. At light speed (186,000 mile/s), this equates to a distance of 3100 miles that a voltage or current signal will propagate in that time. If the velocity factor of the transmission line is less than 1, the propagation velocity will be less than 186,000 miles per second, and the distance less by the same factor. But even if we used the coaxial cable’s velocity factor from the last example (0.66), the distance is still a very long 2046 miles! Whatever distance we calculate for a given frequency is called the *wavelength* of the signal.

A simple formula for calculating wavelength is as follows:

$$\lambda = \frac{v}{f}$$

Where,

λ = Wavelength

v = Velocity of propagation

f = Frequency of signal

The lower-case Greek letter “lambda” (λ) represents wavelength, in whatever unit of length used in the velocity figure (if miles per second, then wavelength in miles; if meters per second, then wavelength in meters). Velocity of propagation is usually the speed of light when calculating signal wavelength in open air or in a vacuum, but will be less if the transmission line has a velocity factor less than 1.

If a “long” line is considered to be one at least 1/4 wavelength in length, you can see why all connecting lines in the circuits discussed thusfar have been assumed “short.” For a 60 Hz AC power system, power lines would have to exceed 775 miles in length before the effects of propagation time became significant. Cables connecting an audio amplifier to speakers would have to be over 4.65 miles in length before line reflections would significantly impact a 10 kHz audio signal!

When dealing with radio-frequency systems, though, transmission line length is far from trivial. Consider a 100 MHz radio signal: its wavelength is a mere 9.8202 feet, even at the full propagation velocity of light (186,000 mile/s). A transmission line carrying this signal would not have to be more than about 2-1/2 feet in length to be considered “long!” With a cable velocity factor of 0.66, this critical length shrinks to 1.62 feet.

When an electrical source is connected to a load via a “short” transmission line, the load’s impedance dominates the circuit. This is to say, when the line is short, its own characteristic impedance is of little consequence to the circuit’s behavior. We see this when testing a coaxial cable with an ohmmeter: the cable reads “open” from center conductor to outer conductor if the cable end is left unterminated. Though the line acts as a resistor for a very brief period of time after the meter is connected (about $50\ \Omega$ for an RG-58/U cable), it immediately thereafter behaves as a simple “open circuit:” the impedance of the line’s open end. Since the combined response time of an ohmmeter and the human being using it *greatly exceeds* the round-trip propagation time up and down the cable, it is “electrically short” for this application, and we only register the terminating (load) impedance. It is the extreme speed of the propagated signal that makes us unable to detect the cable’s $50\ \Omega$ transient impedance with an ohmmeter.

If we use a coaxial cable to conduct a DC voltage or current to a load, and no component in the circuit is capable of measuring or responding quickly enough to “notice” a reflected wave, the cable is considered “electrically short” and its impedance is irrelevant to circuit function. Note how the electrical “shortness” of a cable is relative to the application: in a DC circuit where voltage and current values change slowly, nearly any physical length of cable would be considered “short” from the standpoint of characteristic impedance and reflected waves. Taking the same length of cable, though, and using it to conduct a high-frequency AC signal could result in a vastly different assessment of that cable’s “shortness!”

When a source is connected to a load via a “long” transmission line, the line’s own characteristic impedance dominates over load impedance in determining circuit behavior. In other words, an electrically “long” line acts as the principal component in the circuit, its own characteristics overshadowing the load’s. With a source connected to one end of the cable and a load to the other, current drawn from the source is a function primarily of the line and not the load. This is increasingly true the longer the transmission line is. Consider our hypothetical $50\ \Omega$ cable of infinite length, surely the ultimate example of a “long” transmission line: no matter what kind of load we connect to one end of this line, the source (connected to the other end) will only see $50\ \Omega$ of impedance, because the line’s infinite length prevents the signal from *ever reaching* the end where the load is connected. In this scenario, line impedance exclusively defines circuit behavior, rendering the load completely irrelevant.

The most effective way to minimize the impact of transmission line length on circuit behavior is to match the line’s characteristic impedance to the load impedance. If the load impedance is equal to the line impedance, then *any* signal source connected to the other end of the line will “see” the exact same impedance, and will have the exact same amount of current drawn from it, regardless of line length. In this condition of perfect impedance matching, line length only affects the amount of time delay from signal departure at the source to signal arrival at the load. However, perfect matching of line and load impedances is not always practical or possible.

The next section discusses the effects of “long” transmission lines, especially when line length happens to match specific fractions or multiples of signal wavelength.

Review

- Coaxial cabling is sometimes used in DC and low-frequency AC circuits as well as in high-frequency circuits, for the excellent immunity to induced “noise” that it provides for signals.
- When the period of a transmitted voltage or current signal greatly exceeds the propagation time for a transmission line, the line is considered *electrically short*. Conversely, when the propagation time is a large fraction or multiple of the signal’s period, the line is considered *electrically long*.
- A signal’s *wavelength* is the physical distance it will propagate in the timespan of one period. Wavelength is calculated by the formula $\lambda=v/f$, where “ λ ” is the wavelength, “ v ” is the propagation velocity, and “ f ” is the signal frequency.
- A rule-of-thumb for transmission line “shortness” is that the line must be at least $1/4$ wavelength before it is considered “long.”
- In a circuit with a “short” line, the terminating (load) impedance dominates circuit behavior. The source effectively sees nothing but the load’s impedance, barring any resistive losses in the transmission line.
- In a circuit with a “long” line, the line’s own characteristic impedance dominates circuit behavior. The ultimate example of this is a transmission line of infinite length: since the signal will *never* reach the load impedance, the source only “sees” the cable’s characteristic impedance.
- When a transmission line is terminated by a load precisely matching its impedance, there are no reflected waves and thus no problems with line length.

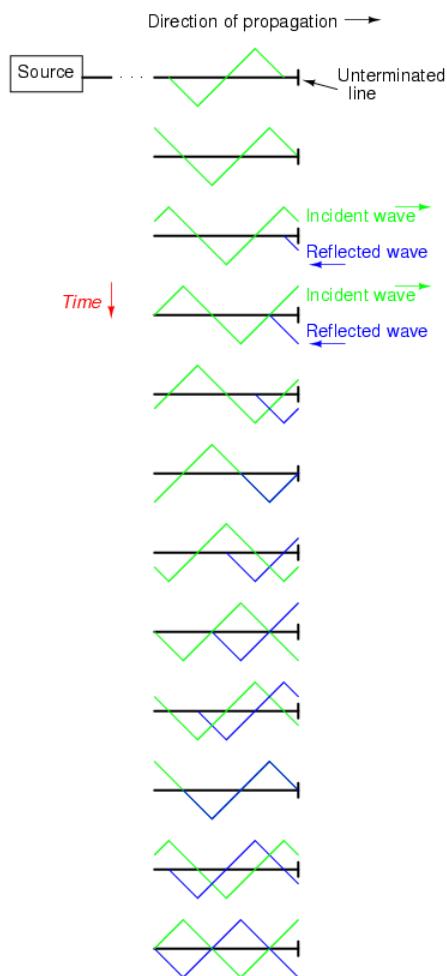
12.5: “Long” and “Short” Transmission Lines is shared under a [GNU Free Documentation License 1.3](#) license and was authored, remixed, and/or curated by LibreTexts.

- **14.5: “Long” and “Short” Transmission Lines** by Tony R. Kuphaldt is licensed [notset](#). Original source: <https://www.allaboutcircuits.com/textbook/alternating-current>.

12.6: Standing Waves and Resonance

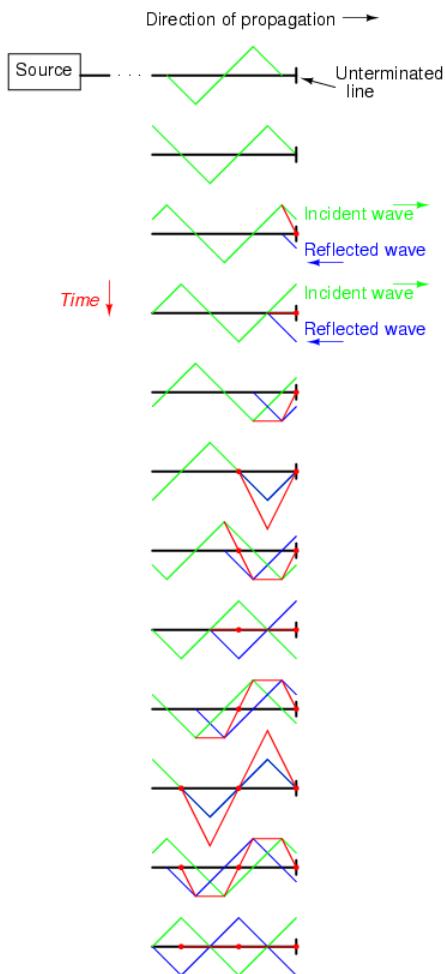
Whenever there is a mismatch of impedance between transmission line and load, reflections will occur. If the incident signal is a continuous AC waveform, these reflections will mix with more of the oncoming incident waveform to produce stationary waveforms called *standing waves*.

The following illustration shows how a triangle-shaped incident waveform turns into a mirror-image reflection upon reaching the line's unterminated end. The transmission line in this illustrative sequence is shown as a single, thick line rather than a pair of wires, for simplicity's sake. The incident wave is shown traveling from left to right, while the reflected wave travels from right to left: (Figure below)



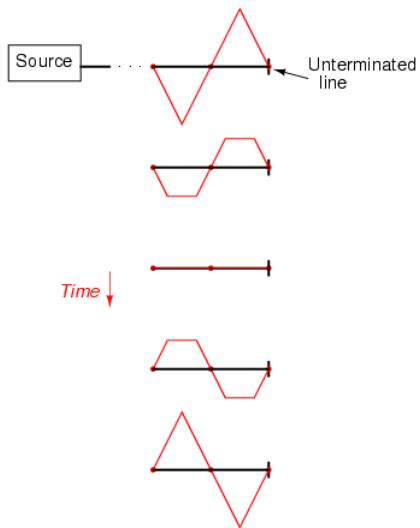
Incident wave reflects off end of unterminated transmission line.

If we add the two waveforms together, we find that a third, stationary waveform is created along the line's length: (Figure below)



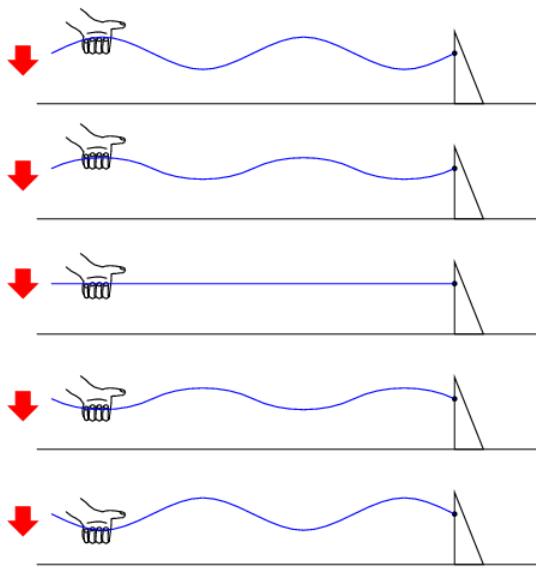
The sum of the incident and reflected waves is a stationary wave.

This third, “standing” wave, in fact, represents the only voltage along the line, being the representative sum of incident and reflected voltage waves. It oscillates in instantaneous magnitude, but does not propagate down the cable’s length like the incident or reflected waveforms causing it. Note the dots along the line length marking the “zero” points of the standing wave (where the incident and reflected waves cancel each other), and how those points never change position: (Figure below)



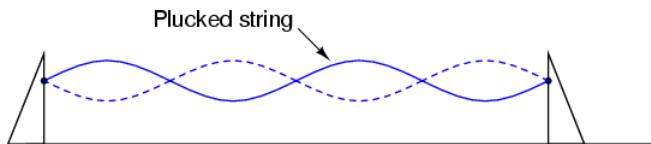
The standing wave does not propagate along the transmission line.

Standing waves are quite abundant in the physical world. Consider a string or rope, shaken at one end, and tied down at the other (only one half-cycle of hand motion shown, moving downward): (Figure below)



Standing waves on a rope.

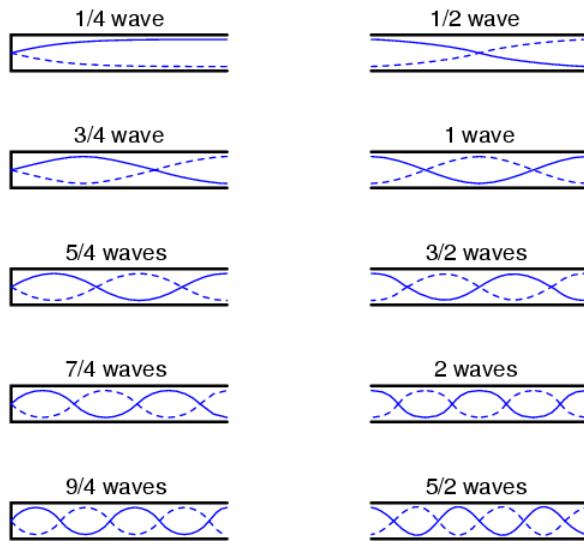
Both the nodes (points of little or no vibration) and the antinodes (points of maximum vibration) remain fixed along the length of the string or rope. The effect is most pronounced when the free end is shaken at just the right frequency. Plucked strings exhibit the same “standing wave” behavior, with “nodes” of maximum and minimum vibration along their length. The major difference between a plucked string and a shaken string is that the plucked string supplies its own “correct” frequency of vibration to maximize the standing-wave effect: (Figure below)



Standing waves on a plucked string.

Wind blowing across an open-ended tube also produces standing waves; this time, the waves are vibrations of air molecules (sound) within the tube rather than vibrations of a solid object. Whether the standing wave terminates in a node (minimum amplitude) or an antinode (maximum amplitude) depends on whether the other end of the tube is open or closed: (Figure below)

Standing sound waves in open-ended tubes

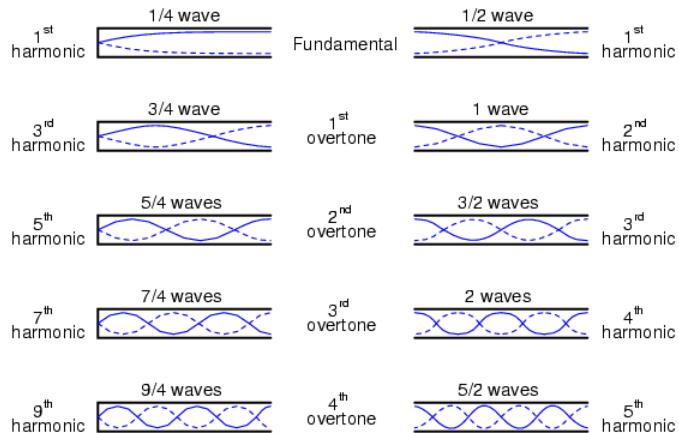


Standing sound waves in open ended tubes.

A closed tube end must be a wave node, while an open tube end must be an antinode. By analogy, the anchored end of a vibrating string must be a node, while the free end (if there is any) must be an antinode.

Note how there is more than one wavelength suitable for producing standing waves of vibrating air within a tube that precisely match the tube's end points. This is true for all standing-wave systems: standing waves will resonate with the system for any frequency (wavelength) correlating to the node/antinode points of the system. Another way of saying this is that there are multiple resonant frequencies for any system supporting standing waves.

All higher frequencies are integer-multiples of the lowest (fundamental) frequency for the system. The sequential progression of harmonics from one resonant frequency to the next defines the **overtone** frequencies for the system: (Figure below)

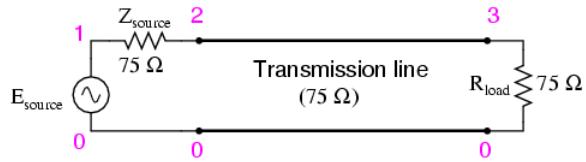


Harmonics (overtones) in open ended pipes

The actual frequencies (measured in Hertz) for any of these harmonics or overtones depends on the physical length of the tube and the waves' propagation velocity, which is the speed of sound in air.

Because transmission lines support standing waves, and force these waves to possess nodes and antinodes according to the type of termination impedance at the load end, they also exhibit resonance at frequencies determined by physical length and propagation velocity. Transmission line resonance, though, is a bit more complex than resonance of strings or of air in tubes, because we must consider both voltage waves and current waves.

This complexity is made easier to understand by way of computer simulation. To begin, let's examine a perfectly matched source, transmission line, and load. All components have an impedance of 75Ω : (Figure below)



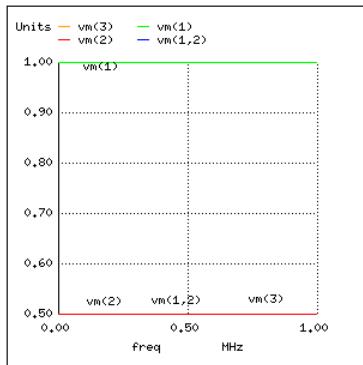
Perfectly matched transmission line.

Using SPICE to simulate the circuit, we'll specify the transmission line (`t1`) with a 75Ω characteristic impedance (`z0=75`) and a propagation delay of 1 microsecond (`td=1u`). This is a convenient method for expressing the physical length of a transmission line: the amount of time it takes a wave to propagate down its entire length. If this were a real 75Ω cable—perhaps a type “RG-59B/U” coaxial cable, the type commonly used for cable television distribution—with a velocity factor of 0.66, it would be about 648 feet long. Since 1 μs is the period of a 1 MHz signal, I'll choose to sweep the frequency of the AC source from (nearly) zero to that figure, to see how the system reacts when exposed to signals ranging from DC to 1 wavelength.

Here is the SPICE netlist for the circuit shown above:

```
Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 75
.ac lin 101 1m 1meg
* Using "Nutmeg" program to plot analysis
.end
```

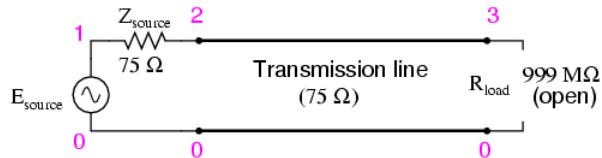
Running this simulation and plotting the source impedance drop (as an indication of current), the source voltage, the line's source-end voltage, and the load voltage, we see that the source voltage—shown as `vm(1)` (voltage magnitude between node 1 and the implied ground point of node 0) on the graphic plot—registers a steady 1 volt, while every other voltage registers a steady 0.5 volts: (Figure below)



No resonances on a matched transmission line.

In a system where all impedances are perfectly matched, there can be no standing waves, and therefore no resonant “peaks” or “valleys” in the Bode plot.

Now, let's change the load impedance to $999 \text{ M}\Omega$, to simulate an open-ended transmission line. (Figure below) We should definitely see some reflections on the line now as the frequency is swept from 1 mHz to 1 MHz: (Figure below)

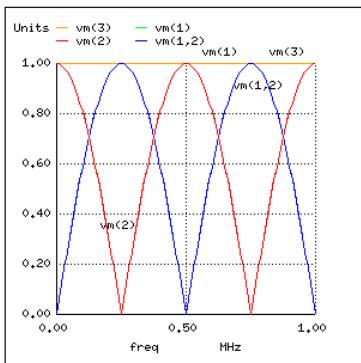


Open ended transmission line.

```

Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 999meg
.ac lin 101 1m 1meg
* Using "Nutmeg" program to plot analysis
.end

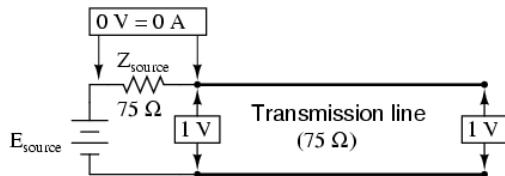
```



Resonances on open transmission line.

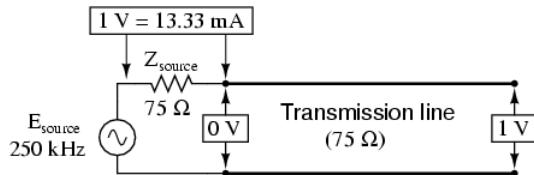
Here, both the supply voltage $vm(1)$ and the line's load-end voltage $vm(3)$ remain steady at 1 volt. The other voltages dip and peak at different frequencies along the sweep range of 1 mHz to 1 MHz. There are five points of interest along the horizontal axis of the analysis: 0 Hz, 250 kHz, 500 kHz, 750 kHz, and 1 MHz. We will investigate each one with regard to voltage and current at different points of the circuit.

At 0 Hz (actually 1 mHz), the signal is practically DC, and the circuit behaves much as it would given a 1-volt DC battery source. There is no circuit current, as indicated by zero voltage drop across the source impedance (Z_{source} : $vm(1,2)$), and full source voltage present at the source-end of the transmission line (voltage measured between node 2 and node 0: $vm(2)$). (Figure below)



At $f=0$: input: $V=1$, $I=0$; end: $V=1$, $I=0$.

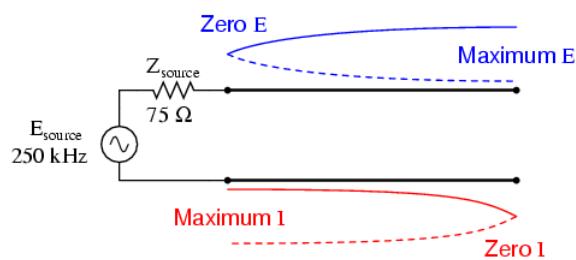
At 250 kHz, we see zero voltage and maximum current at the source-end of the transmission line, yet still full voltage at the load-end: (Figure below)



At $f=250$ KHz: input: $V=0$, $I=13.33$ mA; end: $V=1$ $I=0$.

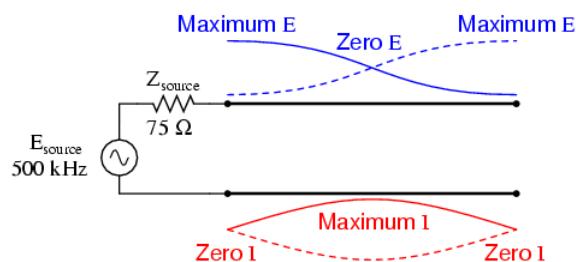
You might be wondering, how can this be? How can we get full source voltage at the line's open end while there is zero voltage at its entrance? The answer is found in the paradox of the standing wave. With a source frequency of 250 kHz, the line's length is precisely right for 1/4 wavelength to fit from end to end. With the line's load end open-circuited, there can be no current, but there

will be voltage. Therefore, the load-end of an open-circuited transmission line is a current node (zero point) and a voltage antinode (maximum amplitude): (Figure below)



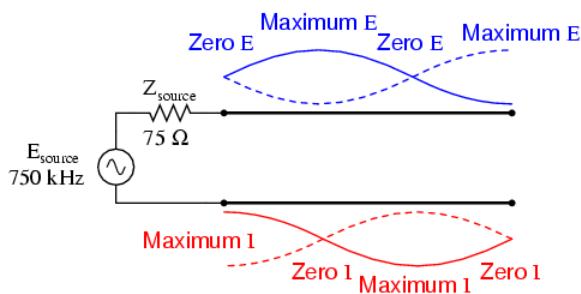
Open end of transmission line shows current node, voltage antinode at open end.

At 500 kHz, exactly one-half of a standing wave rests on the transmission line, and here we see another point in the analysis where the source current drops off to nothing and the source-end voltage of the transmission line rises again to full voltage: (Figure below)



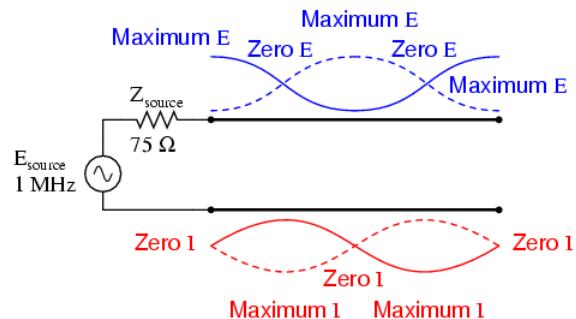
Full standing wave on half wave open transmission line.

At 750 kHz, the plot looks a lot like it was at 250 kHz: zero source-end voltage ($\text{vm}(2)$) and maximum current ($\text{vm}(1,2)$). This is due to 3/4 of a wave poised along the transmission line, resulting in the source “seeing” a short-circuit where it connects to the transmission line, even though the other end of the line is open-circuited: (Figure below)



1 1/2 standing waves on 3/4 wave open transmission line.

When the supply frequency sweeps up to 1 MHz, a full standing wave exists on the transmission line. At this point, the source-end of the line experiences the same voltage and current amplitudes as the load-end: full voltage and zero current. In essence, the source “sees” an open circuit at the point where it connects to the transmission line. (Figure below)



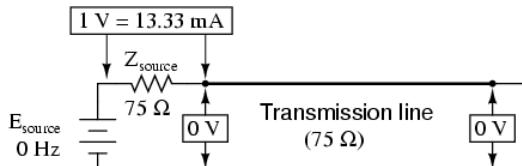
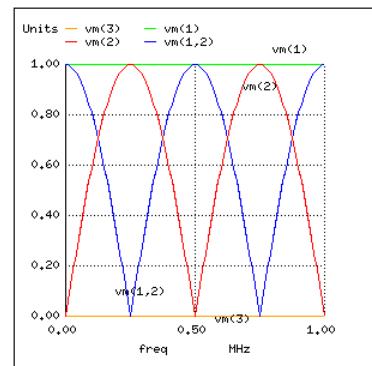
Double standing waves on full wave open transmission line.

In a similar fashion, a short-circuited transmission line generates standing waves, although the node and antinode assignments for voltage and current are reversed: at the shorted end of the line, there will be zero voltage (node) and maximum current (antinode). What follows is the SPICE simulation (circuit Figure below and illustrations of what happens (Figure 2nd-below at resonances) at all the interesting frequencies: 0 Hz (Figure below), 250 kHz (Figure below), 500 kHz (Figure below), 750 kHz (Figure below), and 1 MHz (Figure below). The short-circuit jumper is simulated by a $1 \mu\Omega$ load impedance: (Figure below)

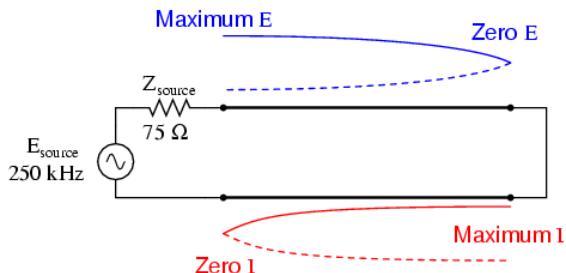


Shorted transmission line.

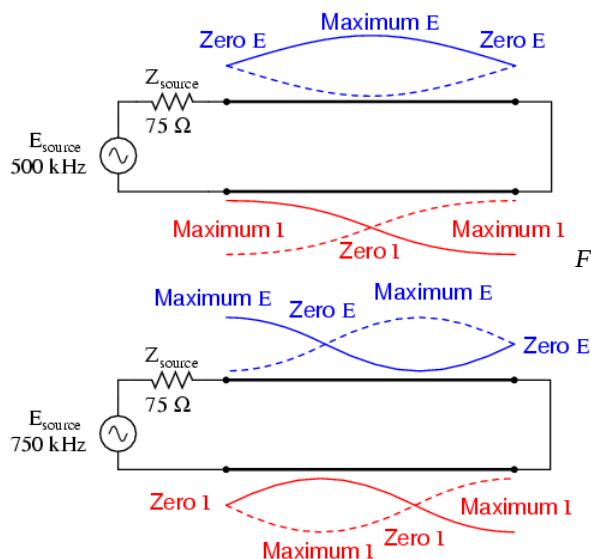
```
Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 1u
.ac lin 101 1m 1meg
* Using "Nutmeg" program to plot analysis
.end
```



Resonances on shorted transmission line
end: $V=0$, $I=13.33 \text{ mA}$.

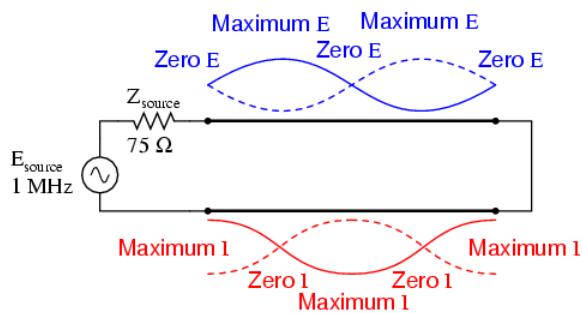


Half wave standing wave pattern on 1/4 wave shorted transmission line.



Full wave standing wave pattern on half wave shorted transmission line.

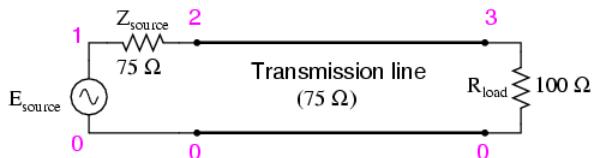
1 1/2 standing wavepattern on 3/4 wave shorted transmission line.



Double standing waves on full wave shorted transmission line.

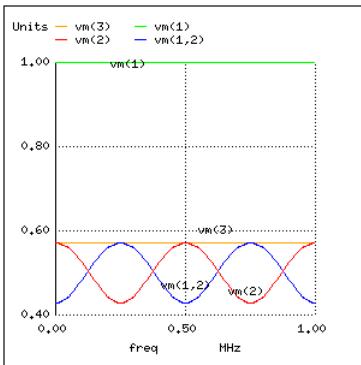
In both these circuit examples, an open-circuited line and a short-circuited line, the energy reflection is total: 100% of the incident wave reaching the line's end gets reflected back toward the source. If, however, the transmission line is terminated in some impedance other than an open or a short, the reflections will be less intense, as will be the difference between minimum and maximum values of voltage and current along the line.

Suppose we were to terminate our example line with a $100\ \Omega$ resistor instead of a $75\ \Omega$ resistor. (Figure below) Examine the results of the corresponding SPICE analysis to see the effects of impedance mismatch at different source frequencies: (Figure below)



Transmission line terminated in a mismatch

```
Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 100
.ac lin 101 1m 1meg
* Using 'Nutmeg' program to plot analysis
.end
```



Weak resonances on a mismatched transmission line

If we run another SPICE analysis, this time printing numerical results rather than plotting them, we can discover exactly what is happening at all the interesting frequencies: (DC, Figure below; 250 kHz, Figure below; 500 kHz, Figure below; 750 kHz, Figure below; and 1 MHz, Figure below).

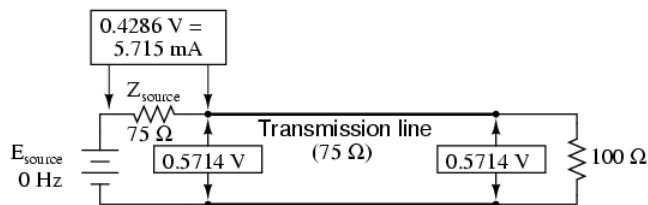
```

Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 100
.ac lin 5 1m 1meg
.print ac v(1,2) v(1) v(2) v(3)
.end

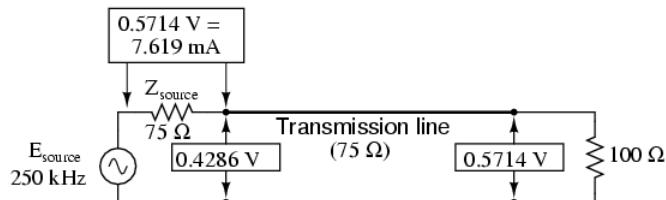
freq v(1,2) v(1) v(2) v(3)
1.000E-03 4.286E-01 1.000E+00 5.714E-01 5.714E-01
2.500E+05 5.714E-01 1.000E+00 4.286E-01 5.714E-01
5.000E+05 4.286E-01 1.000E+00 5.714E-01 5.714E-01
7.500E+05 5.714E-01 1.000E+00 4.286E-01 5.714E-01
1.000E+06 4.286E-01 1.000E+00 5.714E-01 5.714E-01

```

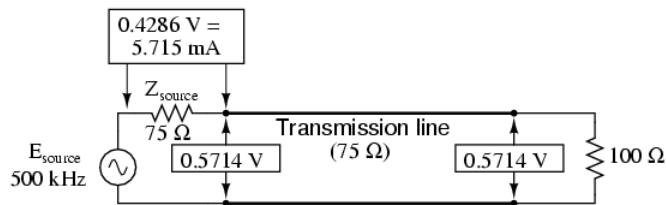
At all frequencies, the source voltage, $v(1)$, remains steady at 1 volt, as it should. The load voltage, $v(3)$, also remains steady, but at a lesser voltage: 0.5714 volts. However, both the line input voltage ($v(2)$) and the voltage dropped across the source's 75 Ω impedance ($v(1,2)$), indicating current drawn from the source) vary with frequency.



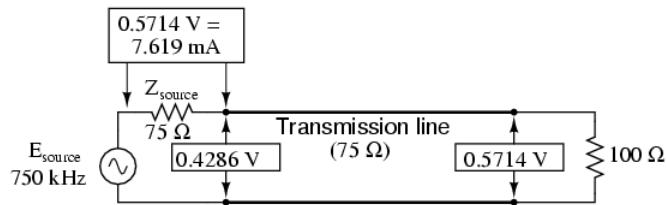
At $f=0$ Hz: input: $V=0.5714$, $I=5.715$ mA; end: $V=0.5714$, $I=5.715$ mA.



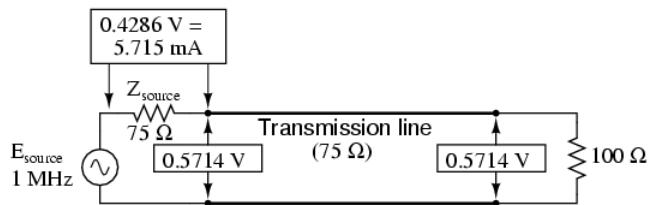
At $f=250$ KHz: input: $V=0.4286$, $I=7.619$ mA; end: $V=0.5714$, $I=7.619$ mA.



At $f=500 \text{ KHz}$: input: $V=0.5714$, $I=5.715 \text{ mA}$; end: $V=0.5714$, $I=5.715 \text{ mA}$.



At $f=750 \text{ KHz}$: input: $V=0.4286$, $I=7.619 \text{ mA}$; end: $V=0.5714$, $I=7.619 \text{ mA}$.



At $f=1 \text{ MHz}$: input: $V=0.5714$, $I=5.715 \text{ mA}$; end: $V=0.5714$, $I=5.715 \text{ mA}$.

At odd harmonics of the fundamental frequency (250 kHz, Figure 3rd-above and 750 kHz, Figure above) we see differing levels of voltage at each end of the transmission line, because at those frequencies the standing waves terminate at one end in a node and at the other end in an antinode. Unlike the open-circuited and short-circuited transmission line examples, the maximum and minimum voltage levels along this transmission line do not reach the same extreme values of 0% and 100% source voltage, but we still have points of “minimum” and “maximum” voltage. (Figure 6th-above) The same holds true for current: if the line’s terminating impedance is mismatched to the line’s characteristic impedance, we will have points of minimum and maximum current at certain fixed locations on the line, corresponding to the standing current wave’s nodes and antinodes, respectively.

One way of expressing the severity of standing waves is as a ratio of maximum amplitude (antinode) to minimum amplitude (node), for voltage or for current. When a line is terminated by an open or a short, this *standing wave ratio*, or *SWR* is valued at infinity, since the minimum amplitude will be zero, and any finite value divided by zero results in an infinite (actually, “undefined”) quotient. In this example, with a 75Ω line terminated by a 100Ω impedance, the SWR will be finite: 1.333, calculated by taking the maximum line voltage at either 250 kHz or 750 kHz (0.5714 volts) and dividing by the minimum line voltage (0.4286 volts).

Standing wave ratio may also be calculated by taking the line’s terminating impedance and the line’s characteristic impedance, and dividing the larger of the two values by the smaller. In this example, the terminating impedance of 100Ω divided by the characteristic impedance of 75Ω yields a quotient of exactly 1.333, matching the previous calculation very closely.

$$\text{SWR} = \frac{E_{\text{maximum}}}{E_{\text{minimum}}} = \frac{I_{\text{maximum}}}{I_{\text{minimum}}}$$

$$\text{SWR} = \frac{Z_{\text{load}}}{Z_0} \quad \text{or} \quad \frac{Z_0}{Z_{\text{load}}}$$

which ever is greater

A perfectly terminated transmission line will have an SWR of 1, since voltage at any location along the line's length will be the same, and likewise for current. Again, this is usually considered ideal, not only because reflected waves constitute energy not delivered to the load, but because the high values of voltage and current created by the antinodes of standing waves may over-stress the transmission line's insulation (high voltage) and conductors (high current), respectively.

Also, a transmission line with a high SWR tends to act as an antenna, radiating electromagnetic energy away from the line, rather than channeling all of it to the load. This is usually undesirable, as the radiated energy may "couple" with nearby conductors, producing signal interference. An interesting footnote to this point is that antenna structures—which typically resemble open- or short-circuited transmission lines—are often designed to operate at *high* standing wave ratios, for the very reason of maximizing signal radiation and reception.

The following photograph (Figure below) shows a set of transmission lines at a junction point in a radio transmitter system. The large, copper tubes with ceramic insulator caps at the ends are rigid coaxial transmission lines of 50Ω characteristic impedance. These lines carry RF power from the radio transmitter circuit to a small, wooden shelter at the base of an antenna structure, and from that shelter on to other shelters with other antenna structures:



Flexible coaxial cables connected to rigid lines.

Flexible coaxial cable connected to the rigid lines (also of 50Ω characteristic impedance) conduct the RF power to capacitive and inductive "phasing" networks inside the shelter. The white, plastic tube joining two of the rigid lines together carries "filling" gas from one sealed line to the other. The lines are gas-filled to avoid collecting moisture inside them, which would be a definite problem for a coaxial line. Note the flat, copper "straps" used as jumper wires to connect the conductors of the flexible coaxial cables to the conductors of the rigid lines. Why flat straps of copper and not round wires? Because of the skin effect, which renders most of the cross-sectional area of a round conductor useless at radio frequencies.

Like many transmission lines, these are operated at low SWR conditions. As we will see in the next section, though, the phenomenon of standing waves in transmission lines is not always undesirable, as it may be exploited to perform a useful function: impedance transformation.

Review

- *Standing waves* are waves of voltage and current which do not propagate (i.e. they are stationary), but are the result of interference between incident and reflected waves along a transmission line.
- A *node* is a point on a standing wave of *minimum* amplitude.
- An *antinode* is a point on a standing wave of *maximum* amplitude.
- Standing waves can only exist in a transmission line when the terminating impedance does not match the line's characteristic impedance. In a perfectly terminated line, there are no reflected waves, and therefore no standing waves at all.
- At certain frequencies, the nodes and antinodes of standing waves will correlate with the ends of a transmission line, resulting in *resonance*.
- The lowest-frequency resonant point on a transmission line is where the line is one quarter-wavelength long. Resonant points exist at every harmonic (integer-multiple) frequency of the fundamental (quarter-wavelength).

- *Standing wave ratio*, or *SWR*, is the ratio of maximum standing wave amplitude to minimum standing wave amplitude. It may also be calculated by dividing termination impedance by characteristic impedance, or vice versa, whichever yields the greatest quotient. A line with no standing waves (perfectly matched: Z_{load} to Z_0) has an SWR equal to 1.
- Transmission lines may be damaged by the high maximum amplitudes of standing waves. Voltage antinodes may break down insulation between conductors, and current antinodes may overheat conductors.

12.6: Standing Waves and Resonance is shared under a [GNU Free Documentation License 1.3](#) license and was authored, remixed, and/or curated by Tony R Kuphaldt.

- **14.6: Standing Waves and Resonance** by [Tony R. Kuphaldt](#) is licensed [notset](#). Original source:
<https://www.allaboutcircuits.com/textbook/alternating-current>.

12.7: Antenna Systems (Summary)

Key Terms

transverse wave	a wave, such as an electromagnetic wave, which oscillates perpendicular to the axis along the line of travel
standing wave	a wave that oscillates in place, with nodes where no motion happens
wavelength	the distance from one peak to the next in a wave
amplitude	the height, or magnitude, of an electromagnetic wave
frequency	the number of complete wave cycles (up-down-up) passing a given point within one second (cycles/second)
resonant system	a system that displays enhanced oscillation when subjected to a periodic disturbance of the same frequency as its natural frequency
oscillate	to fluctuate back and forth in a steady beat

Summary

Production of Electromagnetic Waves

- Electromagnetic waves are created by oscillating charges (which radiate whenever accelerated) and have the same frequency as the oscillation.
- Since the electric and magnetic fields in most electromagnetic waves are perpendicular to the direction in which the wave moves, it is ordinarily a transverse wave.
- The strengths of the electric and magnetic parts of the wave are related by

$$\frac{\mathbf{E}}{\mathbf{B}} = c,$$

which implies that the magnetic field \mathbf{B} is very weak relative to the electric field \mathbf{E} .

12.7: Antenna Systems (Summary) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax.

- [24.2: Production of Electromagnetic Waves](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/college-physics>.

CHAPTER OVERVIEW

13: Propagation of Electromagnetic Waves

- 13.1: Introduction
- 13.2: Ray and Wave Models of Propagation
- 13.3: Reflection of Rays
- 13.4: Refraction of Rays
- 13.5: Application- Line-of-Sight Transmission
- 13.6: Diffraction of Waves
- 13.7: Interference of Waves
- 13.8: Double-Slit Interference
- 13.9: Propagation of Electromagnetic Waves (Summary)
- 13.10: Propagation of Electromagnetic Waves (Exercises)
- 13.11: Propagation of Electromagnetic Waves (Answers)

13: Propagation of Electromagnetic Waves is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

13.1: Introduction

What happens to electromagnetic waves as they propagate through space and start to interact with their environment? This chapter will describe some of the fundamental physical phenomena that can occur, including reflection, refraction, diffraction, and interference. As many of these phenomena are easiest to observe optically, many of the examples are for visible light, but all of these effects also occur for radio waves.

13.1: Introduction is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Ronald Kumon.

13.2: Ray and Wave Models of Propagation

Learning Objectives

By the end of this section, you will be able to:

- Describe the condition which determines if electromagnetic wave propagation can be accurately described by the ray model.
- List the ways in which light travels from a source to another location in the ray model.
- List phenomena that the wave model of propagation describes better than the ray model.

The Ray Model

In the chapter on [Electromagnetic Waves](#), you have already seen that electromagnetic energy can propagate in the form of waves. However, experiments show that when the electromagnetic wave interacts with an object that is several times as large as the wave's wavelength, it travels in straight lines and acts like a **ray**. The word "ray" comes from mathematics, and here means a straight line that originates at some point. The **ray model** describes the propagation path of the electromagnetic energy as straight lines.

If the electromagnetic wave is light, it is acceptable to think of light rays like the thin beams coming out of a laser. Its wave characteristics are not pronounced in such situations. Since the wavelength of visible light is less than a micron (a thousandth of a millimeter), it acts like a ray in the many common situations in which it encounters objects larger than a micron. For example, when visible light encounters anything large enough to observe with unaided eyes, such as a coin, it acts like a ray, with generally negligible wave characteristics. As we have seen, radio waves can have wavelengths of from tenths of meters to hundreds of meters, so the ray model is perhaps even more widely applicable in many circumstances for radio waves.

In this chapter, we start mainly with the ray characteristics in the context of light, as this approach is easiest to visualize. There are three ways in which light can travel from a source to another location (Figure 13.2.1): (1) It can come directly from the source through empty space, such as from the Sun to Earth. (2) It can also travel through various media, such as air and glass, to the observer. (3) Light can also arrive after being reflected, such as by a mirror. In all of these cases, we can accurately model the path of light as straight lines. Light may change direction when it encounters objects (such as a mirror) or in passing from one material to another (such as in passing from air to glass), but it then continues in a straight line or as a ray.

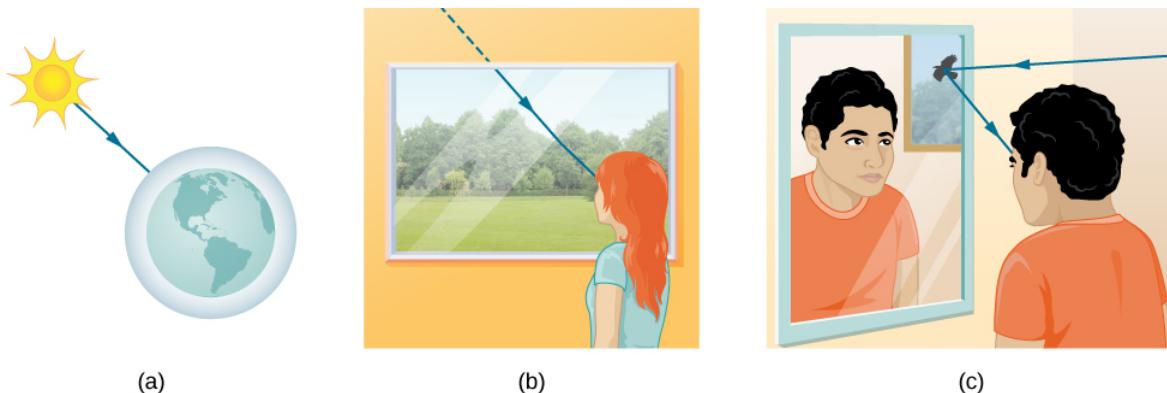


Figure 13.2.1: Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth, traveling through empty space directly from the source. (b) Light can reach a person by traveling through media like air and glass. (c) Light can also reflect from an object like a mirror. In the situations shown here, light interacts with objects large enough that it travels in straight lines, like a ray.

Since light moves in straight lines, changing directions when it interacts with materials, its path is described by geometry and simple trigonometry. This part of optics, where the ray aspect of light dominates, is, therefore, called **geometric optics**. Two laws govern how light changes direction when it interacts with matter. These are the **law of reflection**, for situations in which light bounces off matter, and the **law of refraction**, for situations in which light passes through matter. We will examine more about each of these laws in upcoming sections of this chapter.

The Wave Model

When an electromagnetic wave interacts with objects that are comparable or smaller than its wavelength, then the ray model is no longer appropriate. In the context of light, when the wave aspect dominates, this part of optics is called **wave optics**. When the waves interact with objects or other waves, they can exhibit very prominent wave characteristics such as **diffraction** and **interference**. These phenomena will also be discussed in subsequent sections.

13.2: Ray and Wave Models of Propagation is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Ronald Kumon & OpenStax.

- 1.2: The Propagation of Light by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-3>.

13.3: Reflection of Rays

Learning Objectives

By the end of this section, you will be able to:

- Explain the reflection of light from polished and rough surfaces
- Describe the principle and applications of corner reflectors

Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at a piece of white paper, you are seeing light scattered from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

The **law of reflection** states that the angle of reflection equals the angle of incidence:

$$\theta_r = \theta_i \quad (13.3.1)$$

The law of reflection is illustrated in Figure 13.3.1, which also shows how the angle of incidence and angle of reflection are measured relative to the perpendicular to the surface at the point where the light ray strikes.

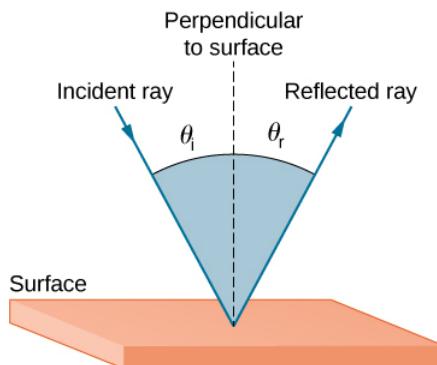


Figure 13.3.1: The law of reflection states that the angle of reflection equals the angle of incidence— $\theta_r = \theta_i$. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.

We expect to see reflections from smooth surfaces, but Figure 13.3.2 illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as shown in Figure 13.3.1a.

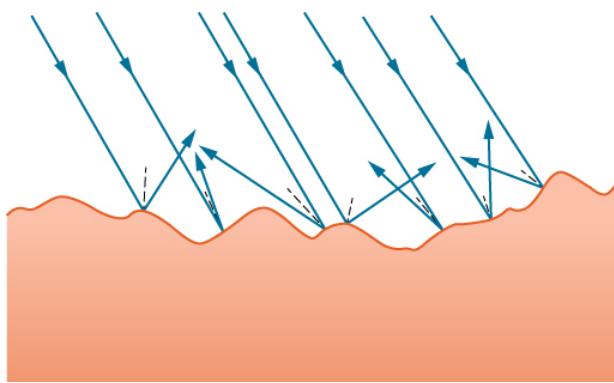


Figure 13.3.2: Light is diffused when it reflects from a rough surface. Here, many parallel rays are incident, but they are reflected at many different angles, because the surface is rough.

People, clothing, leaves, and walls all have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in Figure 13.3.3b. When the Moon reflects from a lake, as shown in Figure 13.3.1c, a combination of these effects takes place.

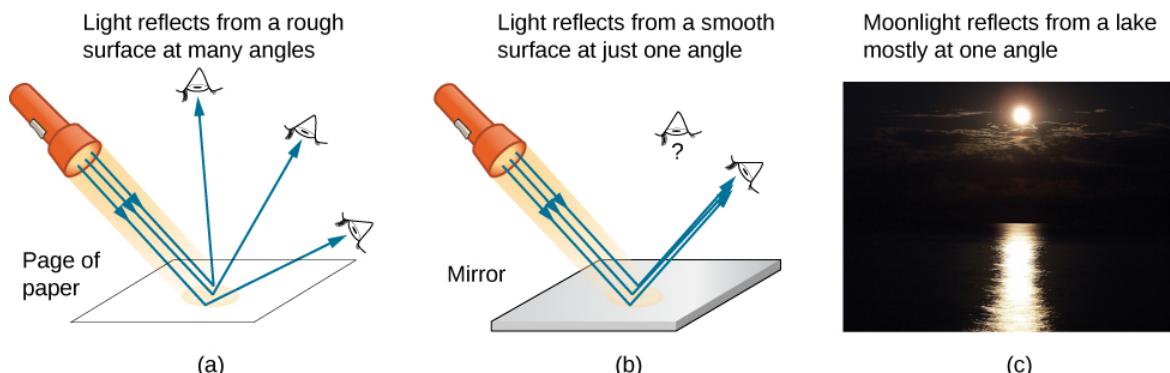


Figure 13.3.3: (a) When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light. (b) A mirror illuminated by many parallel rays reflects them in only one direction, because its surface is very smooth. Only the observer at a particular angle sees the reflected light. (c) Moonlight is spread out when it is reflected by the lake, because the surface is shiny but uneven. (credit c: modification of work by Diego Torres Silvestre)

When you see yourself in a mirror, it appears that the image is actually behind the mirror (Figure 13.3.4). We see the light coming from a direction determined by the law of reflection. The angles are such that the image is exactly the same distance behind the mirror as you stand in front of the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of your imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (which are optical instruments themselves). The precise manner in which images are formed by mirrors and lenses is discussed in an upcoming chapter on [Geometric Optics and Image Formation](#).

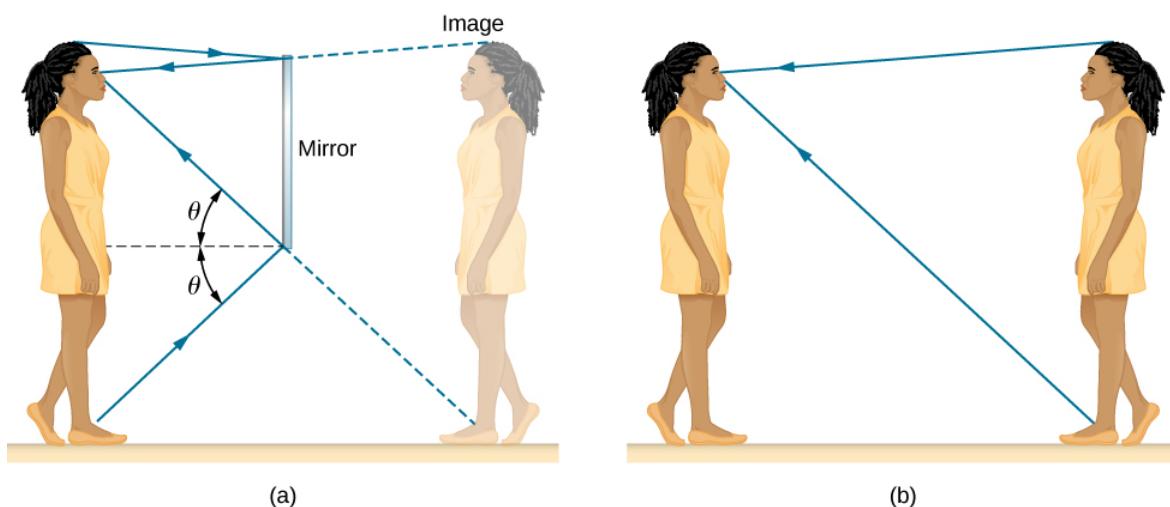


Figure 13.3.4: (a) Your image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be behind the mirror at the same distance away as (b) if you were looking at your twin directly, with no mirror.

Corner Reflectors (Retroreflectors)

A light ray that strikes an object consisting of two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came (Figure 13.3.5). This is true whenever the reflecting surfaces are perpendicular, and it is independent of the angle of incidence. Such an object is called a **corner reflector**, since the light bounces from its inside corner. Corner reflectors are a subclass of retroreflectors, which all reflect rays back in the directions from which they came. Although the geometry of the proof is much more complex, corner reflectors can also be built with three mutually perpendicular reflecting surfaces and are useful in three-dimensional applications.

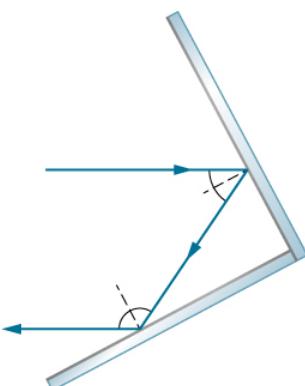
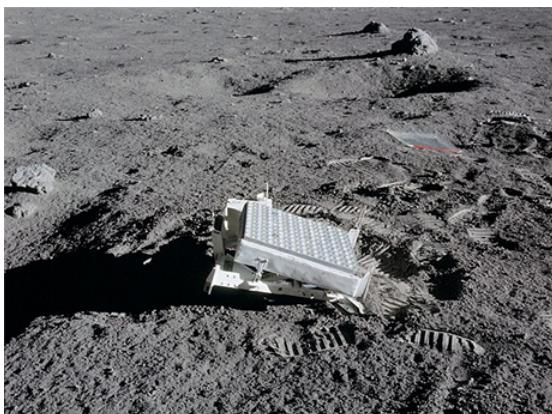


Figure 13.3.5: A light ray that strikes two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came.

Many inexpensive reflector buttons on bicycles, cars, and warning signs have corner reflectors designed to return light in the direction from which it originated. Rather than simply reflecting light over a wide angle, retroreflection ensures high visibility if the observer and the light source are located together, such as a car's driver and headlights. The Apollo astronauts placed a true corner reflector on the Moon (Figure 13.3.6). Laser signals from Earth can be bounced from that corner reflector to measure the gradually increasing distance to the Moon of a few centimeters per year.



(a)



(b)

Figure 13.3.6: (a) Astronauts placed a corner reflector on the Moon to measure its gradually increasing orbital distance. (b) The bright spots on these bicycle safety reflectors are reflections of the flash of the camera that took this picture on a dark night. (credit a: modification of work by NASA; credit b: modification of work by "Julo"/Wikimedia Commons)

Working on the same principle as these optical reflectors, corner reflectors are routinely used as radar reflectors (Figure 13.3.7) for radio-frequency applications. Under most circumstances, small boats made of fiberglass or wood do not strongly reflect radio waves emitted by radar systems. To make these boats visible to radar (to avoid collisions, for example), radar reflectors are attached to boats, usually in high places.



Figure 13.3.7: A radar reflector hoisted on a sailboat is a type of corner reflector. (credit: Tim Sheerman-Chase)

As a counterexample, if you are interested in building a stealth airplane, radar reflections should be minimized to evade detection. One of the design considerations would then be to avoid building 90°90° corners into the airframe.

This page titled [13.3: Reflection of Rays](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [1.3: The Law of Reflection](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-3>.

13.4: Refraction of Rays

Learning Objectives

By the end of this section, you will be able to:

- Determine the index of refraction, given the speed of light in a medium
- Describe how rays change direction upon entering a medium
- Apply the law of refraction in problem solving

You may often notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places (Figure 13.4.1). This happens because light coming from the fish to you changes direction when it leaves the tank, and in this case, it can travel two different paths to get to your eyes. The changing of a light ray's direction (loosely called bending) when it passes through substances of different refractive indices is called **refraction** and is related to changes in the speed of light, $v = c/n$. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to data transmission through optical fibers.

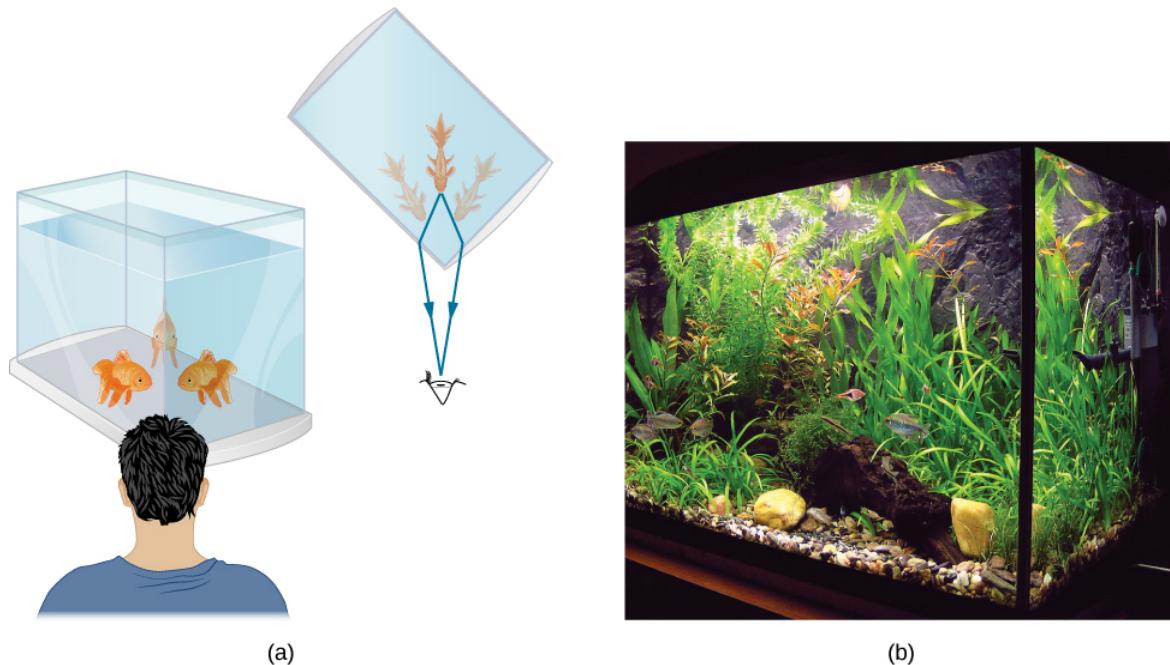


Figure 13.4.1: (a) Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena. (b) This image shows refraction of light from a fish near the top of a fish tank.

The Speed of Light in Matter

Today, the speed of light is known to great precision. In fact, the speed of light in a vacuum c is so important that it is accepted as one of the basic physical quantities and has the value

$$c = 2.99792458 \times 10^8 \text{ m/s} \equiv 3.00 \times 10^8 \text{ m/s} \quad (13.4.1)$$

where the approximate value of $3.00 \times 10^8 \text{ m/s}$ is used whenever three-digit accuracy is sufficient.

The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction varies with different atoms, crystal lattices, and other substructures. We can define a constant of a material that describes the speed of light in it, called the index of refraction n :

$$n = \frac{c}{v} \quad (13.4.2)$$

where v is the observed speed of light in the material.

Since the speed of light is always less than c in matter and equals c only in a vacuum, the index of refraction is always greater than or equal to one; that is, $n \geq 1$. Table 13.4.1 gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors separated by a prism, as we will see in [Dispersion](#).) Note that for gases, n is close to 1.0. This seems reasonable, since atoms in gases are widely separated, and light travels at c in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light v in a medium varies considerably from its value c in a vacuum, it is still a large speed.

Figure 13.4.1: Index of Refraction in Various Media For light with a wavelength of 589 nm in a vacuum

Medium	n
Gases at 0°C, 1 atm	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at 20°C	
Benzene	1.501
Carbon disulfide	1.628
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20°C	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice (at 0°C)0°C)	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

✓ Example 13.4.1

Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy

We can calculate the speed of light in a material v from the index of refraction n of the material, using Equation \red{index}

Solution

Rearranging Equation 13.4.2 for v gives us

$$v = \frac{c}{n}.$$

The index of refraction for zircon is given as 1.923 in Table 13.4.1, and c is given in Equation 13.4.1. Entering these values in the equation gives

$$\begin{aligned} v &= \frac{3.00 \times 10^8 \text{ m/s}}{1.923} \\ &= 1.56 \times 10^8 \text{ m/s.} \end{aligned}$$

Significance

This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in Table 13.4.1 that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

? Exercise 13.4.1

Table 13.4.1 shows that ethanol and fresh water have very similar indices of refraction. By what percentage do the speeds of light in these liquids differ?

Answer

2.1% (to two significant figures)

Figure 13.4.2 shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it. (Some of the incident light is reflected from the surface, but for now we concentrate on the light that is transmitted.) The change in direction of the light ray depends on the relative values of the [indices of refraction](#) of the two media involved. In the situations shown, medium 2 has a greater index of refraction than medium 1. Note that as shown in Figure 13.4.1a, the direction of the ray moves closer to the perpendicular when it progresses from a medium with a lower index of refraction to one with a higher index of refraction. Conversely, as shown in Figure 13.4.1b, the direction of the ray moves away from the perpendicular when it progresses from a medium with a higher index of refraction to one with a lower index of refraction. The path is exactly reversible.

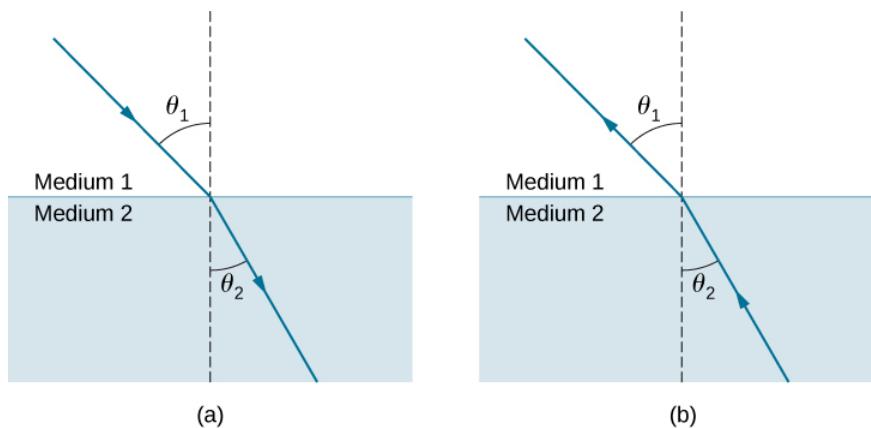


Figure 13.4.2: The change in direction of a light ray depends on how the index of refraction changes when it crosses from one medium to another. In the situations shown here, the index of refraction is greater in medium 2 than in medium 1. (a) A ray of light moves closer to the perpendicular when entering a medium with a higher index of refraction. (b) A ray of light moves away from the perpendicular when entering a medium with a lower index of refraction.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction and thus a large change in angle. The exact mathematical relationship is the law of refraction, or Snell's law, after the Dutch mathematician Willebrord Snell (1591–1626), who discovered it in 1621. The law of refraction is stated in equation form as

$$n_1 \sin \theta_1 = n_2 \sin \theta_2. \quad (13.4.3)$$

Here (n_1) and n_2 are the indices of refraction for media 1 and 2, and θ_1 and θ_2 are the angles between the rays and the perpendicular in media 1 and 2. The incoming ray is called the incident ray, the outgoing ray is called the refracted ray, and the associated angles are the incident angle and the refracted angle, respectively.

Snell's experiments showed that the law of refraction is obeyed and that a characteristic index of refraction n could be assigned to a given medium and its value measured. Snell was not aware that the speed of light varied in different media, a key fact used when we derive the law of refraction theoretically using [Huygens's Principle](#).

✓ Example 13.4.1: Determining the Index of Refraction

Find the index of refraction for medium 2 in Figure 13.4.1a, assuming medium 1 is air and given that the incident angle is 30.0° and the angle of refraction is 22.0° .

Strategy

The index of refraction for air is taken to be 1 in most cases (and up to four significant figures, it is 1.000). Thus, $n_1 = 1.00$ here. From the given information, $\theta_1 = 30.0^\circ$ and $\theta_2 = 22.0^\circ$. With this information, the only unknown in Snell's law is n_2 , so we can use Snell's law (Equation 13.4.3) to find it.

Solution

From Snell's law (Equation 13.4.3), we have

$$\begin{aligned} n_1 \sin \theta_1 &= n_2 \sin \theta_2 \\ n_2 &= n_1 \frac{\sin \theta_1}{\sin \theta_2}. \end{aligned}$$

Entering known values,

$$\begin{aligned} n_2 &= 1.00 \frac{\sin 30.0^\circ}{\sin 22.0^\circ} \\ &= \frac{0.500}{0.375} \\ &= 1.33. \end{aligned}$$

Significance

This is the index of refraction for water, and Snell could have determined it by measuring the angles and performing this calculation. He would then have found 1.33 to be the appropriate index of refraction for water in all other situations, such as when a ray passes from water to glass. Today, we can verify that the index of refraction is related to the speed of light in a medium by measuring that speed directly.

Explore [bending of light](#) between two media with different indices of refraction. Use the “Intro” simulation and see how changing from air to water to glass changes the bending angle. Use the protractor tool to measure the angles and see if you can recreate the configuration in Example 13.4.1. Also by measurement, confirm that the angle of reflection equals the angle of incidence.

✓ Example 13.4.2: A Larger Change in Direction

Suppose that in a situation like that in Example 13.4.1, light goes from air to diamond and that the incident angle is 30.0° . Calculate the angle of refraction θ_2 in the diamond.

Strategy

Again, the index of refraction for air is taken to be $n_1=1.00$, and we are given $\theta_1=30.0^\circ$. We can look up the [index of refraction for diamond](#), finding $n_2=2.419$. The only unknown in Snell’s law is θ_2 , which we wish to determine.

Solution

Solving Snell’s law (Equation 13.4.3) for $\sin \theta_2$ yields

$$\sin \theta_2 = \frac{n_1}{n_2} \sin \theta_1.$$

Entering known values,

$$\sin \theta_2 = \frac{1.00}{2.419} \sin 30.0^\circ = (0.413)(0.500) = 0.207.$$

The angle is thus

$$\theta_2 = \sin^{-1}(0.207) = 11.9^\circ.$$

Significance

For the same 30.0° angle of incidence, the angle of refraction in diamond is significantly smaller than in water (11.9° rather than 22.0° —see Example 13.4.2). This means there is a larger change in direction in diamond. The cause of a large change in direction is a large change in the index of refraction (or speed). In general, the larger the change in speed, the greater the effect on the direction of the ray.

? Exercise 13.4.1: Zircon

The solid with the next highest index of refraction after diamond is zircon. If the diamond in Example 13.4.2 were replaced with a piece of zircon, what would be the new angle of refraction?

Answer

15.1°

13.4: Refraction of Rays is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Ronald Kumon & OpenStax.

- [1.4: Refraction](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.
- [1.2: The Propagation of Light](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.

13.5: Application- Line-of-Sight Transmission

Learning Objectives

- Define line-of-sight communication.
- Determine the maximum distance of line-of-site communication using the ray model.

When a radio wave travels away from its transmitting location, it can be modeled as ray. Even excluding attenuation of the signal as it travels, a signal may not be detectable at a receiving location simply because one antenna may not "see" another because of the earth's curvature.

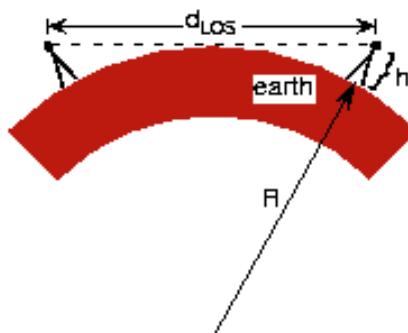


Figure 6.5.1 Two antennae are shown each having the same height. Line-of-sight transmission means the transmitting and receiving antennae can "see" each other as shown. The maximum distance at which they can see each other, d_{LOS} , occurs when the sighting line just grazes the earth's surface.

At the usual radio frequencies, propagating electromagnetic energy does not follow the earth's surface. **Line-of-sight** communication has the transmitter and receiver antennas in visual contact with each other. Assuming both antennas have height

$$d_{LOS} = 2\sqrt{2hR + H^2} \approx 2\sqrt{2Rh}$$

where R is the earth's radius and $R = 6.38 \times 10^6$ m.

? Exercise 13.5.1

Derive the expression of line-of-sight distance using only the Pythagorean Theorem. Generalize it to the case where the antennas have different heights (as is the case with commercial radio and cellular telephone). What is the range of cellular telephone where the handset antenna has essentially zero height?

Solution



Figure 6.5.2

Use the Pythagorean Theorem,

$$(h + R)^2 = R^2 + d^2$$

where h is the antenna height, d is the distance from the top of the earth to a tangency point with the earth's surface, and R the earth's radius. The line-of-sight distance between two earth-based antennae equals

$$d_{LOS} = \sqrt{2h_1R + h_1^2} + \sqrt{2h_2R + h_2^2}$$

As the earth's radius is much larger than the antenna height, we have a good approximation that

$$d_{LOS} = \sqrt{2h_1R} + \sqrt{2h_2R}$$

If one antenna is at ground elevation, say

$$h_2 = 0$$

the other antenna's range is

$$\sqrt{2h_1 R}$$

? Exercise 13.5.1

Can you imagine a situation wherein global wireless communication is possible with only one transmitting antenna? In particular, what happens to wavelength when carrier frequency decreases?

Solution

As frequency decreases, wavelength increases and can approach the distance between the earth's surface and the ionosphere. Assuming a distance between the two of 80 km, the relation $\lambda f = c$ gives a corresponding frequency of 3.75 kHz. Such low carrier frequencies would be limited to low bandwidth analog communication and to low datarate digital communications. The US Navy did use such a communication scheme to reach all of its submarines at once.

Using a 100 m antenna would provide line-of-sight transmission over a distance of 71.4 km. Using such very tall antennas would provide wireless communication within a town or between closely spaced population centers. Consequently, **networks** of antennas sprinkle the countryside (each located on the highest hill possible) to provide long-distance wireless communications: Each antenna receives energy from one antenna and retransmits to another. This kind of network is known as a **relay network**.

13.5: Application- Line-of-Sight Transmission is shared under a CC BY 1.0 license and was authored, remixed, and/or curated by Don H. Johnson.

- 6.5: Line-of-Sight Transmission by Don H. Johnson is licensed CC BY 1.0. Original source:
<https://cnx.org/contents/d442r0wh9.72:g9deOnx519>.

13.6: Diffraction of Waves

Learning Objectives

By the end of this section, you will be able to:

- Describe Huygens's principle
- Use Huygens's principle to explain the law of reflection
- Use Huygens's principle to explain the law of refraction
- Use Huygens's principle to explain diffraction

So far in this chapter, we have been discussing optical phenomena using the ray model of light. However, some phenomena require analysis and explanations based on the wave characteristics of light. This is particularly true when the wavelength is not negligible compared to the dimensions of an optical device, such as a slit in the case of **diffraction**. Huygens's principle is an indispensable tool for this analysis.

Figure 13.6.1 shows how a transverse wave looks as viewed from above and from the side. A light wave can be imagined to propagate like this, although we do not actually see it wiggling through space. From above, we view the wave fronts (or wave crests) as if we were looking down on ocean waves. The side view would be a graph of the electric or magnetic field. The view from above is perhaps more useful in developing concepts about wave optics.

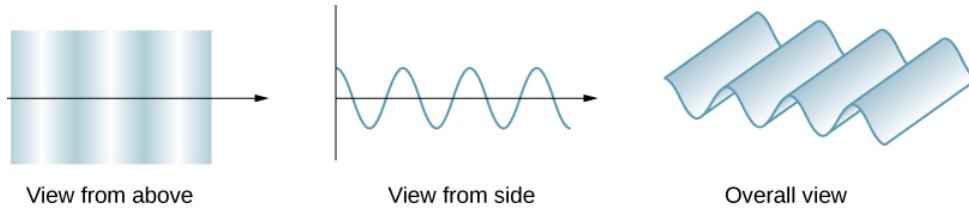


Figure 13.6.1: A transverse wave, such as an electromagnetic light wave, as viewed from above and from the side. The direction of propagation is perpendicular to the wave fronts (or wave crests) and is represented by a ray.

The Dutch scientist Christiaan Huygens (1629–1695) developed a useful technique for determining in detail how and where waves propagate. Starting from some known position, Huygens's principle states that every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wave front is tangent to all of the wavelets.

Figure 13.6.2 shows how Huygens's principle is applied. A wave front is the long edge that moves, for example, with the crest or the trough. Each point on the wave front emits a semicircular wave that moves at the propagation speed v . We can draw these wavelets at a time t later, so that they have moved a distance $s = vt$. The new wave front is a plane tangent to the wavelets and is where we would expect the wave to be a time t later. Huygens's principle works for all types of waves, including water waves, sound waves, and light waves. It is useful not only in describing how light waves propagate but also in explaining the laws of reflection and refraction. In addition, we will see that Huygens's principle tells us how and where light rays interfere.

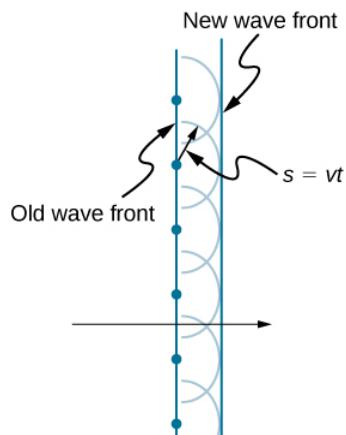


Figure 13.6.2: Huygens's principle applied to a straight wave front. Each point on the wave front emits a semicircular wavelet that moves a distance $s=vt$. The new wave front is a line tangent to the wavelets.

Reflection

Figure 13.6.3 shows how a mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection. As the wave front strikes the mirror, wavelets are first emitted from the left part of the mirror and then from the right. The wavelets closer to the left have had time to travel farther, producing a wave front traveling in the direction shown.

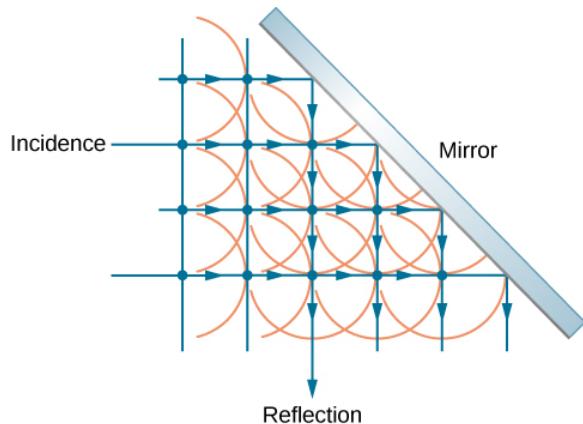


Figure 13.6.3: Huygens's principle applied to a plane wave front striking a mirror. The wavelets shown were emitted as each point on the wave front struck the mirror. The tangent to these wavelets shows that the new wave front has been reflected at an angle equal to the incident angle. The direction of propagation is perpendicular to the wave front, as shown by the downward-pointing arrows.

Refraction

The law of refraction can be explained by applying Huygens's principle to a wave front passing from one medium to another (Figure 13.6.4). Each wavelet in the figure was emitted when the wave front crossed the interface between the media. Since the speed of light is smaller in the second medium, the waves do not travel as far in a given time, and the new wave front changes direction as shown. This explains why a ray changes direction to become closer to the perpendicular when light slows down. Snell's law can be derived from the geometry in Figure 13.6.5 (Example 13.6.1).

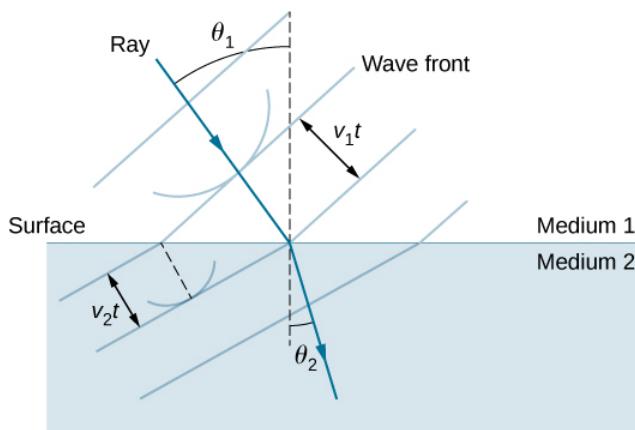


Figure 13.6.4: Huygens's principle applied to a plane wave front traveling from one medium to another, where its speed is less. The ray bends toward the perpendicular, since the wavelets have a lower speed in the second medium.

Example 13.6.1: Deriving the Law of Refraction

By examining the geometry of the wave fronts, derive the law of refraction.

Strategy

Consider Figure 13.6.5, which expands upon Figure 13.6.4. It shows the incident wave front just reaching the surface at point A, while point B is still well within medium 1. In the time Δt it takes for a wavelet from B to reach B' on the surface at speed $v_1 = c/n_1$, a wavelet from A travels into medium 2 a distance of $AA' = v_2\Delta t$, where $v_2 = c/n_2$. Note that in this example, v_2 is slower than v_1 because $n_1 < n_2$.

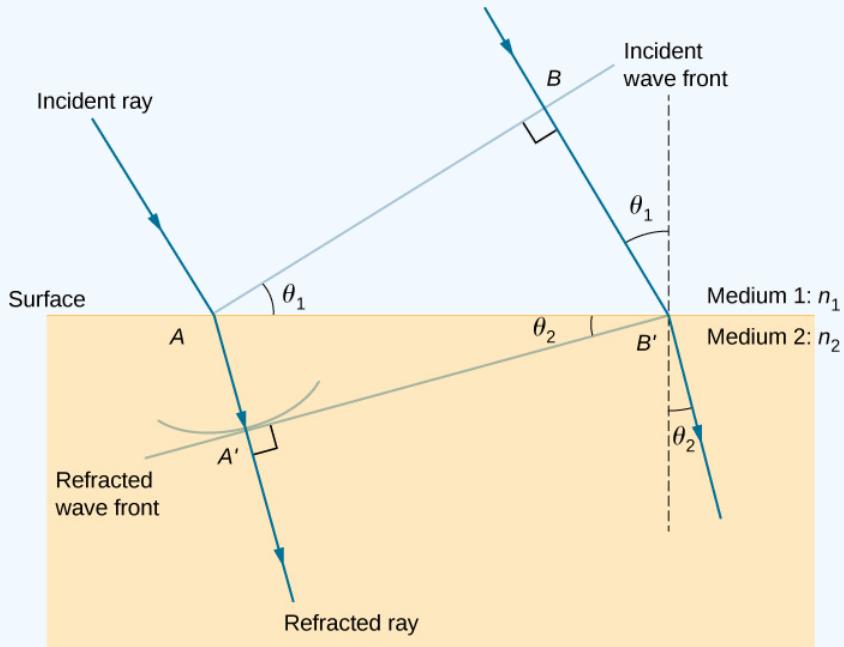


Figure 13.6.5: Geometry of the law of refraction from medium 1 to medium 2.

Solution

The segment on the surface AB' is shared by both the triangle ABB' inside medium 1 and the triangle AA'B' inside medium 2. Note that from the geometry, the angle $\angle BAB'$ is equal to the angle of incidence, θ_1 . Similarly, $\angle AB'A'$ is θ_2 .

The length of AB' is given in two ways as

$$AB' = \frac{BB'}{\sin \theta_1} = \frac{AA'}{\sin \theta_2}.$$

Inverting the equation and substituting $AA' = c\Delta t/n_2$ from above and similarly $BB' = c\Delta t/n_1$, we obtain

$$\frac{\sin \theta_1}{c\Delta t/n_1} = \frac{\sin \theta_2}{c\Delta t/n_2}.$$

Cancellation of $c\Delta t$ allows us to simplify this equation into the familiar form

$$\frac{n_1 \sin \theta_1}{n_2 \sin \theta_2}.$$

Snell's law

Significance

Although the law of refraction was established experimentally by [Snell](#), its derivation here requires Huygens's principle and the understanding that the speed of light is different in different media.

Exercise 13.6.1

In Example 13.6.1, we had $n_1 < n_2$. If n_2 were decreased such that $n_1 > n_2$ and the speed of light in medium 2 is faster than in medium 1, what would happen to the length of AA'? What would happen to the wave front A'B' and the direction of the refracted ray?

Answer

AA' becomes longer, A'B' tilts further away from the surface, and the refracted ray tilts away from the normal.

This [applet](#) by Walter Fendt shows an animation of reflection and refraction using Huygens's wavelets while you control the parameters. Be sure to click on "Next step" to display the wavelets. You can see the reflected and refracted wave fronts forming.

Diffraction

What happens when a wave passes through an opening, such as light shining through an open door into a dark room? For light, we observe a sharp shadow of the doorway on the floor of the room, and no visible light bends around corners into other parts of the room. When sound passes through a door, we hear it everywhere in the room and thus observe that sound spreads out when passing through such an opening (Figure 13.6.6). What is the difference between the behavior of sound waves and light waves in this case? The answer is that light has very short wavelengths and acts like a ray. Sound has wavelengths on the order of the size of the door and bends around corners (for frequency of 1000 Hz,

$$\lambda = \frac{c}{f} = \frac{330 \text{ m/s}}{1000 \text{ s}^{-1}} = 0.33 \text{ m},$$

about three times smaller than the width of the doorway).

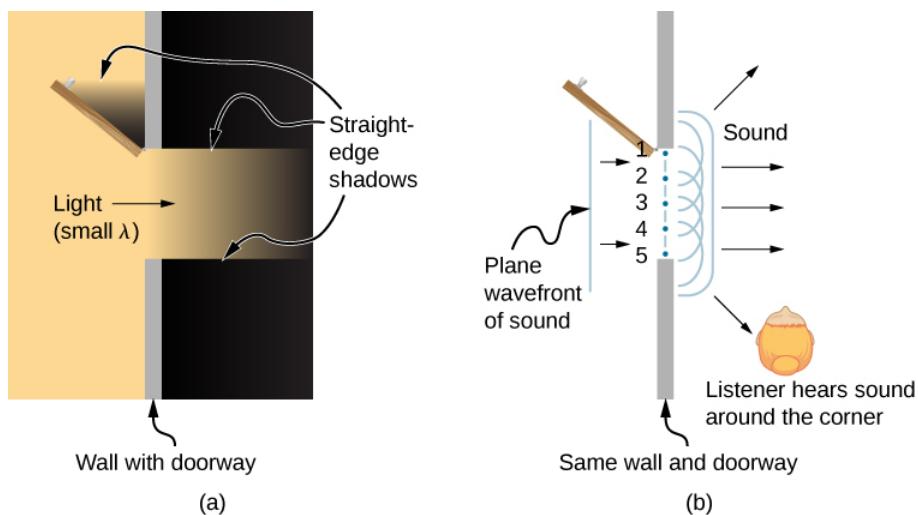


Figure 13.6.6: (a) Light passing through a doorway makes a sharp outline on the floor. Since light's wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the door.

If we pass light through smaller openings such as slits, we can use Huygens's principle to see that light bends as sound does (Figure 13.6.7). The bending of a wave around the edges of an opening or an obstacle is called diffraction. **Diffraction** is a wave characteristic and occurs for all types of waves. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Thus, the horizontal diffraction of the laser beam after it passes through the slits in Figure 13.6.7 is evidence that light is a wave.

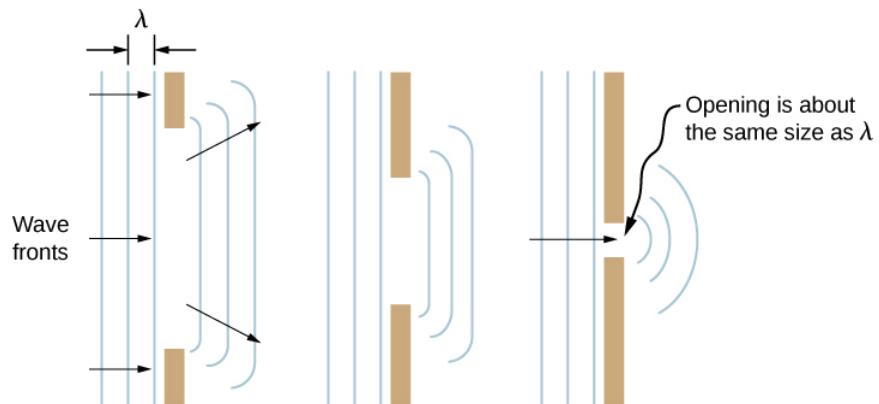


Figure 13.6.7: Huygens's principle applied to a plane wave front striking an opening. The edges of the wave front bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

13.6: Diffraction of Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 1.7: Huygens's Principle by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.

13.7: Interference of Waves

Learning Objectives

- Explain how waves are reflected and transmitted at the boundaries of a medium
- Define the terms interference and superposition
- Find the resultant wave of two identical sinusoidal waves that differ only by a phase shift

Up to now, we have been studying waves that propagate continuously through a medium, but we have not discussed what happens when waves encounter the boundary of the medium or what happens when a wave encounters another wave propagating through the same medium. Waves do interact with boundaries of the medium, and all or part of the wave can be reflected. For example, when you stand some distance from a rigid cliff face and yell, you can hear the sound waves reflect off the rigid surface as an echo. Waves can also interact with other waves propagating in the same medium. If you throw two rocks into a pond some distance from one another, the circular ripples that result from the two stones seem to pass through one another as they propagate out from where the stones entered the water. This phenomenon is known as interference. In this section, we examine what happens to waves encountering a boundary of a medium or another wave propagating in the same medium. We will see that their behavior is quite different from the behavior of particles and rigid bodies.

Reflection and Transmission

When a wave propagates through a medium, it reflects when it encounters the boundary of the medium. The wave before hitting the boundary is known as the incident wave. The wave after encountering the boundary is known as the reflected wave. How the wave is reflected at the boundary of the medium depends on the boundary conditions.

For example, mechanical waves will react differently if the boundary of the medium is fixed in place or free to move (Figure 13.7.1). A **fixed boundary condition** exists when the medium at a boundary is fixed in place so it cannot move. A **free boundary condition** exists when the medium at the boundary is free to move.

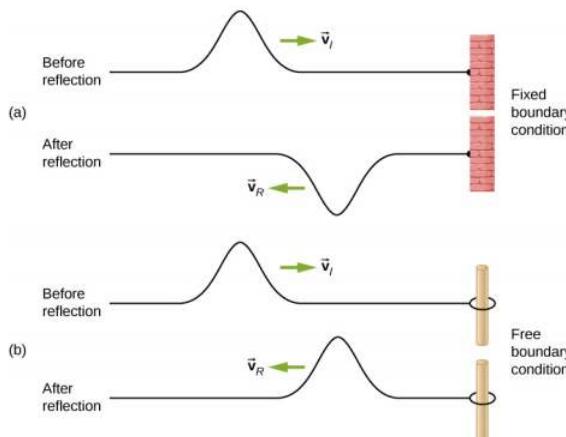


Figure 13.7.1: (a) One end of a string is fixed so that it cannot move. A wave propagating on the string, encountering this fixed boundary condition, is reflected $180^\circ(\pi \text{ rad})$ out of phase with respect to the incident wave. (b) One end of a string is tied to a solid ring of negligible mass on a frictionless lab pole, where the ring is free to move. A wave propagating on the string, encountering this free boundary condition, is reflected in phase $0^\circ(0 \text{ rad})$ with respect to the wave.

Figure 13.7.1a shows a fixed boundary condition. Here, one end of the string is fixed to a wall so the end of the string is fixed in place and the medium (the string) at the boundary cannot move. When the wave is reflected, the amplitude of the reflected wave is exactly the same as the amplitude of the incident wave, but the reflected wave is reflected $180^\circ(\pi \text{ rad})$ out of phase with respect to the incident wave. The phase change can be explained using Newton's third law: Recall that Newton's third law states that when object A exerts a force on object B, then object B exerts an equal and opposite force on object A. As the incident wave encounters the wall, the string exerts an upward force on the wall and the wall reacts by exerting an equal and opposite force on the string. The reflection at a fixed boundary is inverted. Note that the figure shows a crest of the incident wave reflected as a trough. If the incident wave were a trough, the reflected wave would be a crest.

Figure 13.7.1b shows a free boundary condition. Here, one end of the string is tied to a solid ring of negligible mass on a frictionless pole, so the end of the string is free to move up and down. As the incident wave encounters the boundary of the medium, it is also reflected. In the case of a free boundary condition, the reflected wave is in phase with respect to the incident wave. In this case, the wave encounters the free boundary applying an upward force on the ring, accelerating the ring up. The ring travels up to the maximum height equal to the amplitude of the wave and then accelerates down towards the equilibrium position due to the tension in the string. The figure shows the crest of an incident wave being reflected in phase with respect to the incident wave as a crest. If the incident wave were a trough, the reflected wave would also be a trough. The amplitude of the reflected wave would be equal to the amplitude of the incident wave.

In some situations, the boundary of the medium is neither fixed nor free. Consider Figure 13.7.2a, where a low-linear mass density string is attached to a string of a higher linear mass density. In this case, the reflected wave is out of phase with respect to the incident wave. There is also a transmitted wave that is in phase with respect to the incident wave. Both the transmitted and reflected waves have amplitudes less than the amplitude of the incident wave. If the tension is the same in both strings, the wave speed is higher in the string with the lower linear mass density.

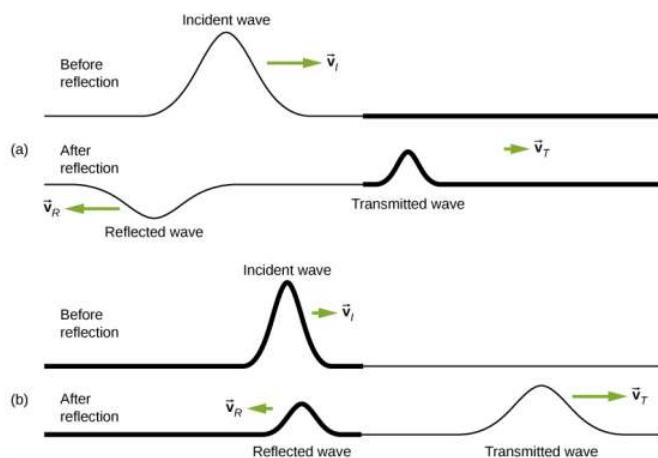


Figure 13.7.2: Waves traveling along two types of strings: a thick string with a high linear density and a thin string with a low linear density. Both strings are under the same tension, so a wave moves faster on the low-density string than on the high-density string. (a) A wave moving from a low-speed to a high-speed medium results in a reflected wave that is $180^\circ(\pi \text{ rad})$ out of phase with respect to the incident pulse (or wave) and a transmitted wave that is in phase with the incident wave. (b) When a wave moves from a low-speed medium to a high-speed medium, both the reflected and transmitted wave are in phase with respect to the incident wave.

13.7.2b shows a high-linear mass density string is attached to a string of a lower linear density. In this case, the reflected wave is in phase with respect to the incident wave. There is also a transmitted wave that is in phase with respect to the incident wave. Both the incident and the reflected waves have amplitudes less than the amplitude of the incident wave. Here you may notice that if the tension is the same in both strings, the wave speed is higher in the string with the lower linear mass density.

Superposition and Interference

Most waves do not look very simple. Complex waves are more interesting, even beautiful, but they look formidable. Most interesting mechanical waves consist of a combination of two or more traveling waves propagating in the same medium. The principle of superposition can be used to analyze the combination of waves.

Consider two simple pulses of the same amplitude moving toward one another in the same medium, as shown in Figure 13.7.3. Eventually, the waves overlap, producing a wave that has twice the amplitude, and then continue on unaffected by the encounter. The pulses are said to interfere, and this phenomenon is known as **interference**.

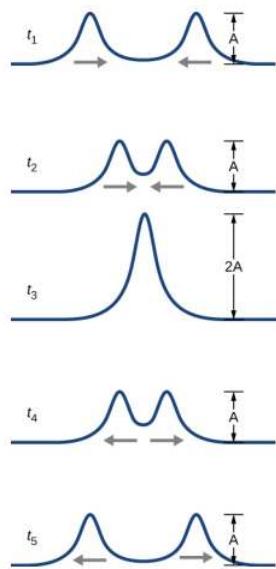


Figure 13.7.3: Two pulses moving toward one another experience interference. The term interference refers to what happens when two waves overlap.

To analyze the interference of two or more waves, we use the principle of superposition. For mechanical waves, the principle of **superposition** states that if two or more traveling waves combine at the same point, the resulting position of the mass element of the medium, at that point, is the algebraic sum of the position due to the individual waves. This property is exhibited by many waves observed, such as waves on a string, sound waves, and surface water waves. Electromagnetic waves also obey the superposition principle, but the electric and magnetic fields of the combined wave are added instead of the displacement of the medium. Waves that obey the superposition principle are linear waves; waves that do not obey the superposition principle are said to be nonlinear waves. In this chapter, we deal with linear waves, in particular, sinusoidal waves.

The superposition principle can be understood by considering the linear wave equation. In [Mathematics of a Wave](#), we defined a linear wave as a wave whose mathematical representation obeys the linear wave equation. For a transverse wave on a string with an elastic restoring force, the linear wave equation is

$$\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y(x,t)}{\partial t^2}. \quad (13.7.1)$$

Any wave function $y(x, t) = y(x \mp vt)$, where the argument of the function is linear ($x \mp vt$) is a solution to the linear wave equation and is a linear wave function. If wave functions $y_1(x, t)$ and $y_2(x, t)$ are solutions to the linear wave equation, the sum of the two functions $y_1(x, t) + y_2(x, t)$ is also a solution to the linear wave equation. Mechanical waves that obey superposition are normally restricted to waves with amplitudes that are small with respect to their wavelengths. If the amplitude is too large, the medium is distorted past the region where the restoring force of the medium is linear.

Waves can interfere constructively or destructively. Figure 13.7.4 shows two identical sinusoidal waves that arrive at the same point exactly in phase. Figure 13.7.4a and 13.7.4b show the two individual waves, Figure 13.7.4c shows the resultant wave that results from the algebraic sum of the two linear waves. The crests of the two waves are precisely aligned, as are the troughs. This superposition produces **constructive interference**. Because the disturbances add, constructive interference produces a wave that has twice the amplitude of the individual waves, but has the same wavelength.

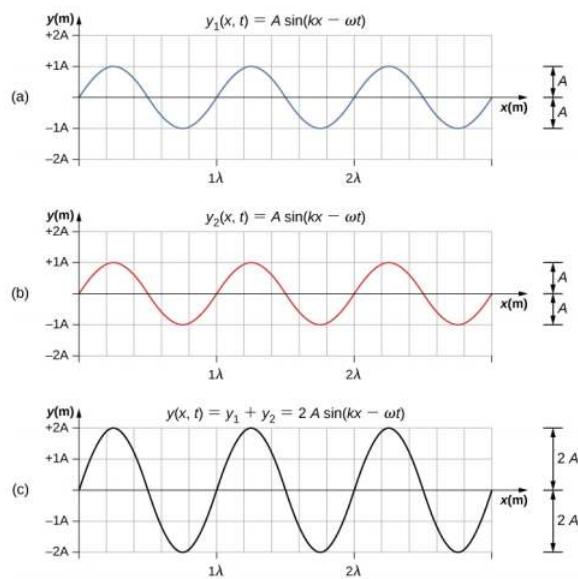


Figure 13.7.4: Constructive interference of two identical waves produces a wave with twice the amplitude, but the same wavelength.

Figure 13.7.5 shows two identical waves that arrive exactly 180° out of phase, producing **destructive interference**. Figure 13.7.5a and 13.7.5b show the individual waves, and Figure 13.7.5c shows the superposition of the two waves. Because the troughs of one wave add the crest of the other wave, the resulting amplitude is zero for destructive interference—the waves completely cancel.

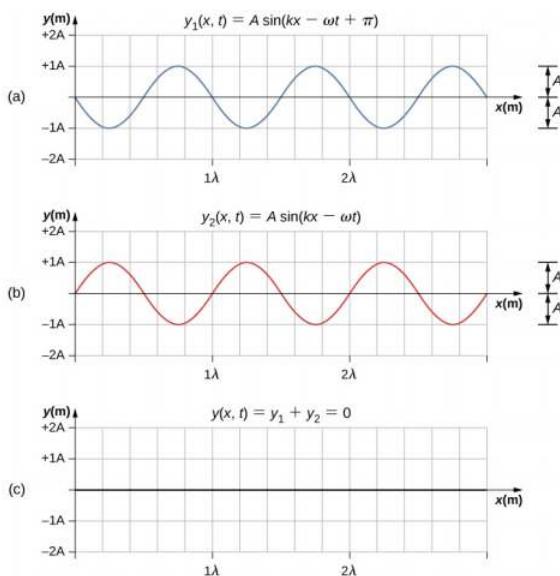


Figure 13.7.5: Destructive interference of two identical waves, one with a phase shift of $180^\circ (\pi \text{ rad})$, produces zero amplitude, or complete cancellation.

When linear waves interfere, the resultant wave is just the algebraic sum of the individual waves as stated in the principle of superposition. Figure 13.7.6 shows two waves (red and blue) and the resultant wave (black). The resultant wave is the algebraic sum of the two individual waves.

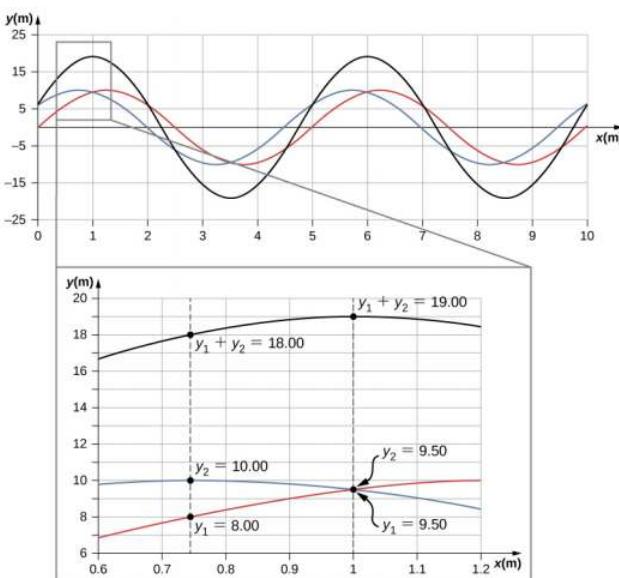


Figure 13.7.6: When two linear waves in the same medium interfere, the height of resulting wave is the sum of the heights of the individual waves, taken point by point. This plot shows two waves (red and blue) added together, along with the resulting wave (black). These graphs represent the height of the wave at each point. The waves may be any linear wave, including ripples on a pond, disturbances on a string, sound, or electromagnetic waves.

The superposition of most waves produces a combination of constructive and destructive interference, and can vary from place to place and time to time. Sound from a stereo, for example, can be loud in one spot and quiet in another. Varying loudness means the sound waves add partially constructively and partially destructively at different locations. A stereo has at least two speakers creating sound waves, and waves can reflect from walls. All these waves interfere, and the resulting wave is the superposition of the waves.

We have shown several examples of the superposition of waves that are similar. Figure 13.7.7 illustrates an example of the superposition of two dissimilar waves. Here again, the disturbances add, producing a resultant wave.

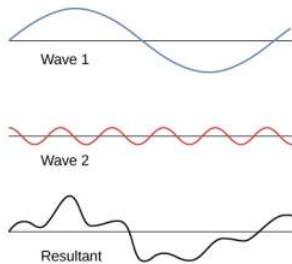


Figure 13.7.7: Superposition of nonidentical waves exhibits both constructive and destructive interference.

At times, when two or more mechanical waves interfere, the pattern produced by the resulting wave can be rich in complexity, some without any readily discernable patterns. For example, plotting the sound wave of your favorite music can look quite complex and is the superposition of the individual sound waves from many instruments; it is the complexity that makes the music interesting and worth listening to. At other times, waves can interfere and produce interesting phenomena, which are complex in their appearance and yet beautiful in simplicity of the physical principle of superposition, which formed the resulting wave. One example is the phenomenon known as standing waves, produced by two identical waves moving in different directions. We will look more closely at this phenomenon in the next section.

Simulation

Try this [simulation](#) to make waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern. You can observe one source or two sources. Using two sources, you can observe the interference patterns that result from varying the frequencies and the amplitudes of the sources.

Superposition of Sinusoidal Waves that Differ by a Phase Shift

Many examples in physics consist of two sinusoidal waves that are identical in amplitude, wave number, and angular frequency, but differ by a phase shift:

$$\begin{aligned}y_1(x, t) &= A \sin(kx - \omega t + \phi), \\y_2(x, t) &= A \sin(kx - \omega t).\end{aligned}\quad (13.7.2)$$

When these two waves exist in the same medium, the resultant wave resulting from the superposition of the two individual waves is the sum of the two individual waves:

$$y_R(x, t) = y_1(x, t) + y_2(x, t) = A \sin(kx - \omega t + \phi) + A \sin(kx - \omega t). \quad (13.7.3)$$

The resultant wave can be better understood by using the trigonometric identity:

$$\sin u + \sin v = 2 \sin\left(\frac{u+v}{2}\right) \cos\left(\frac{u-v}{2}\right), \quad (13.7.4)$$

where $u = kx - \omega t + \phi$ and $v = kx - \omega t$. The resulting wave becomes

$$\begin{aligned}y_R(x, t) &= y_1(x, t) + y_2(x, t) = A \sin(kx - \omega t + \phi) + A \sin(kx - \omega t) \\&= 2A \sin\left(\frac{(kx - \omega t + \phi) + (kx - \omega t)}{2}\right) \cos\left(\frac{(kx - \omega t + \phi) - (kx - \omega t)}{2}\right) \\&= 2A \sin\left(kx - \omega t + \frac{\phi}{2}\right) \cos\left(\frac{\phi}{2}\right).\end{aligned}\quad (13.7.5)$$

This equation is usually written as

$$y_R(x, t) = 2A \cos\left(\frac{\phi}{2}\right) \sin\left(kx - \omega t + \frac{\phi}{2}\right). \quad (13.7.6)$$

The resultant wave has the same wave number and angular frequency, an amplitude of $A_R = [2A \cos(\frac{\phi}{2})]$, and a phase shift equal to half the original phase shift. Examples of waves that differ only in a phase shift are shown in Figure 13.7.7.

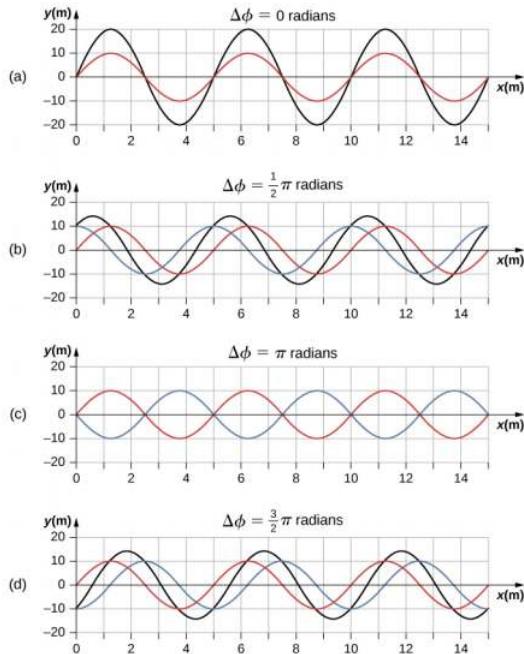


Figure 13.7.8: Superposition of two waves with identical amplitudes, wavelengths, and frequency, but that differ in a phase shift. The red wave is defined by the wave function $y_1(x, t) = A \sin(kx - \omega t)$ and the blue wave is defined by the wave function $y_2(x, t) = A \sin(kx - \omega t + \phi)$. The black line shows the result of adding the two waves. The phase difference between the two waves are (a) 0.00 rad, (b) $\frac{\pi}{2}$ rad, (c) π rad, and (d) $\frac{3\pi}{2}$ rad.

The red and blue waves each have the same amplitude, wave number, and angular frequency, and differ only in a phase shift. They therefore have the same period, wavelength, and frequency. The green wave is the result of the superposition of the two waves.

When the two waves have a phase difference of zero, the waves are in phase, and the resultant wave has the same wave number and angular frequency, and an amplitude equal to twice the individual amplitudes (part (a)). This is constructive interference. If the phase difference is 180° , the waves interfere in destructive interference (part (c)). The resultant wave has an amplitude of zero. Any other phase difference results in a wave with the same wave number and angular frequency as the two incident waves but with a phase shift of $\frac{\phi}{2}$ and an amplitude equal to $2A \cos\left(\frac{\phi}{2}\right)$. Examples are shown in parts (b) and (d).

13.7: Interference of Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Ronald Kumon & OpenStax.

- 16.6: Interference of Waves by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-1>.

13.8: Double-Slit Interference

Learning Objectives

By the end of this section, you will be able to:

- Explain the phenomenon of interference
- Define constructive and destructive interference for a double slit

The Dutch physicist Christiaan **Huygens** (1629–1695) thought that light was a wave, but Isaac Newton did not. Newton thought that there were other explanations for color, and for the interference and diffraction effects that were observable at the time. Owing to Newton's tremendous reputation, his view generally prevailed; the fact that Huygens's principle worked was not considered direct evidence proving that light is a wave. The acceptance of the wave character of light came many years later in 1801, when the English physicist and physician Thomas **Young** (1773–1829) demonstrated optical interference with his now-classic double-slit experiment.

If there were not one but two sources of waves, the waves could be made to interfere, as in the case of waves on water (Figure 13.8.1). If light is an electromagnetic wave, it must therefore exhibit interference effects under appropriate circumstances. In Young's experiment, sunlight was passed through a pinhole on a board. The emerging beam fell on two pinholes on a second board. The light emanating from the two pinholes then fell on a screen where a pattern of bright and dark spots was observed. This pattern, called fringes, can only be explained through interference, a wave phenomenon.

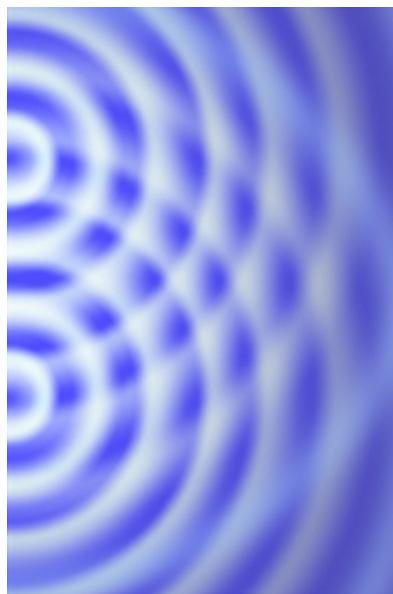


Figure 13.8.1: Photograph of an interference pattern produced by circular water waves in a ripple tank. Two thin plungers are vibrated up and down in phase at the surface of the water. Circular water waves are produced by and emanate from each plunger.

We can analyze double-slit interference with the help of Figure 13.8.2, which depicts an apparatus analogous to Young's. Light from a monochromatic source falls on a slit S_0 . The light emanating from S_0 is incident on two other slits S_1 and S_2 that are equidistant from S_0 . A pattern of **interference fringes** on the screen is then produced by the light emanating from S_1 and S_2 . All slits are assumed to be so narrow that they can be considered secondary point sources for Huygens' wavelets ([The Nature of Light](#)). Slits S_1 and S_2 are a distance d apart ($d \leq 1\text{ mm}$), and the distance between the screen and the slits is $D(\approx 1\text{ m})$, which is much greater than d .

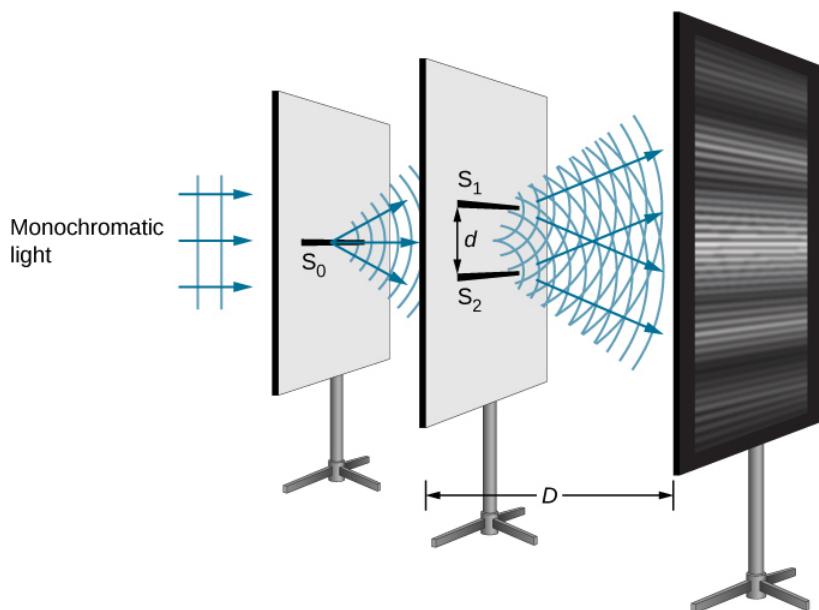


Figure 13.8.2: The double-slit interference experiment using monochromatic light and narrow slits. Fringes produced by interfering Huygens wavelets from slits S_1 and S_2 are observed on the screen.

Since S_0 is assumed to be a point source of monochromatic light, the secondary Huygens wavelets leaving S_1 and S_2 always maintain a constant phase difference (zero in this case because S_1 and S_2 are equidistant from S_0) and have the same frequency. The sources S_1 and S_2 are then said to be coherent. By coherent waves, we mean the waves are in phase or have a definite phase relationship. The term incoherent means the waves have random phase relationships, which would be the case if S_1 and S_2 were illuminated by two independent light sources, rather than a single source S_0 . Two independent light sources (which may be two separate areas within the same lamp or the Sun) would generally not emit their light in unison, that is, not coherently. Also, because S_1 and S_2 are the same distance from S_0 , the amplitudes of the two Huygens wavelets are equal.

Young used sunlight, where each wavelength forms its own pattern, making the effect more difficult to see. In the following discussion, we illustrate the double-slit experiment with monochromatic light (single λ) to clarify the effect. Figure 13.8.3 shows the pure constructive and destructive interference of two waves having the same wavelength and amplitude.

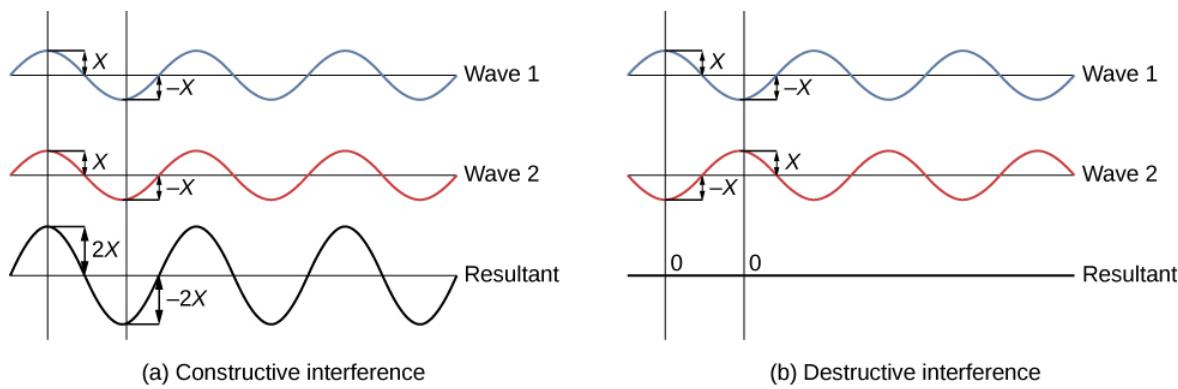


Figure 13.8.3: The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

When light passes through narrow slits, the slits act as sources of coherent waves and light spreads out as semicircular waves, as shown in Figure 13.8.1a. Pure **constructive interference** occurs where the waves are crest to crest or trough to trough. Pure **destructive interference** occurs where they are crest to trough. The light must fall on a screen and be scattered into our eyes for us to see the pattern. An analogous pattern for water waves is shown in Figure 13.8.1. Note that regions of constructive and destructive interference move out from the slits at well-defined angles to the original beam. These angles depend on wavelength and the distance between the slits, as we shall see below.

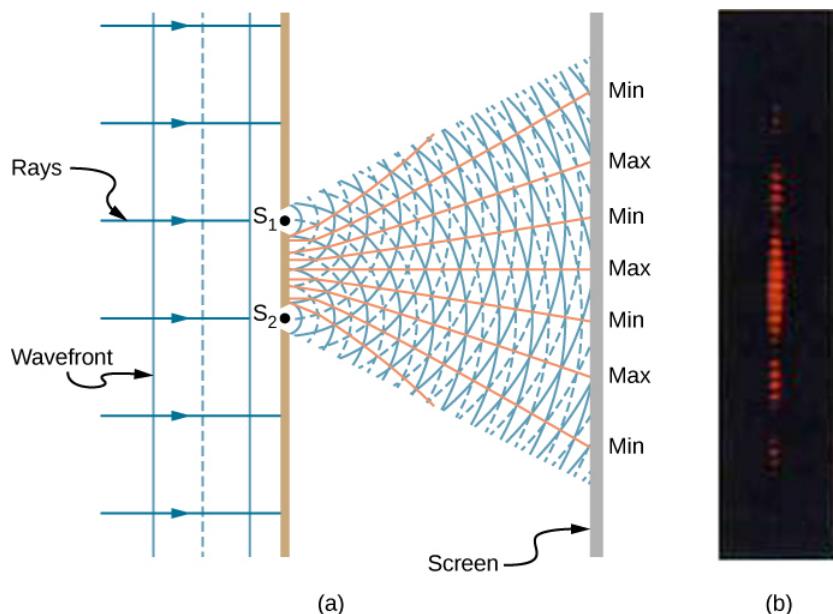


Figure 13.8.4: Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto a screen and is scattered into our eyes. (b) When light that has passed through double slits falls on a screen, we see a pattern such as this.

To understand the double-slit interference pattern, consider how two waves travel from the slits to the screen (Figure 13.8.5). Each slit is a different distance from a given point on the screen. Thus, different numbers of wavelengths fit into each path. Waves start out from the slits in phase (crest to crest), but they may end up out of phase (crest to trough) at the screen if the paths differ in length by half a wavelength, interfering destructively. If the paths differ by a whole wavelength, then the waves arrive in phase (crest to crest) at the screen, interfering constructively. More generally, if the path length difference Δl between the two waves is any half-integral number of wavelengths [($1/2\lambda$), ($3/2\lambda$), ($5/2\lambda$), etc.], then destructive interference occurs. Similarly, if the path length difference is any integral number of wavelengths (λ , 2λ , 3λ , etc.), then constructive interference occurs. These conditions can be expressed as equations:

$$\underbrace{\Delta l = m\lambda}_{\text{constructive interference}}$$

for $m = 0, \pm 1, \pm 2, \pm 3 \dots$

$$\Delta l = \underbrace{\left(m + \frac{1}{2} \right) \lambda}_{\text{destructive interference}}$$

for $m = 0, \pm 1, \pm 2, \pm 3 \dots$

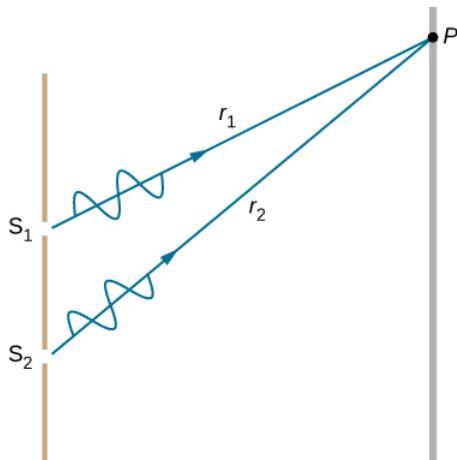


Figure 13.8.5: Waves follow different paths from the slits to a common point P on a screen. Destructive interference occurs where one path is a half wavelength longer than the other—the waves start in phase but arrive out of phase. Constructive interference occurs where one path is a whole wavelength longer than the other—the waves start out and arrive in phase.

13.8: Double-Slit Interference is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 3.2: Young's Double-Slit Interference by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.

13.9: Propagation of Electromagnetic Waves (Summary)

Key Terms

corner reflector	object consisting of two (or three) mutually perpendicular reflecting surfaces, so that the light that enters is reflected back exactly parallel to the direction from which it came
constructive interference	when two waves arrive at the same point exactly in phase; that is, the crests of the two waves are precisely aligned, as are the troughs
destructive interference	when two identical waves arrive at the same point exactly out of phase; that is, precisely aligned crest to trough
geometric optics	part of optics dealing with the ray aspect of light
Huygens's principle	every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself; the new wave front is a plane tangent to all of the wavelets
index of refraction	for a material, the ratio of the speed of light in a vacuum to that in a material
interference	overlap of two or more waves at the same point and time
law of reflection	angle of reflection equals the angle of incidence
law of refraction	when a light ray crosses from one medium to another, it changes direction by an amount that depends on the index of refraction of each medium and the sines of the angle of incidence and angle of refraction
ray	straight line that originates at some point
refraction	changing of a light ray's direction when it passes through variations in matter
superposition	phenomenon that occurs when two or more waves arrive at the same point
wave optics	part of optics dealing with the wave aspect of light

Key Equations

Speed of light	$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$
Index of refraction	$n = \frac{c}{v}$
Law of reflection	$\theta_r = \theta_i$
Law of refraction (Snell's law)	$n_1 \sin \theta_1 = n_2 \sin \theta_2$

Summary

Ray and Wave Models of Propagation

- The speed of light in a vacuum is $c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$.
- The index of refraction of a material is $n = c/v$, where v is the speed of light in a material and c is the speed of light in a vacuum.
- The ray model of light describes the path of light as straight lines. The part of optics dealing with the ray aspect of light is called geometric optics.

- Light can travel in three ways from a source to another location: (1) directly from the source through empty space; (2) through various media; and (3) after being reflected from a mirror.

Reflection of Rays

- When a light ray strikes a smooth surface, the angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.

Refraction of Rays

- The change of a light ray's direction when it passes through variations in matter is called refraction.
- The law of refraction, also called Snell's law, relates the indices of refraction for two media at an interface to the change in angle of a light ray passing through that interface.

Application: Line-of-Sight Transmission

- Using the ray model, there is a maximum distance of transmission of a ray due to the curvature of the earth.

Diffraction of Waves

- According to Huygens's principle, every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wave front is tangent to all of the wavelets.
- A mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection.
- The law of refraction can be explained by applying Huygens's principle to a wave front passing from one medium to another.
- The bending of a wave around the edges of an opening or an obstacle is called diffraction.

Interference of Waves

- Superposition is the combination of two waves at the same location.
- Constructive interference occurs from the superposition of two identical waves that are in phase.
- Destructive interference occurs from the superposition of two identical waves that are 180° (π radians) out of phase.
- The wave that results from the superposition of two sine waves that differ only by a phase shift is a wave with an amplitude that depends on the value of the phase difference.

Double-Slit Interference

- Young's double-slit experiment gave definitive proof of the wave character of light.
- An interference pattern is obtained by the superposition of light from two slits.

13.9: Propagation of Electromagnetic Waves (Summary) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

- 1.S: The Nature of Light (Summary) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.
- 16.S: Waves (Summary) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-1>.
- 3.S: Interference (Summary) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.

13.10: Propagation of Electromagnetic Waves (Exercises)

Conceptual Questions

1.1 The Propagation of Light

1. Under what conditions can light be modeled like a ray? Like a wave?
2. Why is the index of refraction always greater than or equal to 1?
3. Does the fact that the light flash from lightning reaches you before its sound prove that the speed of light is extremely large or simply that it is greater than the speed of sound? Discuss how you could use this effect to get an estimate of the speed of light.
4. Speculate as to what physical process might be responsible for light traveling more slowly in a medium than in a vacuum.

1.2 The Law of Reflection

5. Using the law of reflection, explain how powder takes the shine off of a person's nose. What is the name of the optical effect?

1.3 Refraction

6. Diffusion by reflection from a rough surface is described in this chapter. Light can also be diffused by refraction. Describe how this occurs in a specific situation, such as light interacting with crushed ice.
7. Will light change direction toward or away from the perpendicular when it goes from air to water? Water to glass? Glass to air?
8. Explain why an object in water always appears to be at a depth shallower than it actually is?
9. Explain why a person's legs appear very short when wading in a pool. Justify your explanation with a ray diagram showing the path of rays from the feet to the eye of an observer who is out of the water.
10. Explain why an oar that is partially submerged in water appears bent.

1.4 Total Internal Reflection

11. A ring with a colorless gemstone is dropped into water. The gemstone becomes invisible when submerged. Can it be a diamond? Explain.
12. The most common type of mirage is an illusion that light from faraway objects is reflected by a pool of water that is not really there. Mirages are generally observed in deserts, when there is a hot layer of air near the ground. Given that the refractive index of air is lower for air at higher temperatures, explain how mirages can be formed.
13. How can you use total internal reflection to estimate the index of refraction of a medium?

1.5 Dispersion

14. Is it possible that total internal reflection plays a role in rainbows? Explain in terms of indices of refraction and angles, perhaps referring to that shown below. Some of us have seen the formation of a double rainbow; is it physically possible to observe a triple rainbow? A photograph of a double rainbow.



15. A high-quality diamond may be quite clear and colorless, transmitting all visible wavelengths with little absorption. Explain how it can sparkle with flashes of brilliant color when illuminated by white light.

1.6 Huygens's Principle

16. How do wave effects depend on the size of the object with which the wave interacts? For example, why does sound bend around the corner of a building while light does not?
17. Does Huygens's principle apply to all types of waves?
18. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Does the reverse hold true? That is, if diffraction is not observed, does that mean the phenomenon is not a wave?

1.7 Polarization

19. Can a sound wave in air be polarized? Explain.
20. No light passes through two perfect polarizing filters with perpendicular axes. However, if a third polarizing filter is placed between the original two, some light can pass. Why is this? Under what circumstances does most of the light pass?
21. Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.
22. When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to $\frac{1}{\lambda}$. Does this mean there is more scattering for small λ than large λ ? How does this relate to the fact that the sky is blue?
23. Using the information given in the preceding question, explain why sunsets are red.
24. When light is reflected at Brewster's angle from a smooth surface, it is 100% polarized parallel to the surface. Part of the light will be refracted into the surface. Describe how you would do an experiment to determine the polarization of the refracted light. What direction would you expect the polarization to have and would you expect it to be 100%?
25. If you lie on a beach looking at the water with your head tipped slightly sideways, your polarized sunglasses do not work very well. Why not?

Problems

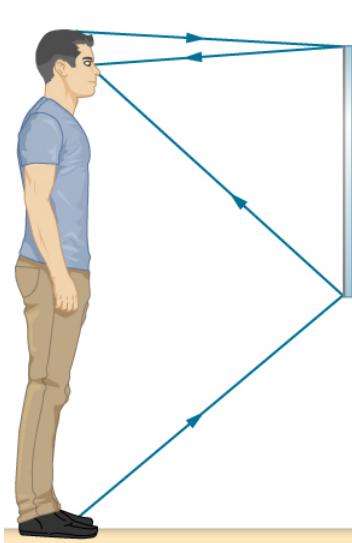
1.1 The Propagation of Light

26. What is the speed of light in water? In glycerine?
27. What is the speed of light in air? In crown glass?
28. Calculate the index of refraction for a medium in which the speed of light is $2.012 \times 10^8 \text{ m/s}$, and identify the most likely substance based on Table 1.1.
29. In what substance in Table 1.1 is the speed of light $2.290 \times 10^8 \text{ m/s}$?
30. There was a major collision of an asteroid with the Moon in medieval times. It was described by monks at Canterbury Cathedral in England as a red glow on and around the Moon. How long after the asteroid hit the Moon, which is $3.84 \times 10^5 \text{ km}$ away, would the light first arrive on Earth?

31. Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?
32. Compare the time it takes for light to travel 1000 m on the surface of Earth and in outer space.
33. How far does light travel underwater during a time interval of $1.50 \times 10^{-6} s$?

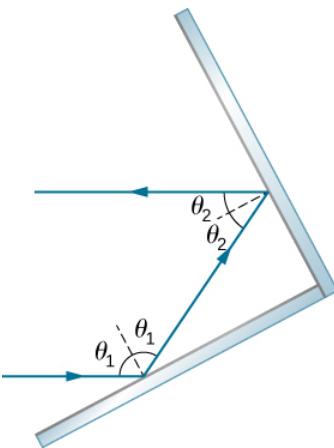
1.2 The Law of Reflection

34. Suppose a man stands in front of a mirror as shown below. His eyes are 1.65 m above the floor and the top of his head is 0.13 m higher. Find the height above the floor of the top and bottom of the smallest mirror in which he can see both the top of his head and his feet. How is this distance related to the man's height?



The figure is a drawing of a man standing in front of a mirror and looking at his image. The mirror is about half as tall as the man, with the top of the mirror above his eyes but below the top of his head. The light rays from his feet reach the bottom of the mirror and reflect to his eyes. The rays from the top of his head reach the top of the mirror and reflect to his eyes.

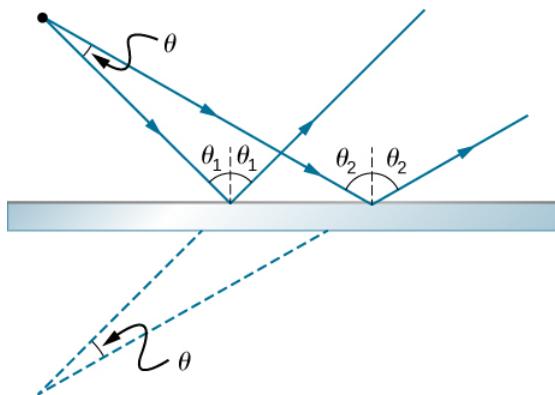
35. Show that when light reflects from two mirrors that meet each other at a right angle, the outgoing ray is parallel to the incoming ray, as illustrated below.



Two mirrors meet each other at a right angle. An incoming ray of light hits one mirror at an angle of theta one to the normal, is reflected at the same angle of theta one on the other side of the normal, then hits the other mirror at an angle of theta two to the normal and reflects at the same angle of theta two on the other side of the normal, such that the outgoing ray is parallel to the incoming ray.

36. On the Moon's surface, lunar astronauts placed a corner reflector, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. What percent correction is needed to account for the delay in time due to the slowing of light in Earth's atmosphere? Assume the distance to the Moon is precisely $3.84 \times 10^8 \text{ m}$ and Earth's atmosphere (which varies in density with altitude) is equivalent to a layer 30.0 km thick with a constant index of refraction $n = 1.000293$.

37. A flat mirror is neither converging nor diverging. To prove this, consider two rays originating from the same point and diverging at an angle θ (see below). Show that after striking a plane mirror, the angle between their directions remains θ .



Light rays diverging from a point at an angle theta are incident on a mirror at two different places and their reflected rays diverge. One ray hits at an angle theta one from the normal, and reflects at the same angle theta one on the other side of the normal. The other ray hits at a larger angle theta two from the normal, and reflects at the same angle theta two on the other side of the normal. When the reflected rays are extended backwards from their points of reflection, they meet at a point behind the mirror, at the same angle theta with which they left the source.

1.3 Refraction

Unless otherwise specified, for problems 1 through 10, the indices of refraction of glass and water should be taken to be 1.50 and 1.333, respectively.

38. A light beam in air has an angle of incidence of 35° at the surface of a glass plate. What are the angles of reflection and refraction?

39. A light beam in air is incident on the surface of a pond, making an angle of $20^\circ 20'$ with respect to the surface. What are the angles of reflection and refraction?

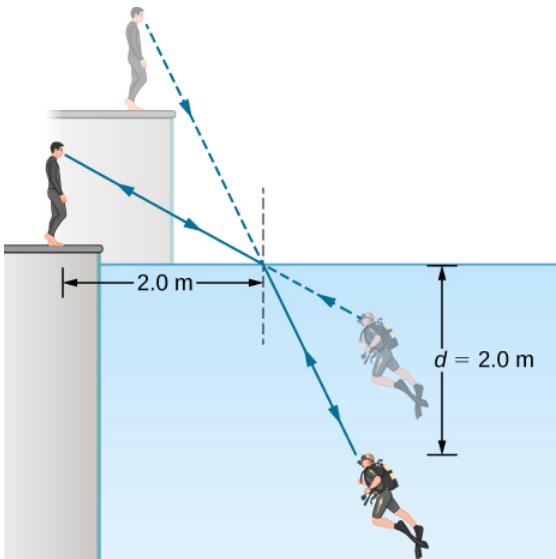
40. When a light ray crosses from water into glass, it emerges at an angle of 30° with respect to the normal of the interface. What is its angle of incidence?

41. A pencil flashlight submerged in water sends a light beam toward the surface at an angle of incidence of 30° . What is the angle of refraction in air?

42. Light rays from the Sun make a 30° angle to the vertical when seen from below the surface of a body of water. At what angle above the horizon is the Sun?

43. The path of a light beam in air goes from an angle of incidence of 35° to an angle of refraction of 22° when it enters a rectangular block of plastic. What is the index of refraction of the plastic?

44. A scuba diver training in a pool looks at his instructor as shown below. What angle does the ray from the instructor's face make with the perpendicular to the water at the point where the ray enters? The angle between the ray in the water and the perpendicular to the water is 25.0° .



A scuba diver and his trainer look at each other. They see each other at the locations given by straight line extrapolations of the rays reaching their eyes. To the trainer, the scuba diver appears less deep than he actually is, and to the diver, the trainer appears higher than he actually is. To the trainer, the scuba diver's feet appear to be at a depth of two point zero meters. The incident ray from the trainer strikes the water surface at a horizontal distance of two point zero meters from the trainer. The diver's head is a vertical distance of d equal to two point zero meters below the surface of the water.

45. (a) Using information in the preceding problem, find the height of the instructor's head above the water, noting that you will first have to calculate the angle of incidence.

(b) Find the apparent depth of the diver's head below water as seen by the instructor.

1.4 Total Internal Reflection

46. Verify that the critical angle for light going from water to air is 48.6° , as discussed at the end of Example 1.4, regarding the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air.

47. (a) At the end of Example 1.4, it was stated that the critical angle for light going from diamond to air is 24.4° . Verify this.

(b) What is the critical angle for light going from zircon to air?

48. An optical fiber uses flint glass clad with crown glass. What is the critical angle?

49. At what minimum angle will you get total internal reflection of light traveling in water and reflected from ice?

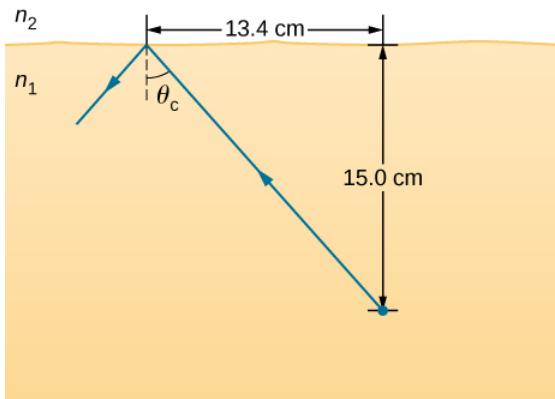
50. Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is 45.0° , what must be the minimum index of refraction of the material from which the reflector is made?

51. You can determine the index of refraction of a substance by determining its critical angle.

(a) What is the index of refraction of a substance that has a critical angle of 68.4° when submerged in water? What is the substance, based on Table 1.1?

(b) What would the critical angle be for this substance in air?

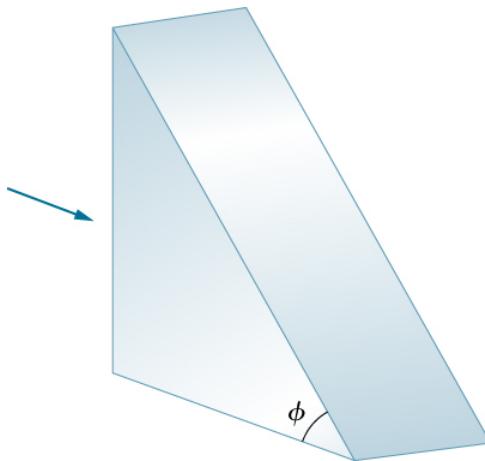
52. A ray of light, emitted beneath the surface of an unknown liquid with air above it, undergoes total internal reflection as shown below. What is the index of refraction for the liquid and its likely identification?



A light ray travels from an object placed in a medium n_1 at 15.0 centimeters below the horizontal interface with medium n_2 . This ray gets totally internally reflected with theta c as critical angle. The horizontal distance between the object and the point of incidence is 13.4 centimeters.

53. Light rays fall normally on the vertical surface of the glass prism ($n = 1.50$ shown below).

- (a) What is the largest value for ϕ such that the ray is totally reflected at the slanted face?
 (b) Repeat the calculation of part (a) if the prism is immersed in water.



A right angle triangular prism has a horizontal base and a vertical side. The hypotenuse of the triangle makes an angle of phi with the horizontal base. A horizontal light rays is incident normally on the vertical surface of the prism.

1.5 Dispersion

54. (a) What is the ratio of the speed of red light to violet light in diamond, based on Table 1.2?

- (b) What is this ratio in polystyrene?
 (c) Which is more dispersive?

55. A beam of white light goes from air into water at an incident angle of 75.0° . At what angles are the red (660 nm) and violet (410 nm) parts of the light refracted?

56. By how much do the critical angles for red (660 nm) and violet (410 nm) light differ in a diamond surrounded by air?

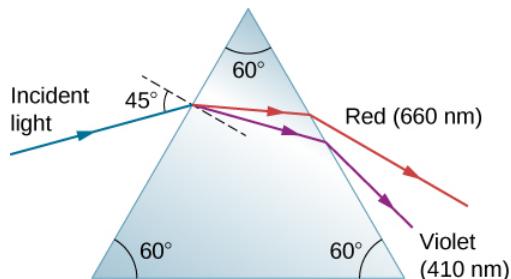
57. (a) A narrow beam of light containing yellow (580 nm) and green (550 nm) wavelengths goes from polystyrene to air, striking the surface at a 30.0° incident angle. What is the angle between the colors when they emerge?

- (b) How far would they have to travel to be separated by 1.00 mm?

- 58.** A parallel beam of light containing orange (610 nm) and violet (410 nm) wavelengths goes from fused quartz to water, striking the surface between them at a 60.0° incident angle. What is the angle between the two colors in water?
- 59.** A ray of 610-nm light goes from air into fused quartz at an incident angle of 55.0° . At what incident angle must 470 nm light enter flint glass to have the same angle of refraction?
- 60.** A narrow beam of light containing red (660 nm) and blue (470 nm) wavelengths travels from air through a 1.00-cm-thick flat piece of crown glass and back to air again. The beam strikes at a 30.0° incident angle.

- (a) At what angles do the two colors emerge?
- (b) By what distance are the red and blue separated when they emerge?

- 61.** A narrow beam of white light enters a prism made of crown glass at a 45.0° incident angle, as shown below. At what angles, θ_R and θ_V , do the red (660 nm) and violet (410 nm) components of the light emerge from the prism?



A blue incident light ray at an angle of incidence equal to 45 degrees to the normal falls on an equilateral triangular prism whose corners are all at angles equal to 60 degrees. At the first surface, the ray refracts and splits into red and violet rays. These rays hit the second surface and emerge from the prism. The red light with 660 nanometers bends less than the violet light with 410 nanometers.

1.7 Polarization

- 62.** What angle is needed between the direction of polarized light and the axis of a polarizing filter to cut its intensity in half?
- 63.** The angle between the axes of two polarizing filters is 45.0° . By how much does the second filter reduce the intensity of the light coming through the first?
- 64.** Two polarizing sheets P_1 and P_2 are placed together with their transmission axes oriented at an angle θ to each other. What is θ when only 25 of the maximum transmitted light intensity passes through them?
- 65.** Suppose that in the preceding problem the light incident on P_1 is unpolarized. At the determined value of θ , what fraction of the incident light passes through the combination?
- 66.** If you have completely polarized light of intensity $150W/m^2$, what will its intensity be after passing through a polarizing filter with its axis at an 89.0° angle to the light's polarization direction?
- 67.** What angle would the axis of a polarizing filter need to make with the direction of polarized light of intensity $1.00kW/m^2$ to reduce the intensity to $10.0W/m^2$?
- 68.** At the end of Example 1.7, it was stated that the intensity of polarized light is reduced to 90.0 of its original value by passing through a polarizing filter with its axis at an angle of 18.4° to the direction of polarization. Verify this statement.
- 69.** Show that if you have three polarizing filters, with the second at an angle of 45.0° to the first and the third at an angle of 90.0° to the first, the intensity of light passed by the first will be reduced to 25.0 of its value. (This is in contrast to having only the first and third, which reduces the intensity to zero, so that placing the second between them increases the intensity of the transmitted light.)
- 70.** Three polarizing sheets are placed together such that the transmission axis of the second sheet is oriented at 25.0° to the axis of the first, whereas the transmission axis of the third sheet is oriented at 40.0° (in the same sense) to the axis of the first. What fraction of the intensity of an incident unpolarized beam is transmitted by the combination?

71. In order to rotate the polarization axis of a beam of linearly polarized light by 90.0° , a student places sheets P_1 and P_2 with their transmission axes at 45.0° and 90.0° , respectively, to the beam's axis of polarization.

- (a) What fraction of the incident light passes through P_1 and
- (b) through the combination?
- (c) Repeat your calculations for part (b) for transmission-axis angles of 30.0° and 90.0° , respectively.

72. It is found that when light traveling in water falls on a plastic block, Brewster's angle is 50.0° . What is the refractive index of the plastic?

73. At what angle will light reflected from diamond be completely polarized?

74. What is Brewster's angle for light traveling in water that is reflected from crown glass?

75. A scuba diver sees light reflected from the water's surface. At what angle will this light be completely polarized?

Additional Problems

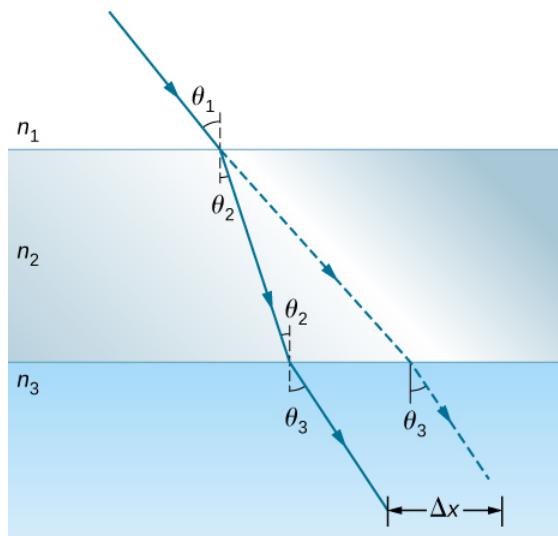
76. From his measurements, Roemer estimated that it took 22 min for light to travel a distance equal to the diameter of Earth's orbit around the Sun.

- (a) Use this estimate along with the known diameter of Earth's orbit to obtain a rough value of the speed of light.
- (b) Light actually takes 16.5 min to travel this distance. Use this time to calculate the speed of light.

77. Cornu performed Fizeau's measurement of the speed of light using a wheel of diameter 4.00 cm that contained 180 teeth. The distance from the wheel to the mirror was 22.9 km. Assuming he measured the speed of light accurately, what was the angular velocity of the wheel?

78. Suppose you have an unknown clear substance immersed in water, and you wish to identify it by finding its index of refraction. You arrange to have a beam of light enter it at an angle of 45.0° , and you observe the angle of refraction to be 40.3° . What is the index of refraction of the substance and its likely identity?

79. Shown below is a ray of light going from air through crown glass into water, such as going into a fish tank. Calculate the amount the ray is displaced by the glass (Δx), given that the incident angle is 40.0° , and the glass is 1.00 cm thick.



The figure illustrates refraction occurring when light travels from medium n_1 to n_3 through an intermediate medium n_2 . The incident ray makes an angle θ_1 with a perpendicular drawn at the point of incidence at the interface between n_1 and n_2 . The light ray entering n_2 bends towards the perpendicular line making an angle θ_2 with it on the n_2 side. The ray arrives at the interface between n_2 and n_3 at an angle of θ_2 to a perpendicular drawn at the point of incidence at this interface, and the transmitted ray bends away from the perpendicular, making an angle of theta three to the perpendicular on the n_3 side. A straight line extrapolation of the original

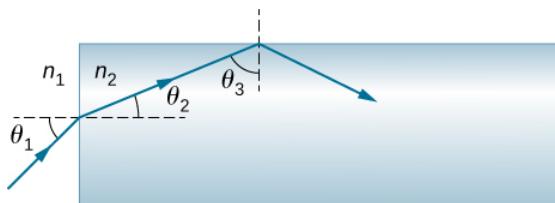
incident ray is shown as a dotted line. This line is parallel to the refracted ray in the third medium, n_3 , and is shifted a distance delta x from the refracted ray. The extrapolated ray is at the same angle theta three to the perpendicular in medium n_3 as the refracted ray.

80. Considering the previous problem, show that θ_3 is the same as it would be if the second medium were not present.
81. At what angle is light inside crown glass completely polarized when reflected from water, as in a fish tank?
82. Light reflected at 55.6° from a window is completely polarized. What is the window's index of refraction and the likely substance of which it is made?
83. (a) Light reflected at 62.5° from a gemstone in a ring is completely polarized. Can the gem be a diamond?
 (b) At what angle would the light be completely polarized if the gem was in water?
84. If θ_b is Brewster's angle for light reflected from the top of an interface between two substances, and θ'_b is Brewster's angle for light reflected from below, prove that $\theta_b + \theta'_b = 90.0^\circ$.
85. **Unreasonable results** Suppose light travels from water to another substance, with an angle of incidence of 10.0° and an angle of refraction of 14.9° .
 - (a) What is the index of refraction of the other substance?
 - (b) What is unreasonable about this result?
 - (c) Which assumptions are unreasonable or inconsistent?
86. **Unreasonable results** Light traveling from water to a gemstone strikes the surface at an angle of 80.0° and has an angle of refraction of 15.2° .
 - (a) What is the speed of light in the gemstone?
 - (b) What is unreasonable about this result?
 - (c) Which assumptions are unreasonable or inconsistent?
87. If a polarizing filter reduces the intensity of polarized light to 50.0 of its original value, by how much are the electric and magnetic fields reduced?
88. Suppose you put on two pairs of polarizing sunglasses with their axes at an angle of 15.0° . How much longer will it take the light to deposit a given amount of energy in your eye compared with a single pair of sunglasses? Assume the lenses are clear except for their polarizing characteristics.
89. (a) On a day when the intensity of sunlight is 1.00 kW/m^2 , a circular lens 0.200 m in diameter focuses light onto water in a black beaker. Two polarizing sheets of plastic are placed in front of the lens with their axes at an angle of 20.0° . Assuming the sunlight is unpolarized and the polarizers are 100 efficient, what is the initial rate of heating of the water in $^\circ\text{C/s}$, assuming it is 80.0 absorbed? The aluminum beaker has a mass of 30.0 grams and contains 250 grams of water.
 (b) Do the polarizing filters get hot? Explain.

Challenge Problems

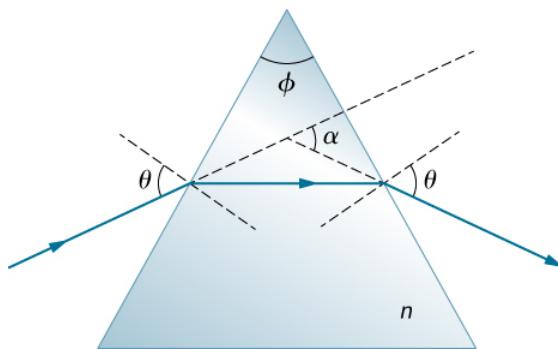
90. Light shows staged with lasers use moving mirrors to swing beams and create colorful effects. Show that a light ray reflected from a mirror changes direction by 2θ when the mirror is rotated by an angle θ .
91. Consider sunlight entering Earth's atmosphere at sunrise and sunset—that is, at a 90.0° incident angle. Taking the boundary between nearly empty space and the atmosphere to be sudden, calculate the angle of refraction for sunlight. This lengthens the time the Sun appears to be above the horizon, both at sunrise and sunset. Now construct a problem in which you determine the angle of refraction for different models of the atmosphere, such as various layers of varying density. Your instructor may wish to guide you on the level of complexity to consider and on how the index of refraction varies with air density.

92. A light ray entering an optical fiber surrounded by air is first refracted and then reflected as shown below. Show that if the fiber is made from crown glass, any incident ray will be totally internally reflected.



The figure shows light traveling from n_1 and incident onto the left face of a rectangular block of material n_2 . The ray is incident at an angle of incidence θ_1 , measured relative to the normal to the surface where the ray enters. The angle of refraction is θ_2 , again, relative to the normal to the surface. The refracted ray falls onto the upper face of the block and gets totally internally reflected with θ_3 as the angle of incidence.

93. A light ray falls on the left face of a prism (see below) at the angle of incidence θ for which the emerging beam has an angle of refraction θ at the right face. Show that the index of refraction n of the glass prism is given by $n = \frac{\sin \frac{1}{2}(\alpha + \phi)}{\sin \frac{1}{2}\phi}$ where ϕ is the vertex angle of the prism and α is the angle through which the beam has been deviated. If $\alpha = 37.0^\circ$ and the base angles of the prism are each 50.0° , what is n ?



A light ray falls on the left face of a triangular prism whose upper vertex has an angle of ϕ and whose index of refraction is n . The angle of incidence of the ray relative to the normal to the left face is θ . The ray refracts in the prism. The refracted ray is horizontal, parallel to the base of the prism. The refracted ray reaches the right face of the prism and refracts as it emerges out of the prism. The emerging ray makes an angle of θ with the normal to the right face.

94. If the apex angle ϕ in the previous problem is 20.0° and $n = 1.50$, what is the value of α ?
95. The light incident on polarizing sheet P_1 is linearly polarized at an angle of 30.0° with respect to the transmission axis of P_1 . Sheet P_2 is placed so that its axis is parallel to the polarization axis of the incident light, that is, also at 30.0° with respect to P_1 .
- What fraction of the incident light passes through P_1 ?
 - What fraction of the incident light is passed by the combination?
 - By rotating P_2 , a maximum in transmitted intensity is obtained. What is the ratio of this maximum intensity to the intensity of transmitted light when P_2 is at 30.0° with respect to P_1 ?
96. Prove that if I is the intensity of light transmitted by two polarizing filters with axes at an angle θ and I' is the intensity when the axes are at an angle $90.0^\circ - \theta$, then $I + I' = I_0$, the original intensity. (Hint: Use the trigonometric identities $\cos 90.0^\circ - \theta = \sin \theta$ and $\cos^2 \theta + \sin^2 \theta = 1$.)

- **1.E: The Nature of Light (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.

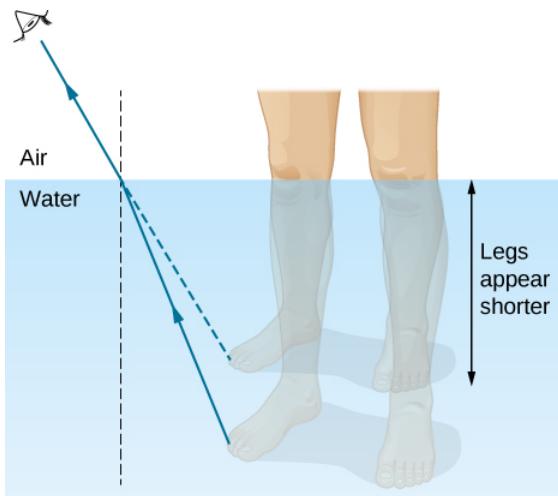
13.11: Propagation of Electromagnetic Waves (Answers)

Check Your Understanding

- 1.1. 2.1% (to two significant figures)
- 1.2. 15.1°
- 1.3. air to water, because the condition that the second medium must have a smaller index of refraction is not satisfied
- 1.4. 9.3 cm
- 1.5. AA' becomes longer, $A'B'$ tilts further away from the surface, and the refracted ray tilts away from the normal.
- 1.6. also 90.0
- 1.7. There will be only refraction but no reflection.

Conceptual Questions

1. model as a ray when devices are large compared to wavelength, as a wave when devices are comparable or small compared to wavelength
3. This fact simply proves that the speed of light is greater than that of sound. If one knows the distance to the location of the lightning and the speed of sound, one could, in principle, determine the speed of light from the data. In practice, because the speed of light is so great, the data would have to be known to impractically high precision.
5. Powder consists of many small particles with randomly oriented surfaces. This leads to diffuse reflection, reducing shine.
7. “toward” when increasing n (air to water, water to glass); “away” when decreasing n (glass to air)
9. A ray from a leg emerges from water after refraction. The observer in air perceives an apparent location for the source, as if a ray traveled in a straight line. See the dashed ray below.



The figure is illustration of the formation of the image of a leg under water, as seen by a viewer in the air above the water. A ray is shown leaving the leg and refracting at the water air interface. The refracted ray bends away from the normal. Extrapolating the refracted ray back into the water, the extrapolated ray is above the actual ray so that the image of the leg is above the actual leg and the leg appears shorter.

11. The gemstone becomes invisible when its index of refraction is the same, or at least similar to, the water surrounding it. Because diamond has a particularly high index of refraction, it can still sparkle as a result of total internal reflection, not invisible.
13. One can measure the critical angle by looking for the onset of total internal reflection as the angle of incidence is varied. Equation 1.5 can then be applied to compute the index of refraction.

15. In addition to total internal reflection, rays that refract into and out of diamond crystals are subject to dispersion due to varying values of n across the spectrum, resulting in a sparkling display of colors.
17. yes
19. No. Sound waves are not transverse waves.
21. Energy is absorbed into the filters.
23. Sunsets are viewed with light traveling straight from the Sun toward us. When blue light is scattered out of this path, the remaining red light dominates the overall appearance of the setting Sun.
25. The axis of polarization for the sunglasses has been rotated 90° .

Problems

27. $2.99705 \times 10^8 \text{ m/s}$; $1.97 \times 10^8 \text{ m/s}$

29. ice at 0°C

31. 1.03 ns

33. 337 m

35. proof

37. proof

39. reflection, 70° ; refraction, 45°

41. 42°

43. 1.53

45. a. 2.9 m;

b. 1.4 m

47. a. 24.42° ;

b. 31.33°

49. 79.11°

51. a. 1.43, fluorite;

b. 44.2°

53. a. 48.2° ;

b. 27.3°

55. 46.5° for red, 46.0° for violet

57. a. 0.04° ;

b. 1.3 m

59. 72.8°

61. 53.5° for red, 55.2° for violet

63. 0.500

65. 0.125 or 1/8

67. 84.3°

69. $0.250I_0$

71. a. 0.500;

- b. 0.250;
- c. 0.187

73. 67.54°

75. 53.1°

Additional Problems

77. 114 radian/s

79. 3.72 mm

81. 41.2°

83. a. 1.92. The gem is not a diamond (it is zircon).

- b. 55.2°

85. a. 0.898;

b. We cannot have $n < 1.00$, since this would imply a speed greater than c.

c. The refracted angle is too big relative to the angle of incidence.

87. $0.707B_1$

89. a. $1.69 \times 10^{-2} \text{ C/s}$;

- b. yes

Challenge Problems

91. First part: 88.6° . The remainder depends on the complexity of the solution the reader constructs.

93. proof; 1.33

95. a. 0.750;

- b. 0.563;

- c. 1.33

This page titled [13.11: Propagation of Electromagnetic Waves \(Answers\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [1.A: The Nature of Light \(Answers\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-3>.

CHAPTER OVERVIEW

14: Introduction to Semiconductor Devices

- [14.1: Introduction](#)
- [14.2: Band Theory of Solids](#)
- [14.3: Semiconductors and Doping](#)
- [14.4: Introduction to Semiconductor Devices](#)
- [14.5: Junction Diodes](#)
- [14.6: Light Emitting Diode](#)
- [14.7: Solar Cells](#)
- [14.8: Bipolar Junction Transistors](#)
- [14.9: Junction Field-effect Transistors](#)

[14: Introduction to Semiconductor Devices](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

14.1: Introduction

Semiconductors play a critical role in modern electronics. The electrical properties of these materials can be used to form the solid-state electronic components which are used in nearly every electronic device including computers, cellular phones, radios, and televisions. In this chapter, we will provide an introduction to the modern theory of band theory of solids which has proven essential for understanding the behavior of these systems. We will then describe how semiconductor materials can be used to construct diodes, light-emitting diodes, transistors, bipolar junction transistors, and junction field-effect transistors.

This page titled [14.1: Introduction](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Ronald Kumon](#).

14.2: Band Theory of Solids

Learning Objectives

By the end of this section, you will be able to:

- Describe two main approaches to determining the energy levels of an electron in a crystal
- Explain the presence of energy bands and gaps in the energy structure of a crystal
- Explain why some materials are good conductors and others are good insulators
- Differentiate between an insulator and a semiconductor

The free electron model explains many important properties of conductors but is weak in at least two areas. First, it assumes a constant potential energy within the solid. (Recall that a constant potential energy is associated with no forces.) Figure 14.2.1 compares the assumption of a constant potential energy (dotted line) with the periodic Coulomb potential, which drops as $-1/r$ at each lattice point, where r is the distance from the ion core (solid line). Second, the free electron model assumes an impenetrable barrier at the surface. This assumption is not valid, because under certain conditions, electrons can escape the surface—such as in the photoelectric effect. In addition to these assumptions, the free electron model does not explain the dramatic differences in electronic properties of conductors, semiconductors, and insulators. Therefore, a more complete model is needed.

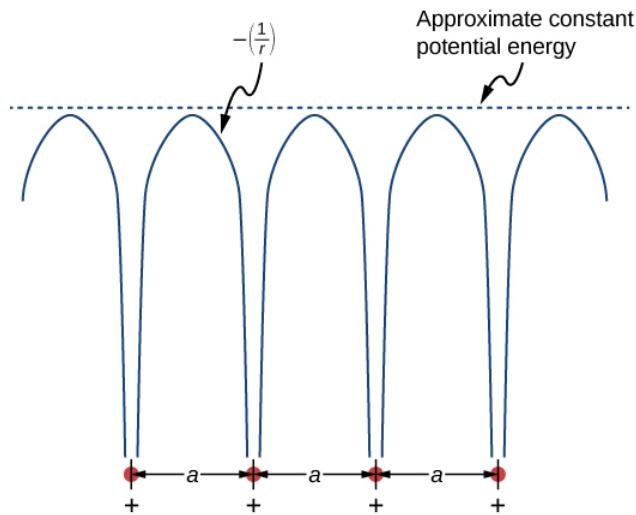


Figure 14.2.1: The periodic potential used to model electrons in a conductor. Each ion in the solid is the source of a Coulomb potential. Notice that the free electron model is productive because the average of this field is approximately constant.

We can produce an improved model by solving Schrödinger's equation for the periodic potential shown in Figure 14.2.1. However, the solution requires technical mathematics far beyond our scope. We again seek a qualitative argument based on quantum mechanics to find a way forward.

We first review the argument used to explain the energy structure of a covalent bond. Consider two identical hydrogen atoms so far apart that there is no interaction whatsoever between them. Further suppose that the electron in each atom is in the same ground state: a 1s electron with an energy of **-13.6 eV** (ignore spin). When the hydrogen atoms are brought closer together, the individual wave functions of the electrons overlap and, by the exclusion principle, can no longer be in the same quantum state, which splits the original equivalent energy levels into two different energy levels. The energies of these levels depend on the interatomic distance, a (Figure 14.2.2a).

If four hydrogen atoms are brought together, four levels are formed from the four possible symmetries—a single sine wave “hump” in each well, alternating up and down, and so on. In the limit of a very large number N of atoms, we expect a spread of nearly continuous bands of electronic energy levels in a solid (Figure 14.2.2c). Each of these bands is known as an **energy band**. (The allowed states of energy and wave number are still technically quantized, but for large numbers of atoms, these states are so close together that they are considered to be continuous or “in the continuum.”)

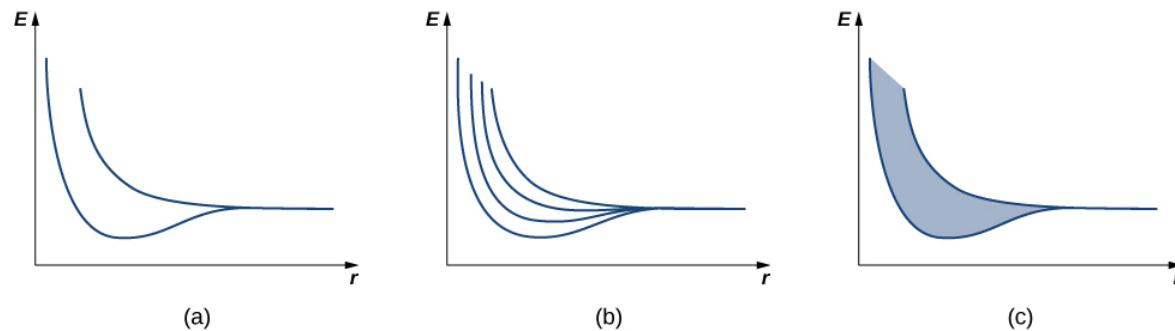


Figure 14.2.2: The dependence of energy-level splitting on the average distance between (a) two atoms, (b) four atoms, and (c) a large number of atoms. For a large number of electrons, a continuous band of energies is produced

Energy bands differ in the number of electrons they hold. In the 1s and 2s energy bands, each energy level holds up to two electrons (spin up and spin down), so this band has a maximum occupancy of $2N$ electrons. In the 2p energy band, each energy level holds up to six electrons, so this band has a maximum occupancy of $6N$ electrons (Figure 14.2.3).

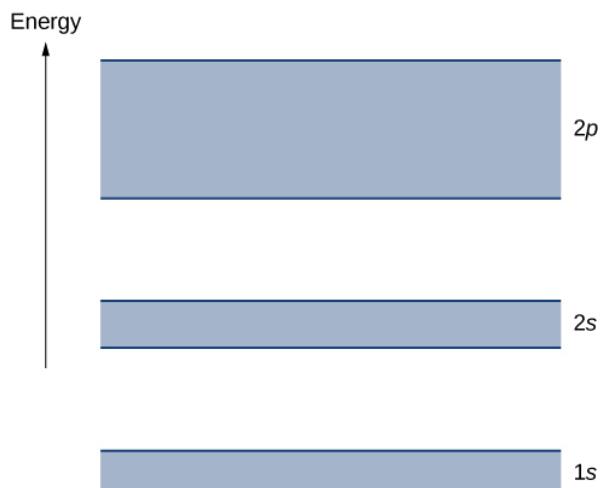


Figure 14.2.3: A simple representation of the energy structure of a solid. Electrons belong to energy bands separated by energy gaps.

Each energy band is separated from the other by an **energy gap**. The electrical properties of conductors and insulators can be understood in terms of energy bands and gaps. The highest energy band that is filled is known as a **valence band**. The next available band in the energy structure is known as a **conduction band**. In a conductor, the highest energy band that contains electrons is partially filled, whereas in an insulator, the highest energy band containing electrons is completely filled. The difference between a conductor and insulator is illustrated in Figure 14.2.4.

A conductor differs from an insulator in how its electrons respond to an applied electric field. If a significant number of electrons are set into motion by the field, the material is a conductor. In terms of the band model, electrons in the partially filled conduction band gain kinetic energy from the electric field by filling higher energy states in the conduction band. By contrast, in an insulator, electrons belong to completely filled bands. When the field is applied, the electrons cannot make such transitions (acquire kinetic energy from the electric field) due to the exclusion principle. As a result, the material does not conduct electricity.

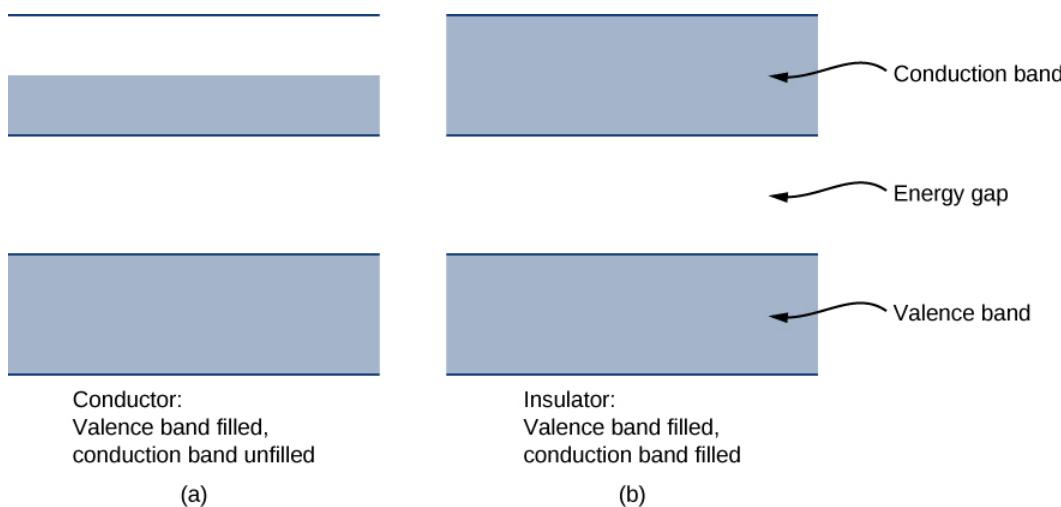


Figure 14.2.4: Comparison of a conductor and insulator. The highest energy band is partially filled in a conductor but completely filled in an insulator.

Simulation

Visit this [simulation](#) to learn about the origin of energy bands in crystals of atoms and how the structure of bands determines how a material conducts electricity. Explore how band structure creates a lattice of many wells.

A semiconductor has a similar energy structure to an insulator except it has a relatively small energy gap between the lowest completely filled band and the next available unfilled band. This type of material forms the basis of modern electronics. At $T = 0\text{ K}$, the semiconductor and insulator both have completely filled bands. The only difference is in the size of the energy gap (or **band gap**) E_g between the highest energy band that is filled (the valence band) and the next-higher empty band (the conduction band). In a semiconductor, this gap is small enough that a substantial number of electrons from the valence band are thermally excited into the conduction band at room temperature. These electrons are then in a nearly empty band and can respond to an applied field. As a general rule of thumb, the band gap of a semiconductor is about 1 eV. (Table 14.2.1 for silicon.) A band gap of greater than approximately 1 eV is considered an insulator. For comparison, the energy gap of diamond (an insulator) is several electron-volts.

Table 14.2.1: Energy Gap for Various Materials at 300 K Note: Except for diamond, the materials listed are all semiconductors.

Material	Energy Gap E_g (eV)
Si	1.14
Ge	0.67
GaAs	1.43
GaP	2.26
GaSb	0.69
InAs	0.35
InP	1.35
InSb	0.16
C(diamond)	5.48

This page titled [14.2: Band Theory of Solids](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [9.6: Band Theory of Solids](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-3>.

14.3: Semiconductors and Doping

Learning Objectives

By the end of this section, you will be able to:

- Describe changes to the energy structure of a semiconductor due to doping
- Distinguish between an n-type and p-type semiconductor
- Describe the Hall effect and explain its significance
- Calculate the charge, drift velocity, and charge carrier number density of a semiconductor using information from a Hall effect experiment

In the preceding section, we considered only the contribution to the electric current due to electrons occupying states in the conduction band. However, moving an electron from the valence band to the conduction band leaves an unoccupied state or **hole** in the energy structure of the valence band, which a nearby electron can move into. As these holes are filled by other electrons, new holes are created. The electric current associated with this filling can be viewed as the collective motion of many negatively charged electrons or the motion of the positively charged electron holes.

To illustrate, consider the one-dimensional lattice in Figure 14.3.1. Assume that each lattice atom contributes one valence electron to the current. As the hole on the right is filled, this hole moves to the left. The current can be interpreted as the flow of positive charge to the left. The density of holes, or the number of holes per unit volume, is represented by p . Each electron that transitions into the conduction band leaves behind a hole. If the conduction band is originally empty, the conduction electron density p is equal to the hole density, that is, $n = p$.

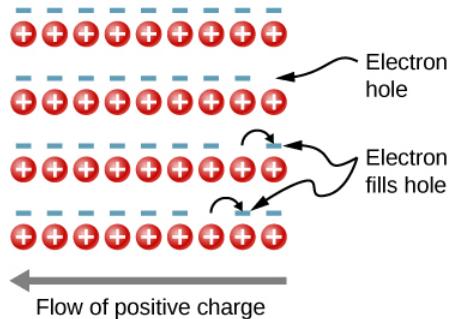


Figure 14.3.1: The motion of holes in a crystal lattice. As electrons shift to the right, an electron hole moves to the left.

As mentioned, a semiconductor is a material with a filled valence band, an unfilled conduction band, and a relatively small energy gap between the bands. Excess electrons or holes can be introduced into the material by the substitution into the crystal lattice of an impurity atom, which is an atom of a slightly different valence number. This process is known as doping. For example, suppose we add an arsenic atom to a crystal of silicon (Figure 14.3.2a).

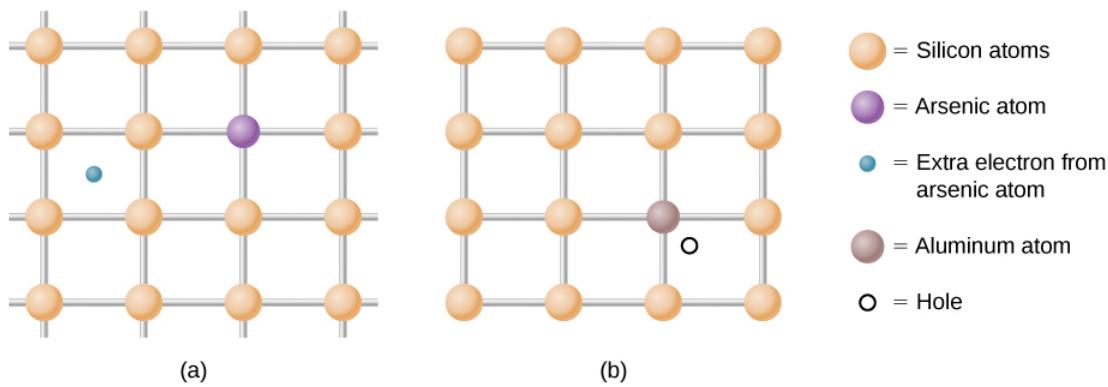


Figure 14.3.2: (a) A donor impurity and (b) an acceptor impurity. The introduction to impurities and acceptors into a semiconductor significantly changes the electronic properties of this material.

Arsenic has five valence electrons, whereas silicon has only four. This extra electron must therefore go into the conduction band, since there is no room in the valence band. The arsenic ion left behind has a net positive charge that weakly binds the delocalized electron. The binding is weak because the surrounding atomic lattice shields the ion's electric field. As a result, the binding energy of the extra electron is only about 0.02 eV. In other words, the energy level of the impurity electron is in the band gap below the conduction band by 0.02 eV, a much smaller value than the energy of the gap, 1.14 eV. At room temperature, this impurity electron is easily excited into the conduction band and therefore contributes to the conductivity (Figure 14.3.3a). An impurity with an extra electron is known as a **donor impurity**, and the doped semiconductor is called an **n-type** semiconductor because the primary carriers of charge (electrons) are negative.

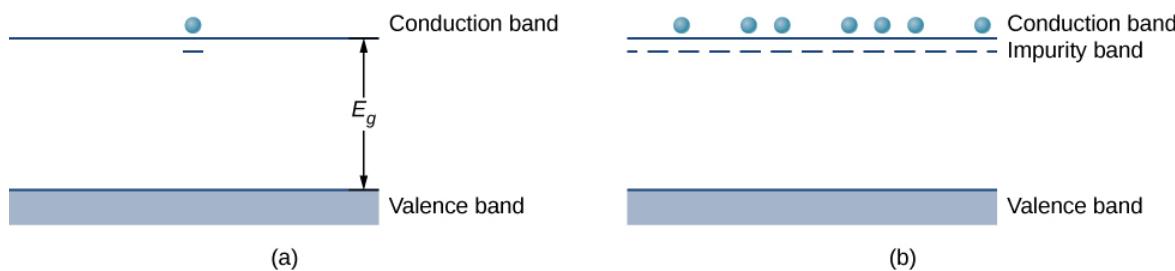


Figure 14.3.3: The extra electron from a donor impurity is excited into the conduction band; (b) formation of an impurity band in an n-type semiconductor.

By adding more donor impurities, we can create an **impurity band**, a new energy band created by semiconductor doping, as shown in Figure 14.3.3b. The Fermi level is now between this band and the conduction band. At room temperature, many impurity electrons are thermally excited into the conduction band and contribute to the conductivity. Conduction can then also occur in the impurity band as vacancies are created there. Note that changes in the energy of an electron correspond to a change in the motion (velocities or kinetic energy) of these charge carriers with the semiconductor, but not the bulk motion of the semiconductor itself.

Doping can also be accomplished using impurity atoms that typically have one **fewer** valence electron than the semiconductor atoms. For example, Al, which has three valence electrons, can be substituted for Si, as shown in Figure 14.3.2b. Such an impurity is known as an **acceptor impurity**, and the doped semiconductor is called a **p-type** semiconductor, because the primary carriers of charge (holes) are positive. If a hole is treated as a positive particle weakly bound to the impurity site, then an empty electron state is created in the band gap just above the valence band. When this state is filled by an electron thermally excited from the valence band (Figure 14.3.1a), a mobile hole is created in the valence band. By adding more acceptor impurities, we can create an impurity band, as shown in Figure 14.3.1b.

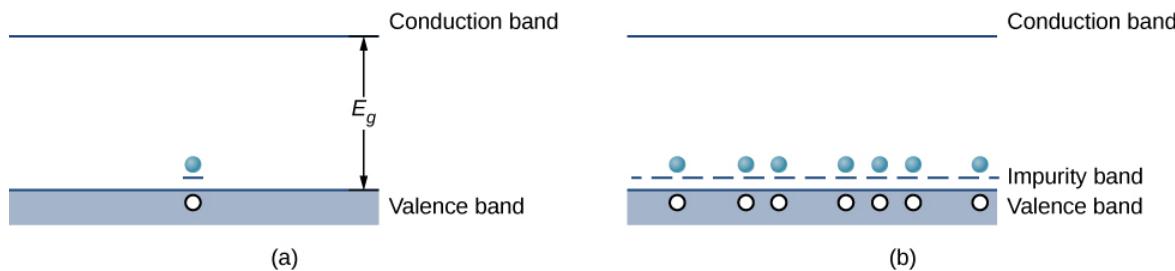


Figure 14.3.4: (a) An electron from the conduction band is excited into the empty state resulting from the acceptor impurity; (b) formation of an impurity band in a p-type semiconductor.

The electric current of a doped semiconductor can be due to the motion of a **majority carrier**, in which holes are contributed by an impurity atom, or due to a **minority carrier**, in which holes are contributed purely by thermal excitations of electrons across the energy gap. In an **n-type** semiconductor, majority carriers are free electrons contributed by impurity atoms, and minority carriers are free electrons produced by thermal excitations from the valence to the conduction band. In a **p-type** semiconductor, the majority carriers are free holes contributed by impurity atoms, and minority carriers are free holes left by the filling of states due to thermal excitation of electrons across the gap. In general, the number of majority carriers far exceeds the minority carriers. The concept of a majority and minority carriers will be used in the next section to explain the operation of diodes and transistors.

Hall Effect

In studying **p**- and **n**-type doping, it is natural to ask: Do “electron holes” really act like particles? The existence of holes in a doped **p**-type semiconductor is demonstrated by the [Hall effect](#). The Hall effect is the production of a potential difference due to the motion of a conductor through an external magnetic field. A schematic of the Hall effect is shown in Figure 14.3.5a.

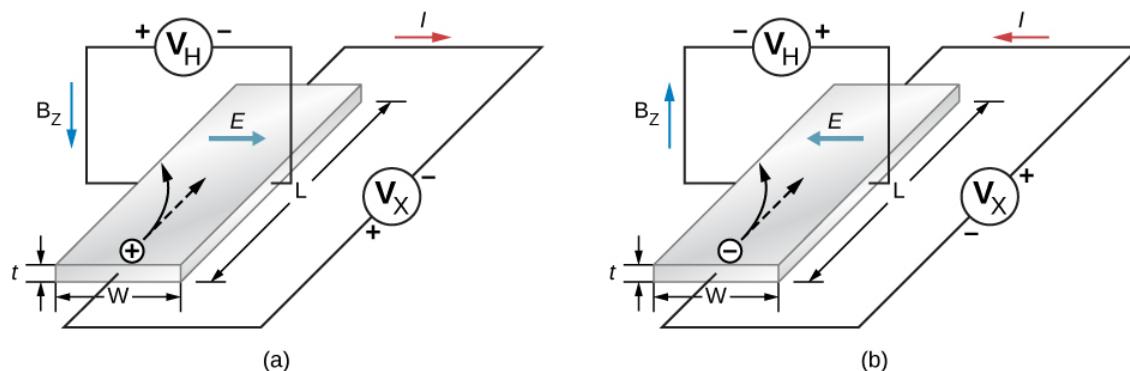


Figure 14.3.5: The Hall effect. (a) Positively charged electron holes are drawn to the left by a uniform magnetic field that points downward. An electric field is generated to the right. (b) Negative charged electrons are drawn to the left by a magnetic field that points up. An electric field is generated to the left.

A semiconductor strip is bathed in a uniform magnetic field (which points into the paper). As the electron holes move from left to right through the semiconductor, a **Lorentz force** drives these charges toward the upper end of the strip. (Recall that the motion of the positively charged carriers is determined by the right-hand rule.) Positive charge continues to collect on the upper edge of the strip until the force associated with the downward electric field between the upper and lower edges of the strip ($F_E = E_q$) just balances the upward magnetic force ($F_B = qvB$). Setting these forces equal to each other, we have $E = vB$. The voltage that develops across the strip is therefore

$$V_H = vBw,$$

where V_H is the Hall voltage; v is the hole’s **drift velocity**, or average velocity of a particle that moves in a partially random fashion; \mathbf{B} is the magnetic field strength; and w is the width of the strip. Note that the Hall voltage is transverse to the voltage that initially produces current through the material. A measurement of the sign of this voltage (or potential difference) confirms the collection of holes on the top side of the strip. The magnitude of the Hall voltage yields the drift velocity (\mathbf{v}) of the majority carriers.

Additional information can also be extracted from the Hall voltage. Note that the electron current density (the amount of current per unit cross-sectional area of the semiconductor strip) is

$$j = nqv, \quad (14.3.1)$$

where q is the magnitude of the charge, n is the number of charge carriers per unit volume, and \mathbf{v} is the drift velocity. The current density is easily determined by dividing the total current by the cross-sectional area of the strip, q is charge of the hole (the magnitude of the charge of a single electron), and \mathbf{u} is determined by Equation 14.3.1. Hence, the above expression for the electron current density gives the number of charge carriers per unit volume, n . A similar analysis can be conducted for negatively charged carriers in an **n**-type material (see Figure 14.3.5).

This page titled [14.3: Semiconductors and Doping](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [9.7: Semiconductors and Doping](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-3>.

14.4: Introduction to Semiconductor Devices

Learning Objectives

By the end of this section, you will be able to:

- Describe what occurs when n- and p-type materials are joined together using the concept of diffusion and drift current (zero applied voltage)
- Explain the response of a p-n junction to a forward and reverse bias voltage
- Describe the function of a transistor in an electric circuit
- Use the concept of a p-n junction to explain its applications in audio amplifiers and computers

Semiconductors have many applications in modern electronics. We describe some basic semiconductor devices in this section. A great advantage of using semiconductors for circuit elements is the fact that many thousands or millions of semiconductor devices can be combined on the same tiny piece of silicon and connected by conducting paths. The resulting structure is called an integrated circuit (ic), and ic chips are the basis of many modern devices, from computers and smartphones to the internet and global communications networks.

Diodes

Perhaps the simplest device that can be created with a semiconductor is a diode. A diode is a circuit element that allows electric current to flow in only one direction, like a one-way valve (see [Model of Conduction in Metals](#)). A diode is created by joining a p-type semiconductor to an n-type semiconductor (Figure 14.4.1). The junction between these materials is called a **p-n junction**. A comparison of the energy bands of a silicon-based diode is shown in Figure 14.4.1b. The positions of the valence and conduction bands are the same, but the impurity levels are quite different. When a p-n junction is formed, electrons from the conduction band of the n-type material diffuse to the p-side, where they combine with holes in the valence band. This migration of charge leaves positive ionized donor ions on the n-side and negative ionized acceptor ions on the p-side, producing a narrow double layer of charge at the p-n junction called the **depletion layer**. The electric field associated with the depletion layer prevents further diffusion. The potential energy for electrons across the p-n junction is given by Figure 14.4.2.

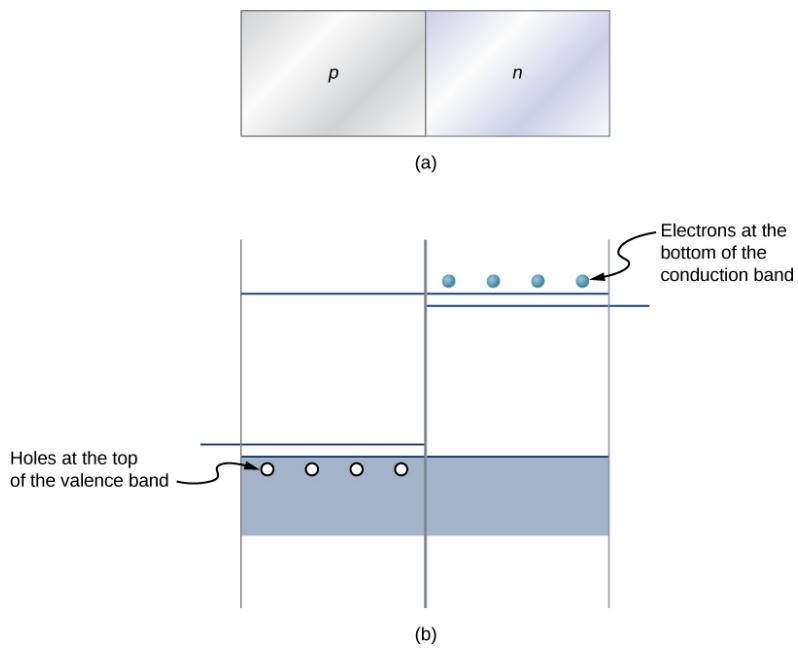


Figure 14.4.1: (a) Representation of a p-n junction. (b) A comparison of the energy bands of p-type and n-type silicon prior to equilibrium.

The behavior of a semiconductor diode can now be understood. If the positive side of the battery is connected to the n-type material, the depletion layer is widened, and the potential energy difference across the p-n junction is increased. Few or none of the

electrons (holes) have enough energy to climb the potential barrier, and current is significantly reduced. This is called the **reverse bias configuration**. On the other hand, if the positive side of a battery is connected to the **p**-type material, the depletion layer is narrowed, the potential energy difference across the **p-n** junction is reduced, and electrons (holes) flow easily. This is called the **forward bias configuration** of the diode. In sum, the diode allows current to flow freely in one direction but prevents current flow in the opposite direction. In this sense, the semiconductor diode is a one-way valve.

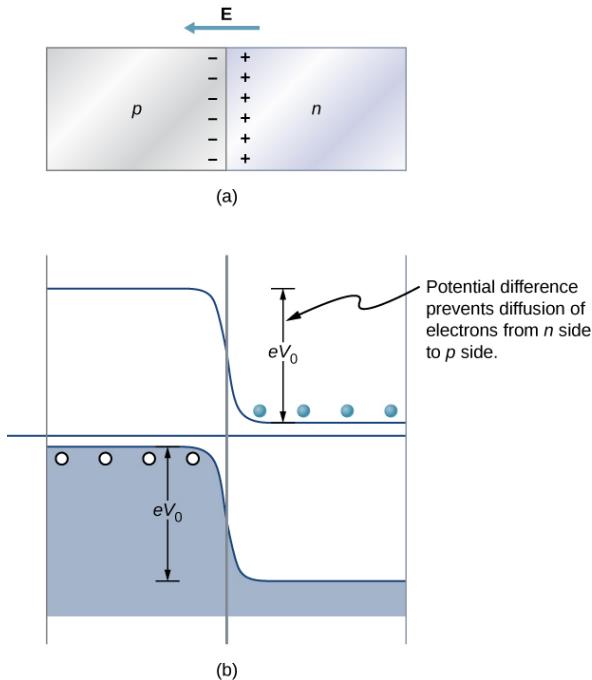


Figure 14.4.2: At equilibrium, (a) excess charge resides near the interface and the net current is zero, and (b) the potential energy difference for electrons (in light blue) prevents further diffusion of electrons into the **p**-side.

We can estimate the mathematical relationship between the current and voltage for a diode using the electric potential concept. Consider N negatively charged majority carriers (electrons donated by impurity atoms) in the **n**-type material and a potential barrier V across the **p-n** junction. According to the Maxwell-Boltzmann distribution, the fraction of electrons that have enough energy to diffuse across the potential barrier is $Ne^{-eV/k_B T}$. However, if a battery of voltage V_b is applied in the forward-bias configuration, this fraction improves to $Ne^{-e(V-V_b)/k_B T}$. The electric current due to the majority carriers from the **n**-side to the **p**-side is therefore

$$I = Ne^{-eV/k_B T} e^{eV_b/k_B T} = I_0 e^{eV_b/k_B T},$$

where I_0 is the current with no applied voltage and T is the temperature. Current due to the minority carriers (thermal excitation of electrons from the valence band to the conduction band on the **p**-side and subsequent attraction to the **n**-side) is $-I_0$, independent of the bias voltage. The net current is therefore

$$I_{\text{net}} = I_0 (e^{eV_b/k_B T} - 1).$$

A sample graph of the current versus bias voltage is given in Figure 14.4.3. In the forward bias configuration, small changes in the bias voltage lead to large changes in the current. In the reverse bias configuration, the current is $I_{\text{net}} \approx -I_0$. For extreme values of reverse bias, the atoms in the material are ionized which triggers an avalanche of current. This case occurs at the **breakdown voltage**.

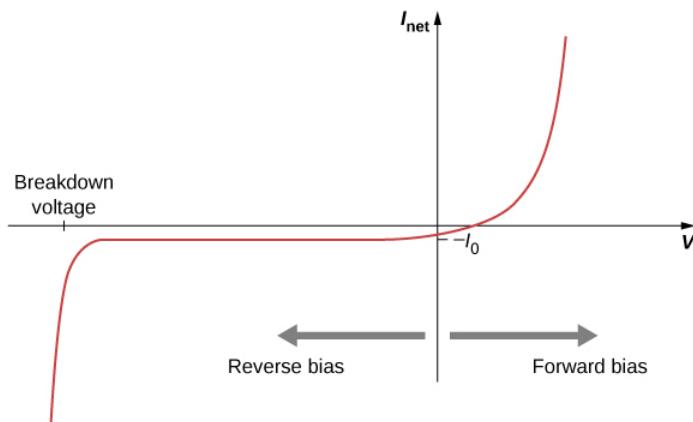


Figure 14.4.3: Current versus voltage across a **p-n** junction (diode). In the forward bias configuration, electric current flows easily. However, in the reverse bias configuration, electric current flow very little.

✓ Example 14.4.1: Diode Current

Attaching the positive end of a battery to the **p**-side and the negative end to the **n**-side of a semiconductor diode produces a current of $4.5 \times 10^{-1} \text{ A}$. The reverse saturation current is $2.2 \times 10^{-8} \text{ A}$. (The reverse saturation current is the current of a diode in a reverse bias configuration such as this.) The battery voltage is 0.12 V. What is the diode temperature?

Strategy

The first arrangement is a forward bias configuration, and the second is the reverse bias configuration.

Solution

The current in the forward and reverse bias configurations is given by

$$I_{\text{net}} = I_0 (e^{eV_b/k_B T} - 1).$$

The current with no bias is related to the reverse saturation current by

$$I_0 \approx -I_{\text{sat}} = 2.2 \times 10^{-8}.$$

Therefore

$$\frac{I_{\text{net}}}{I_0} = \frac{4.5 \times 10^{-1} \text{ A}}{2.2 \times 10^{-8} \text{ A}} = 2.0 \times 10^8.$$

this can be written as

$$\frac{I_{\text{net}}}{I_0} + 1 = e^{eV_b/k_B T}.$$

This ratio is much greater than one, so the second term on the left-hand side of the equation vanishes. Taking the natural log of both sides gives

$$\frac{eV_b}{k_B T} = 19.$$

The temperature is therefore

$$T = \frac{eV_b}{k_B} \left(\frac{1}{19} \right) = \frac{e(0.12 \text{ V})}{8.617 \times 10^{-5} \text{ eV/K}} \left(\frac{1}{19} \right) = 73 \text{ K}.$$

Significance

The current moving through a diode in the forward and reverse bias configuration is sensitive to the temperature of the diode. If the potential energy supplied by the battery is large compared to the thermal energy of the diode's surroundings, $k_B T$, then the forward bias current is very large compared to the reverse saturation current.

? Exercise 14.4.1

How does the magnitude of the forward bias current compare with the reverse bias current?

Solution

The forward bias current is much larger. To a good approximation, diodes permit current flow in only one direction.

Create a **p-n** junction and observe the behavior of a simple circuit for forward and reverse bias voltages. Visit this [site](#) to learn more about semiconductor diodes.

Junction Transistor

If diodes are one-way valves, transistors are one-way valves that can be carefully opened and closed to control current. A special kind of transistor is a junction transistor. A **junction transistor** has three parts, including an **n**-type semiconductor, also called the emitter; a thin **p**-type semiconductor, which is the base; and another **n**-type semiconductor, called the collector (Figure 14.4.4). When a positive terminal is connected to the **p**-type layer (the base), a small current of electrons, called the **base current I_B** , flows to the terminal. This causes a large **collector current I_C** to flow through the collector. The base current can be adjusted to control the large collector current. The current gain is therefore

$$I_c = \beta I_B.$$

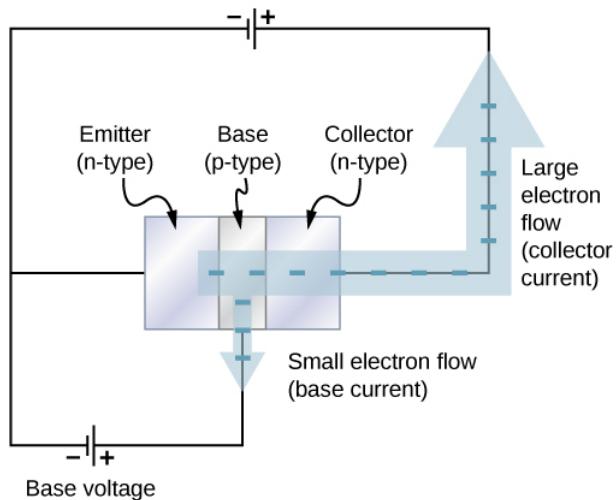


Figure 14.4.4: A junction transistor has three parts: emitter, base, and collector. Voltage applied to the base acts as a valve to control electric current from the emitter to the collector.

A junction transistor can be used to amplify the voltage from a microphone to drive a loudspeaker. In this application, sound waves cause a diaphragm inside the microphone to move in and out rapidly (Figure 14.4.5). When the diaphragm is in the “in” position, a tiny positive voltage is applied to the base of the transistor. This opens the transistor “valve” and allows a large electrical current flow to the loudspeaker. When the diaphragm is in the “out” position, a tiny negative voltage is applied to the base of the transistor, which shuts off the transistor valve so that no current flows to the loudspeaker. This shuts the transistor “valve” off so no current flows to the loudspeaker. In this way, current to the speaker is controlled by the sound waves, and the sound is amplified. Any electric device that amplifies a signal is called an **amplifier**.

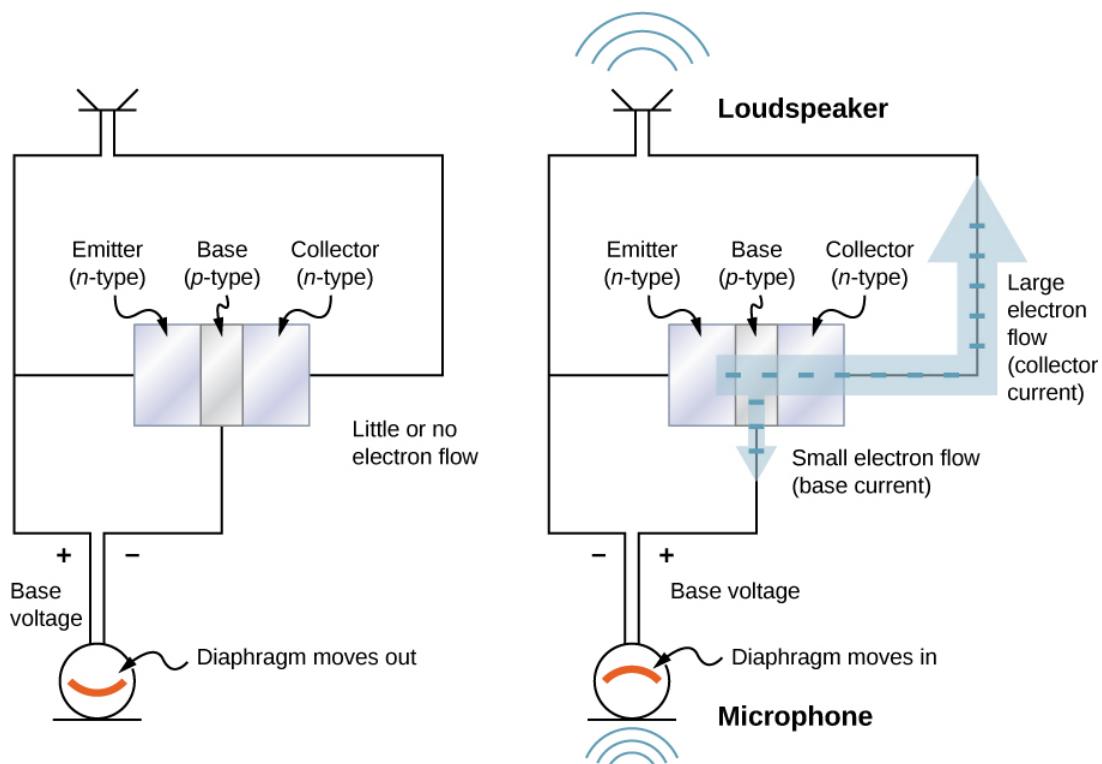


Figure 14.4.5: An audio amplifier based on a junction transistor. Voltage applied to the base by a microphone acts as a valve to control a larger electric current that passes through a loudspeaker.

In modern electronic devices, digital signals are used with diodes and transistors to perform tasks such as data manipulation. Electric circuits carry two types of electrical signals: analog and digital (Figure 14.4.6). An analog signal varies continuously, whereas a digital signal switches between two fixed voltage values, such as plus 1 volt and zero volts. In digital circuits like those found in computers, a transistor behaves like an on-off switch. The transistor is either on, meaning the valve is completely open, or it is off, meaning the valve is completely closed. Integrated circuits contain vast collections of transistors on a single piece of silicon. They are designed to handle digital signals that represent ones and zeroes, which is also known as binary code. The invention of the ic helped to launch the modern computer revolution.

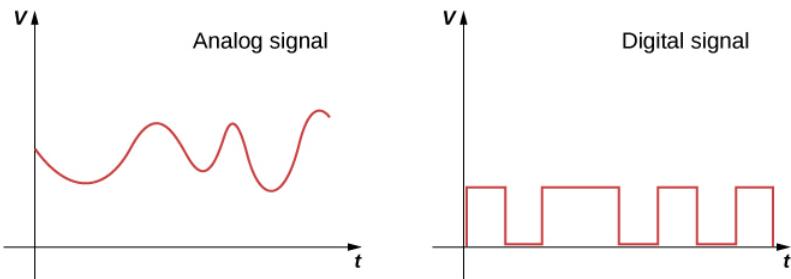


Figure 14.4.6: Real-world data are often analog, meaning data can vary continuously. Intensity values of sound or visual images are usually analog. These data are converted into digital signals for electronic processing in recording devices or computers. The digital signal is generated from the analog signal by requiring certain voltage cut-off value.

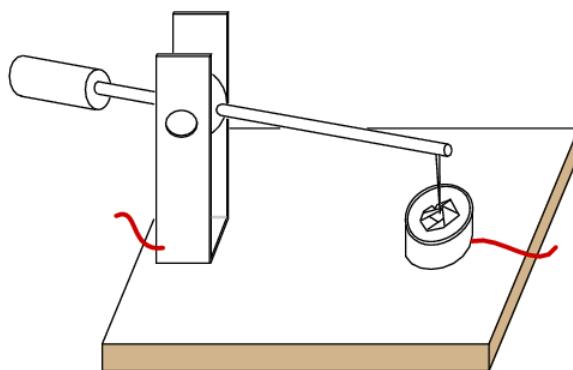
This page titled [14.4: Introduction to Semiconductor Devices](#) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [9.8: Semiconductor Devices](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-3>.

14.5: Junction Diodes

Selenium oxide rectifiers were used before modern power diode rectifiers became available. These and the Cu₂O rectifiers were polycrystalline devices. Photoelectric cells were once made from Selenium.

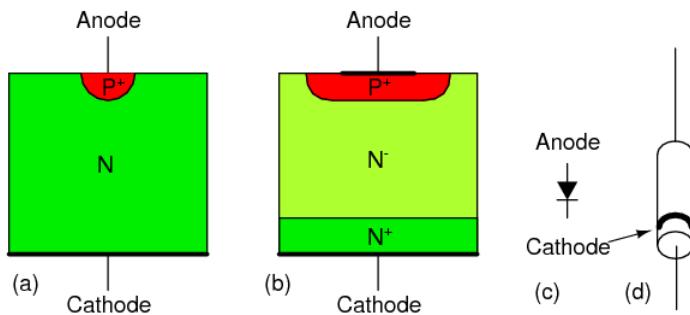
Before the modern semiconductor era, an early diode application was as a radio frequency *detector*, which recovered audio from a radio signal. The “semiconductor” was a polycrystalline piece of the mineral galena, lead sulfide, PbS. A pointed metallic wire known as a *cat whisker* was brought in contact with a spot on a crystal within the polycrystalline mineral. (Figure below) The operator labored to find a “sensitive” spot on the galena by moving the cat whisker about. Presumably, there were P and N-type spots randomly distributed throughout the crystal due to the variability of uncontrolled impurities. Less often the mineral iron pyrites, fools gold, was used, as was the mineral carborundum, silicon carbide, SiC, another detector, part of a Presumably there were P and N-type spots randomly distributed throughout the crystal due to the variability of uncontrolled impurities. Less often the mineral iron pyrites, fools gold, was used, as was the mineral carborundum, silicon carbide, SiC, another detector, part of a *foxhole radio*, consisted of a sharpened pencil lead bound to a bent safety pin, touching a rusty blue-blade disposable razor blade. These all required searching for a sensitive spot, easily lost because of vibration.



Crystal detector

Replacing the mineral with an N-doped semiconductor (Figure below(a)) makes the whole surface sensitive, so that searching for a sensitive spot was no longer required. This device was perfected by G.W.Pickard in 1906. The pointed metal contact produced a localized P-type region within the semiconductor. The metal point was fixed in place, and the whole *point contact diode* encapsulated in a cylindrical body for mechanical and electrical stability. (Figure below(d)) Note that the cathode bar on the schematic corresponds to the bar on the physical package.

Silicon point contact diodes made an important contribution to radar in World War II, detecting giga-hertz radio frequency echo signals in the radar receiver. The concept to be made clear is that the point contact diode preceded the junction diode and modern semiconductors by several decades. Even to this day, the point contact diode is a practical means of microwave frequency detection because of its low capacitance. Germanium point contact diodes were once more readily available than they are today, being preferred for the lower 0.2 V forward voltage in some applications like self-powered crystal radios. Point contact diodes, though sensitive to a wide bandwidth, have a low current capability compared with junction diodes.



Silicon diode cross-section: (a) point contact diode, (b) junction diode, (c) schematic symbol, (d) small signal diode package.

Most diodes today are silicon junction diodes. The cross-section in Figure above(b) looks a bit more complex than a simple PN junction; though, it is still a PN junction. Starting at the cathode connection, the N⁺ indicates this region is heavily doped, having nothing to do with polarity. This reduces the series resistance of the diode. The N⁻ region is lightly doped as indicated by the (-). Light doping produces a diode with a higher reverse breakdown voltage, important for high voltage power rectifier diodes. Lower voltage diodes, even low voltage power rectifiers, would have lower forward losses with heavier doping. The heaviest level of doping produces zener diodes designed for a low reverse breakdown voltage. However, heavy doping increases the reverse leakage current. The P⁺ region at the anode contact is heavily doped P-type semiconductor, a good contact strategy. Glass encapsulated small signal junction diodes are capable of 10's to 100's of mA of current. Plastic or ceramic encapsulated power rectifier diodes handle to 1000's of amperes of current.

Review

- Point contact diodes have superb high-frequency characteristics, usable well into the microwave frequencies.
- Junction diodes range in size from small signal diodes to power rectifiers capable of 1000's of amperes.
- The level of doping near the junction determines the reverse breakdown voltage. Light doping produces a high voltage diode. Heavy doping produces a lower breakdown voltage, and increases reverse leakage current. Zener diodes have a lower breakdown voltage because of heavy doping.

This page titled [14.5: Junction Diodes](#) is shared under a [GNU Free Documentation License 1.3](#) license and was authored, remixed, and/or curated by [Tony R. Kuphaldt \(All About Circuits\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [2.7: Junction Diodes](#) by [Tony R. Kuphaldt](#) is licensed [GNU FDL](#). Original source:
<https://www.allaboutcircuits.com/textbook/semiconductors>.

14.6: Light Emitting Diode

Let's talk about the recombining electrons for a minute. When the electron falls down from the conduction band and fills in a hole in the valence band, there is an obvious loss of energy. The question is; where does that energy go? In silicon, the answer is not very interesting. Silicon is what is known as an **indirect band-gap material**. What this means is that as an electron goes from the bottom of the conduction band to the top of the valence band, it must also undergo a significant change in momentum. This all comes about from the details of the band structure for the material, which we will not concern ourselves with here. As we all know, whenever something changes state, we must still conserve not only energy, but also momentum. In the case of an electron going from the conduction band to the valence band in silicon, both of these things can only be conserved if the transition also creates a quantized set of lattice vibrations, called **phonons**, or "heat". Phonons possess **both** energy and momentum, and their creation upon the recombination of an electron and hole allows for complete conservation of both energy and momentum. All of the energy which the electron gives up in going from the conduction band to the valence band (1.1 eV) ends up in phonons, which is another way of saying that the electron heats up the crystal.

In some other semiconductors, something else occurs. In a class of materials called **direct band-gap semiconductors**, the transition from conduction band to valence band involves essentially no change in momentum. Photons, it turns out, possess a fair amount of energy (several eV per photon in some cases) but they have very little momentum associated with them. Thus, for a direct band gap material, the excess energy of the electron-hole recombination can either be taken away as heat, or more likely, as a photon of light. This radiative transition then conserves energy and momentum by giving off light whenever an electron and hole recombine. This gives rise to (for us) a new type of device, the light emitting diode (LED). Emission of a photon in an LED is shown schematically in Figure 14.6.1.

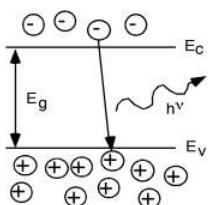


Figure 14.6.1: Radiative recombination in a direct band-gap semiconductor

It was Planck who postulated that the energy of a photon was related to its frequency by a constant, which was later named after him. If the frequency of oscillation is given by the Greek letter "nu" (ν), then the energy of the photon is just $h\nu$ where h is Planck's constant, which has a value of 4.14×10^{-15} eV · seconds.

$$E = h\nu$$

When we talk about light it is conventional to specify its wavelength, λ , instead of its frequency. Visible light has a wavelength on the order of nanometers (red is about 600 nm, green about 500 nm and blue is in the 450 nm region.) A handy "rule of thumb" can be derived from the fact that $\lambda = \frac{c}{\nu}$, where c is the speed of light. Since $c = 3 \times 10^8 \frac{\text{m}}{\text{sec}}$ or $c = 3 \times 10^{17} \frac{\text{nm}}{\text{sec}}$,

$$\begin{aligned}\lambda(\text{nm}) &= \frac{hc}{E(\text{eV})} \\ &= \frac{1242}{E(\text{eV})}\end{aligned}$$

Thus, a semiconductor with a **2 eV** band-gap should give off light at about **620 nm** (in the range of red light). A **3 eV** band-gap material would emit at **414 nm**, in the violet. The human eye, of course, is not equally responsive to all colors. We show this in Figure 14.6.2, where we have also included the materials which are used for important light emitting diodes (LEDs) for each of the different spectral regions.

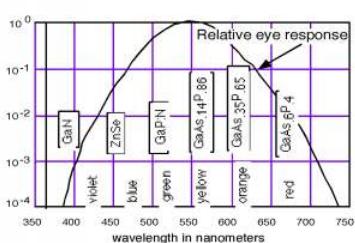


Figure 14.6.2: Relative response of the human eye to various colors

As you no doubt notice, a number of the important LEDs are based on the **GaAsP** system. **GaAs** is a direct band-gap semiconductor with a band gap of **1.42 eV** (in the infrared). **GaP** is an indirect band-gap material with a band gap of **2.26 eV** (550 nm, or green). Both **As** and **P** are group V elements. (Hence the nomenclature of the materials as **III-V compound semiconductors**.) We can replace some of the **As** with **P** in GaAs) and make a mixed compound semiconductor $\text{GaAs}_{1-x}\text{P}_x$. When the mole fraction of phosphorous is less than about **0.45** the band gap is direct, and so we can "engineer" the desired color of LED that we want by simply growing a crystal with the proper phosphorus concentration!

The properties of the GaAsP system are shown in Figure 14.6.3. It turns out that for this system, there are actually **two** different band gaps, as shown in the inset of Figure 14.6.3. One is a direct gap (no change in momentum) and the other is indirect. In **GaAs**, the direct gap has lower energy than the indirect one (like in the inset) and so the transition is a radiative one. As we start adding phosphorous to the system, both the direct and indirect band gaps increase in energy. However, the direct gap energy increases faster with phosphorous fraction than does the indirect one. At a mole fraction x of about **0.45**, the gap energies cross over and the material goes from being a direct gap semiconductor to an indirect gap semiconductor. At $x = 0.35$ the band gap is about **1.97 eV** (630 nm), and so we would only expect to get light up to the red using the **GaAsP** system for making LEDs. Fortunately, people discovered that you could add an impurity (nitrogen) to the **GaAsP** system, which introduced a new level in the system. An electron could go from the indirect conduction band (for a mixture with a mole fraction greater than **0.45**) to the nitrogen site, changing its momentum, but not its energy. It could then make a direct transition to the valence band, and light with colors all the way to the green became possible. The use of a nitrogen **recombination center** is depicted in Figure 14.6.4.

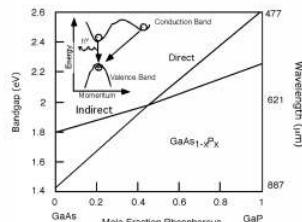


Figure 14.6.3: Band gap for the **GaAsP** system

Addition of a nitrogen recombination center to indirect **GaAs**

If we want colors with wavelengths shorter than the green, we must abandon the **GaAsP** system and look for more suitable materials. A compound semiconductor made from the II-VI elements **Zn** and **Se** make up one promising system, and several research groups have successfully made blue and blue-green LEDs from **ZnSe**. **SiC** is another (weak) blue emitter which is commercially available on the market. Recently, workers at a tiny, unknown chemical company stunned the "display world" by announcing that they had successfully fabricated a blue LED using the II-V material **GaN**. A good blue LED has been the "holy grail" of the display and CD-ROM research community for a number of years now. Obviously, adding blue to the already working green and red LEDs completes the set of 3 primary colors necessary for a full-color flat panel display (hang a TV screen on your wall like a picture?). Using a blue LED or laser in a CD-ROM would more than quadruple its data capacity, as bit diameter scales as λ , and hence the area as λ^2 .

This page titled [14.6: Light Emitting Diode](#) is shared under a [CC BY 1.0](#) license and was authored, remixed, and/or curated by [Bill Wilson](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [1.12: Light Emitting Diode](#) by [Bill Wilson](#) is licensed [CC BY 1.0](#). Original source: <https://cnx.org/contents/IE4zW5wX4.3:1sKbR9Vg22>.

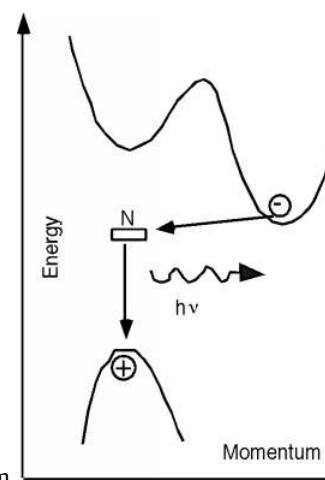


Figure 14.6.4:

14.7: Solar Cells

Now let us look at the opposite process of light generation for a moment. Consider the following situation.

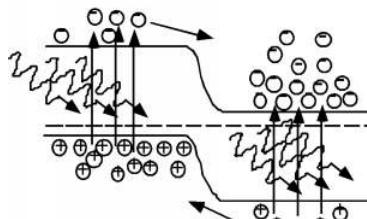


Figure 14.7.1: P-N diode under illumination

We have just a plain old normal p-n junction, only now, instead of applying an external voltage, we imagine that the junction is being illuminated with light whose photon energy is greater than the band-gap. In this situation, instead of recombination, we will get photo-generation of electron hole pairs. The photons simply excite electrons from the full states in the valence band, and "kick" them up into the conduction band, leaving a hole behind. (This is similar to the thermal excitation process we talked about earlier). As you can see from Figure 14.7.1, this creates excess electrons in the conduction band in the p-side of the diode, and excess holes in the valence band of the n-side. These carriers can diffuse over to the junction, where they will be swept across by the built-in electric field in the depletion region. If we were to connect the two sides of the diode together with a wire, a current would flow through that wire as a result of the electrons and holes which move across the junction.

Which way would the current flow? A quick look at Figure 14.7.1 shows that holes (positive charge carriers) are generated on the n-side and they float up to the p-side as they go across the junction. Hence positive current must be coming out of the anode, or p-side of the junction. Likewise, electrons generated on the p-side fall down the junction potential, and come out the n-side, but since they have negative charge, this flow represents current going **into** the cathode. We have constructed a **photovoltaic diode**, or **solar cell**! Figure 14.7.2 is a picture of what this would look like schematically.

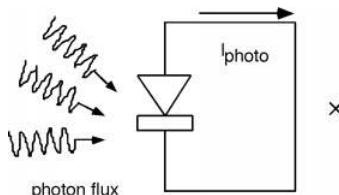


Figure 14.7.2: Schematic representation of a photovoltaic cell

We might like to consider the possibility of using this device as a source of energy, but the way we have things set up now, since the voltage across the diode is zero, and since power equals current times voltage, we see that we are getting nada from the cell. What we need, obviously, is a load resistor, so let's put one in. It should be clear from Figure 14.7.3 that the photo current flowing through the load resistor will develop a voltage which it biases the diode in the **forward** direction, which, of course will cause current to flow back into the anode. This complicates things, it seems we have current coming **out** of the diode and current going **into** the diode all at the same time! How are we going to figure out what is going on?

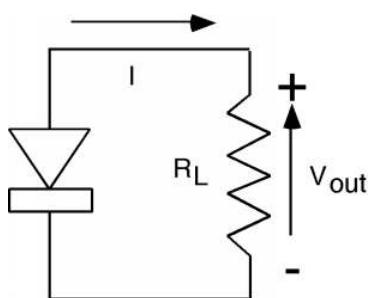


Figure 14.7.3: Photovoltaic cell with a load resistor

The answer is to make a model. The current which arises due to the photon flux can be conveniently represented as a current source. We can leave the diode as a diode, and we have the circuit shown in Figure 14.7.4.

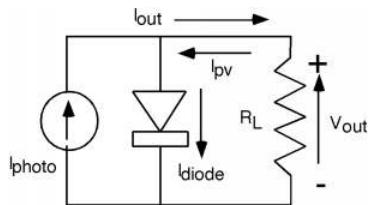


Figure 14.7.4: Model of PV cell

Even though we show I_{out} coming out of the device, we know by the usual polarity convention that when we define V_{out} as being positive at the top, then we should show the current for the photovoltaic, I_{pv} , as going into the diode from the top, which is what was done in Figure 14.7.4. Note that $I_{\text{pv}} = I_{\text{diode}} - I_{\text{photo}}$, so all we need to do is to subtract the two currents; we do this graphically in Figure 14.7.5. Note that we have numbered the four quadrants in the I - V plot of the total PV current. In quadrant I and III, the product of I and V is a positive number, meaning that power is being **dissipated** in the cell. For quadrant II and IV, the product of I and V is negative, and so we are getting power **from** the device. Clearly we want to operate in quadrant IV. In fact, without the addition of an external battery or current source, the circuit will **only** run in the IV'th quadrant. Consider adjusting R_L , the load resistor from 0 (a short) to ∞ (an open). With $R_L = 0$, we would be at point A on Figure 14.7.5. As R_L starts to increase from zero, the voltage across both the diode and the resistor will start to increase also, and we will move to point B, say. As R_L gets bigger and bigger, we keep moving along the curve until, at point C, where R_L is an open and we have the maximum voltage across the device, but, of course, no current coming out!

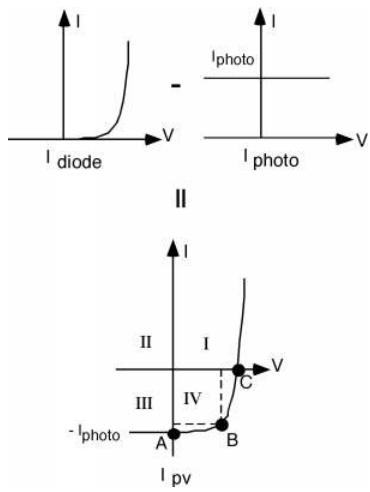


Figure 14.7.5: Combining the diode and the current source

Power is $V \cdot I$ so at B, for instance, the power coming out would be represented by the area enclosed by the two dotted lines and the coordinate axes. Someplace about where I have point B would be where we would be getting the most power out of our solar cell.

Figure 14.7.6 shows you what a real solar cell would look like. They are usually made from a complete wafer of silicon, to maximize the usable area. A shallow ($0.25 \mu\text{m}$) junction is made on the top, and top contacts are applied as stripes of metal conductor as shown. An anti-reflection (AR) coating is applied on top of that, which accounts for the bluish color which a typical solar cell has.

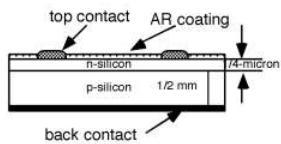
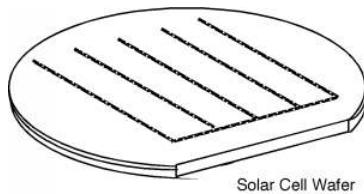


Figure 14.7.6: A real solar cell

The solar power flux on the earth's surface is (conveniently) about $1 \frac{\text{kW}}{\text{m}^2}$, or $100 \frac{\text{mW}}{\text{cm}^2}$. So if we made a solar cell from a 4-inch-diameter wafer (typical) it would have an area of about 81 cm^2 and so would be receiving a flux of about **8.1 Watts**. Typical cell efficiencies run from about **10 %** to maybe **15 %** unless special (and costly) tricks are used. This means that we will get about **1.2 Watts** out of a single wafer. Looking at point B on Figure 14.7.5 we could guess that V_{out} will be about 0.5 to 0.6 volts, thus we could expect to get maybe around 2.5 amps from a 4-inch wafer at 0.5 volts with **15 %** efficiency under the illumination of one sun.

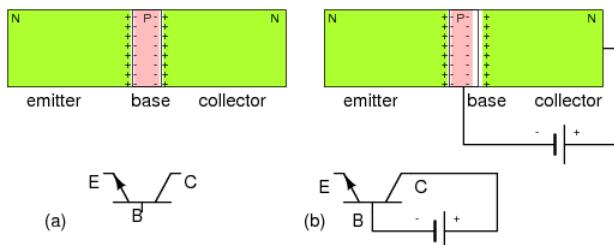
This page titled [14.7: Solar Cells](#) is shared under a [CC BY 1.0](#) license and was authored, remixed, and/or curated by [Bill Wilson](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [1.14: Solar Cells](#) by [Bill Wilson](#) is licensed CC BY 1.0. Original source: <https://cnx.org/contents/IE4zW5wX4.3:1sKbR9Vg22>.

14.8: Bipolar Junction Transistors

The bipolar junction transistor shown in Figure below(a) is an NPN three layer semiconductor sandwich with an *emitter* and *collector* at the ends, and a *base* in between. It is as if a third layer were added to a two layer diode. If this were the only requirement, we would have no more than a pair of back-to-back diodes. In fact, it is far easier to build a pair of back-to-back diodes. The key to the fabrication of a bipolar junction transistor is to make the middle layer, the base, as thin as possible without shorting the outside layers, the emitter, and collector. We cannot over emphasize the importance of the thin base region.

The device in Figure below(a) has a pair of junctions, emitter to base and base to collector, and two depletion regions.



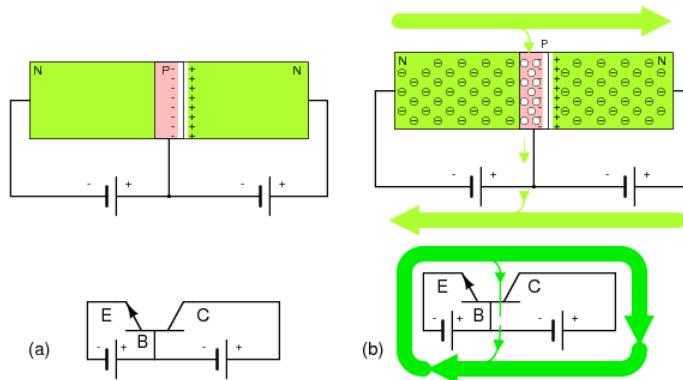
(a) NPN junction bipolar transistor. (b) Apply reverse bias to collector-base junction.

It is customary to reverse bias the base-collector junction of a bipolar junction transistor as shown in (Figure above(b)). Note that this increases the width of the depletion region. The reverse bias voltage could be a few volts to tens of volts for most transistors. There is no current flow, except leakage current, in the collector circuit.

In Figure below(a), a voltage source has been added to the emitter base circuit. Normally we forward bias the emitter-base junction, overcoming the 0.6 V potential barrier. This is similar to forward biasing a junction diode. This voltage source needs to exceed 0.6 V for majority carriers (electrons for NPN) to flow from the emitter into the base becoming minority carriers in the P-type semiconductor.

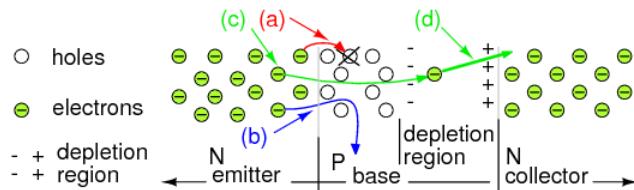
If the base region were thick, as in a pair of back-to-back diodes, all the current entering the base would flow out the base lead. In our NPN transistor example, electrons leaving the emitter for the base would combine with holes in the base, making room for more holes to be created at the (+) battery terminal on the base as electrons exit.

However, the base is manufactured thin. A few majority carriers in the emitter, injected as minority carriers into the base, actually recombine. See Figure below(b). Few electrons injected by the emitter into the base of an NPN transistor fall into holes. Also, few electrons entering the base flow directly through the base to the positive battery terminal. Most of the emitter current of electrons diffuses through the thin base into the collector. Moreover, modulating the small base current produces a larger change in collector current. If the base voltage falls below approximately 0.6 V for a silicon transistor, the large emitter-collector current ceases to flow.



NPN junction bipolar transistor with reverse biased collector-base: (a) Adding forward bias to base-emitter junction, results in (b) a small base current and large emitter and collector currents.

In Figure below we take a closer look at the current amplification mechanism. We have an enlarged view of an NPN junction transistor with emphasis on the thin base region. Though not shown, we assume that external voltage sources 1) forward bias the emitter-base junction, 2) reverse bias the base-collector junction. Electrons, majority carriers, enter the emitter from the (-) battery terminal. The base current flow corresponds to electrons leaving the base terminal for the (+) battery terminal. This is but a small current compared to the emitter current.



Disposition of electrons entering base: (a) Lost due to recombination with base holes. (b) Flows out base lead. (c) Most diffuse from emitter through thin base into base-collector depletion region, and (d) are rapidly swept by the strong depletion region electric field into the collector.

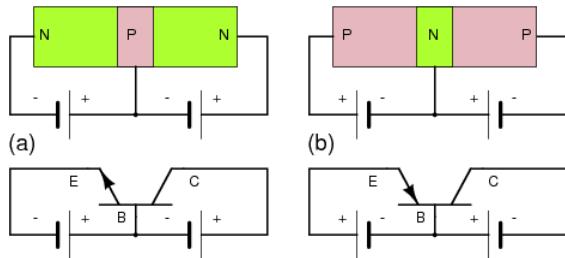
Majority carriers within the N-type emitter are electrons, becoming minority carriers when entering the P-type base. These electrons face four possible fates entering the thin P-type base. A few at Figure above(a) fall into holes in the base that contribute to base current flow to the (+) battery terminal. Not shown, holes in the base may diffuse into the emitter and combine with electrons, contributing to base terminal current. Few at (b) flow on through the base to the (+) battery terminal as if the base were a resistor. Both (a) and (b) contribute to the very small base current flow. Base current is typically 1% of emitter or collector current for small signal transistors. Most of the emitter electrons diffuse right through the thin base (c) into the base-collector depletion region. Note the polarity of the depletion region surrounding the electron at (d). The strong electric field sweeps the electron rapidly into the collector. The strength of the field is proportional to the collector battery voltage. Thus 99% of the emitter current flows into the collector. It is controlled by the base current, which is 1% of the emitter current. This is a potential current gain of 99, the ratio of I_C/I_B , also known as beta, β .

This magic, the diffusion of 99% of the emitter carriers through the base, is only possible if the base is very thin. What would be the fate of the base minority carriers in a base 100 times thicker? One would expect the recombination rate, electrons falling into holes, to be much higher. Perhaps 99%, instead of 1%, would fall into holes, never getting to the collector. The second point to make is that the base current may control 99% of the emitter current, only if 99% of the emitter current diffuses into the collector. If it all flows out the base, no control is possible.

Another feature accounting for passing 99% of the electrons from emitter to collector is that real bipolar junction transistors use a small heavily doped emitter. The high concentration of emitter electrons forces many electrons to diffuse into the base. The lower doping concentration in the base means fewer holes diffuse into the emitter, which would increase the base current. Diffusion of carriers from emitter to base is strongly favored.

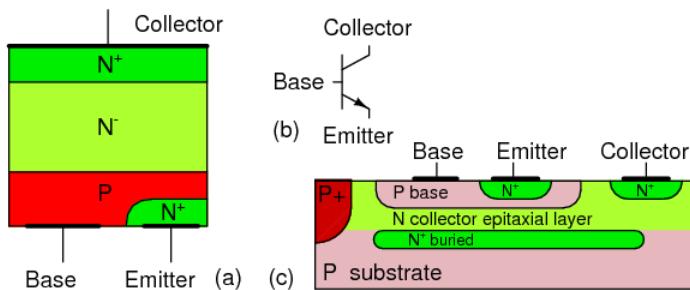
The thin base and the heavily doped emitter help keep the *emitter efficiency* high, 99% for example. This corresponds to 100% emitter current splitting between the base as 1% and the collector as 99%. The emitter efficiency is known as $\alpha = I_C/I_E$.

Bipolar junction transistors are available as PNP as well as NPN devices. We present a comparison of these two in Figure below. The difference is the polarity of the base-emitter diode junctions, as signified by the direction of the schematic symbol emitter arrow. It points in the same direction as the anode arrow for a junction diode, against electron current flow. See diode junction, Figure previous. The point of the arrow and bar correspond to P-type and N-type semiconductors, respectively. For NPN and PNP emitters, the arrow points away and toward the base respectively. There is no schematic arrow on the collector. However, the base-collector junction is the same polarity as the base-emitter junction compared to a diode. Note, we speak of diode, not power supply, polarity.



Compare NPN transistor at (a) with the PNP transistor at (b). Note direction of emitter arrow and supply polarity.

The voltage sources for PNP transistors are reversed compared with an NPN transistors as shown in Figure above. The base-emitter junction must be forward biased in both cases. The base on a PNP transistor is biased negative (b) compared with positive (a) for an NPN. In both cases the base-collector junction is reverse biased. The PNP collector power supply is negative compared with positive for an NPN transistor.



Bipolar junction transistor: (a) discrete device cross-section, (b) schematic symbol, (c) integrated circuit cross-section.

Note that the BJT in Figure above(a) has heavy doping in the emitter as indicated by the N^+ notation. The base has a normal P-dopant level. The base is much thinner than the not-to-scale cross-section shows. The collector is lightly doped as indicated by the N^- notation. The collector needs to be lightly doped so that the collector-base junction will have a high breakdown voltage. This translates into a high allowable collector power supply voltage. Small signal silicon transistors have a 60-80 V breakdown voltage. Though, it may run to hundreds of volts for high voltage transistors. The collector also needs to be heavily doped to minimize ohmic losses if the transistor must handle high current. These contradicting requirements are met by doping the collector more heavily at the metallic contact area. The collector near the base is lightly doped as compared with the emitter. The heavy doping in the emitter gives the emitter-base a low approximate 7 V breakdown voltage in small signal transistors. The heavily doped emitter makes the emitter-base junction have zener diode like characteristics in reverse bias.

The BJT *die*, a piece of a sliced and diced semiconductor wafer, is mounted collector down to a metal case for power transistors. That is, the metal case is electrically connected to the collector. A small signal die may be encapsulated in epoxy. In power transistors, aluminum bonding wires connect the base and emitter to package leads. Small signal transistor dies may be mounted directly to the lead wires. Multiple transistors may be fabricated on a single die called an *integrated circuit*. Even the collector may be bonded out to a lead instead of the case. The integrated circuit may contain internal wiring of the transistors and other integrated components. The integrated BJT shown in (Figure (c) above) is much thinner than the “not to scale” drawing. The P^+ region isolates multiple transistors in a single die. An aluminum metallization layer (not shown) interconnects multiple transistors and other components. The emitter region is heavily doped, N^+ compared to the base and collector to improve emitter efficiency.

Discrete PNP transistors are almost as high quality as the NPN counterpart. However, integrated PNP transistors are not nearly as good as the NPN variety within the same integrated circuit die. Thus, integrated circuits use the NPN variety as much as possible.

Review

- Bipolar transistors conduct current using both electrons and holes in the same device.
- Operation of a bipolar transistor as a current amplifier requires that the collector-base junction be reverse biased and the emitter-base junction be forward biased.

- A transistor differs from a pair of back to back diodes in that the base, the center layer, is very thin. This allows majority carriers from the emitter to diffuse as minority carriers through the base into the depletion region of the base-collector junction, where the strong electric field collects them.
- Emitter efficiency is improved by heavier doping compared with the collector. Emitter efficiency: $\alpha = I_C/I_E$, 0.99 for small signal devices
- Current gain is $\beta=I_C/I_B$, 100 to 300 for small signal transistors.

This page titled [14.8: Bipolar Junction Transistors](#) is shared under a [GNU Free Documentation License 1.3](#) license and was authored, remixed, and/or curated by [Tony R. Kuphaldt \(All About Circuits\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

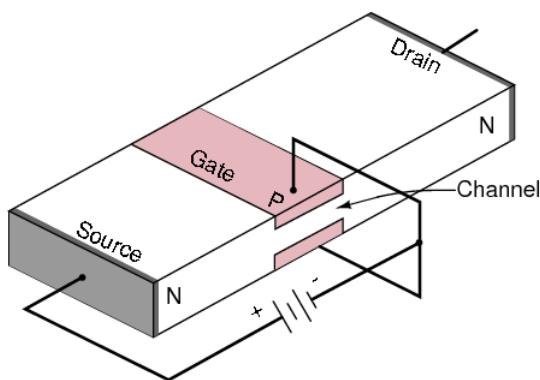
- [2.8: Bipolar Junction Transistors](#) by [Tony R. Kuphaldt](#) is licensed [GNU FDL](#). Original source:
<https://www.allaboutcircuits.com/textbook/semiconductors>.

14.9: Junction Field-effect Transistors

A *field effect transistor* (FET) is a *unipolar* device, conducting a current using only one kind of charge carrier. If based on an N-type slab of semiconductor, the carriers are electrons. Conversely, a P-type based device uses only holes.

At the circuit level, field effect transistor operation is simple. A voltage applied to the *gate*, input element, controls the resistance of the *channel*, the unipolar region between the gate regions. (Figure below) In an N-channel device, this is a lightly doped N-type slab of silicon with terminals at the ends. The *source* and *drain* terminals are analogous to the emitter and collector, respectively, of a BJT. In an N-channel device, a heavy P-type region on both sides of the center of the slab serves as a control electrode, the *gate*. The gate is analogous to the base of a BJT.

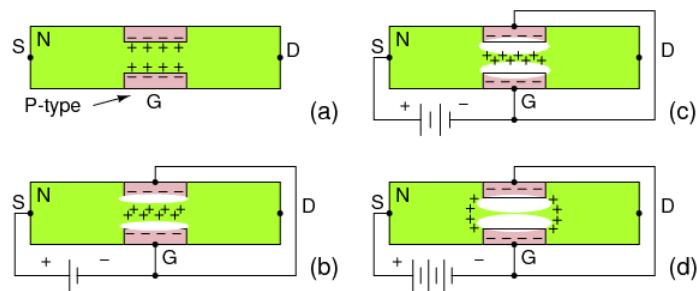
“Cleanliness is next to godliness” applies to the manufacture of field effect transistors. Though it is possible to make bipolar transistors outside of a *clean room*, it is a necessity for field effect transistors. Even in such an environment, manufacture is tricky because of contamination control issues. The unipolar field effect transistor is conceptually simple, but difficult to manufacture. Most transistors today are a metal oxide semiconductor variety (later section) of the field effect transistor contained within integrated circuits. However, discrete JFET devices are available.



Junction field effect transistor cross-section.

A properly biased N-channel junction field effect transistor (JFET) is shown in Figure above. The gate constitutes a diode junction to the source to drain semiconductor slab. The gate is reverse biased. If a voltage (or an ohmmeter) were applied between the source and drain, the N-type bar would conduct in either direction because of the doping. Neither gate nor gate bias is required for conduction. If a gate junction is formed as shown, conduction can be controlled by the degree of reverse bias.

Figure below(a) shows the depletion region at the gate junction. This is due to diffusion of holes from the P-type gate region into the N-type channel, giving the charge separation about the junction, with a non-conductive depletion region at the junction. The depletion region extends more deeply into the channel side due to the heavy gate doping and light channel doping.

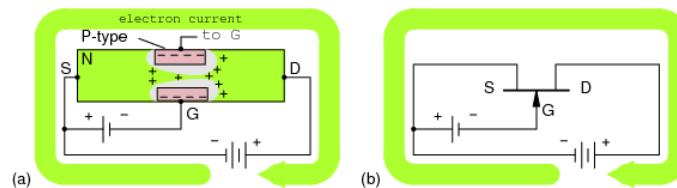


N-channel JFET: (a) Depletion at gate diode. (b) Reverse biased gate diode increases depletion region. (c) Increasing reverse bias enlarges depletion region. (d) Increasing reverse bias pinches-off the S-D channel.

The thickness of the depletion region can be increased Figure above(b) by applying moderate reverse bias. This increases the resistance of the source to drain channel by narrowing the channel. Increasing the reverse bias at (c) increases the depletion region, decreases the channel width, and increases the channel resistance. Increasing the reverse bias V_{GS} at (d) will *pinch-off* the channel

current. The channel resistance will be very high. This V_{GS} at which pinch-off occurs is V_P , the pinch-off voltage. It is typically a few volts. In summation, the channel resistance can be controlled by the degree of reverse biasing on the gate.

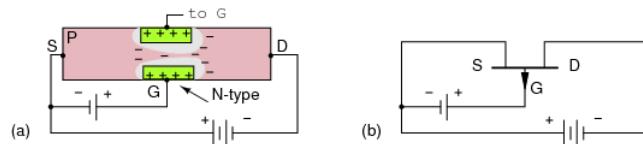
The source and drain are interchangeable, and the source to drain current may flow in either direction for low level drain battery voltage (< 0.6 V). That is, the drain battery may be replaced by a low voltage AC source. For a high drain power supply voltage, to 10's of volts for small signal devices, the polarity must be as indicated in Figure below(a). This drain power supply, not shown in previous figures, distorts the depletion region, enlarging it on the drain side of the gate. This is a more correct representation for common DC drain supply voltages, from a few to tens of volts. As drain voltage V_{DS} increased, the gate depletion region expands toward the drain. This increases the length of the narrow channel, increasing its resistance a little. We say “a little” because large resistance changes are due to changing gate bias. Figure below(b) shows the schematic symbol for an N-channel field effect transistor compared to the silicon cross-section at (a). The gate arrow points in the same direction as a junction diode. The “pointing” arrow and “non-pointing” bar correspond to P and N-type semiconductors, respectively.



N-channel JFET electron current flow from source to drain in (a) cross-section, (b) schematic symbol.

Figure above shows a large electron current flow from (-) battery terminal, to FET source, out the drain, returning to the (+) battery terminal. This current flow may be controlled by varying the gate voltage. A load in series with the battery sees an amplified version of the changing gate voltage.

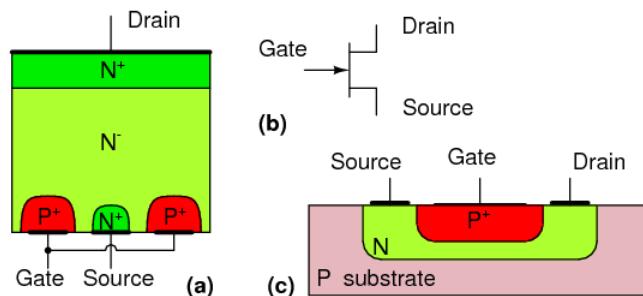
P-channel field effect transistors are also available. The channel is made of P-type material. The gate is a heavily doped N-type region. All the voltage sources are reversed in the P-channel circuit (Figure below) as compared with the more popular N-channel device. Also note, the arrow points out of the gate of the schematic symbol (b) of the P-channel field effect transistor.



P-channel JFET: (a) N-type gate, P-type channel, reversed voltage sources compared with N-channel device. (b) Note reversed gate arrow and voltage sources on schematic.

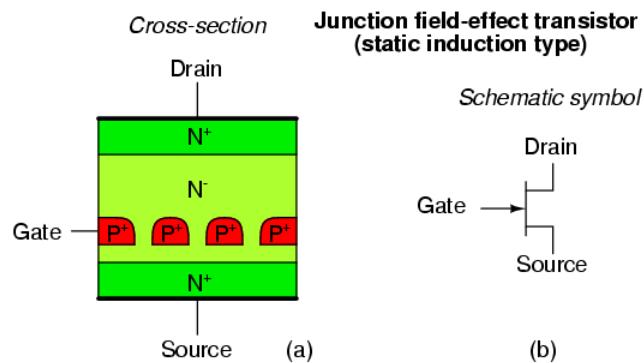
As the positive gate bias voltage is increased, the resistance of the P-channel increases, decreasing the current flow in the drain circuit.

Discrete devices are manufactured with the cross-section shown in Figure below. The cross-section, oriented so that it corresponds to the schematic symbol, is upside down with respect to a semiconductor wafer. That is, the gate connections are on the top of the wafer. The gate is heavily doped, P^+ , to diffuse holes well into the channel for a large depletion region. The source and drain connections in this N-channel device are heavily doped, N^+ to lower connection resistance. However, the channel surrounding the gate is lightly doped to allow holes from the gate to diffuse deeply into the channel. That is the N^- region.



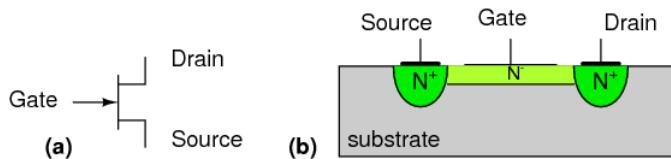
Junction field effect transistor: (a) Discrete device cross-section, (b) schematic symbol, (c) integrated circuit device cross-section.

All three FET terminals are available on the top of the die for the integrated circuit version so that a metalization layer (not shown) can interconnect multiple components. (Figure above(c)) Integrated circuit FET's are used in analog circuits for the high gate input resistance.. The N-channel region under the gate must be very thin so that the intrinsic region about the gate can control and pinch-off the channel. Thus, gate regions on both sides of the channel are not necessary.



Junction field effect transistor (static induction type): (a) Cross-section, (b) schematic symbol.

The static induction field effect transistor (SIT) is a short channel device with a buried gate. (Figure above) It is a power device, as opposed to a small signal device. The low gate resistance and low gate to source capacitance make for a fast switching device. The SIT is capable of hundreds of amps and thousands of volts. And, is said to be capable of an incredible frequency of 10 gHz.



Metal semiconductor field effect transistor (MESFET): (a) schematic symbol, (b) cross-section.

The *Metal semiconductor field effect transistor (MESFET)* is similar to a JFET except the gate is a schottky diode instead of a junction diode. A *schottky diode* is a metal rectifying contact to a semiconductor compared with a more common ohmic contact. In Figure above the source and drain are heavily doped (N^+). The channel is lightly doped (N^-). MESFET's are higher speed than JFET's. The MESET is a depletion mode device, normally on, like a JFET. They are used as microwave power amplifiers to 30 gHz. MESFET's can be fabricated from silicon, gallium arsenide, indium phosphide, silicon carbide, and the diamond allotrope of carbon.

Review

- The unipolar junction field effect transistor (FET or JFET) is so called because conduction in the channel is due to one type of carrier
- The JFET source, gate, and drain correspond to the BJT's emitter, base, and collector, respectively.
- Application of reverse bias to the gate varies the channel resistance by expanding the gate diode depletion region.

This page titled [14.9: Junction Field-effect Transistors](#) is shared under a [GNU Free Documentation License 1.3](#) license and was authored, remixed, and/or curated by [Tony R. Kuphaldt \(All About Circuits\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [2.9: Junction Field-effect Transistors](#) by [Tony R. Kuphaldt](#) is licensed [GNU FDL](#). Original source: <https://www.allaboutcircuits.com/textbook/semiconductors>.

CHAPTER OVERVIEW

15: Part 2 - Detailed and/or Advanced Content

15: Part 2 - Detailed and/or Advanced Content is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

16: Direct Calculation of Electrical Quantities from Charge Distributions

[16.1: Introduction](#)

[16.2: Electric Dipoles](#)

[16.3: Calculating Electric Fields of Charge Distributions](#)

[16.4: Calculating Electric Potential of Charge Distributions](#)

[16.5: Direct Calculation of Electrical Quantities from Charge Distributions \(Summary\)](#)

[16.6: Direct Calculation of Electrical Quantities from Charge Distributions \(Exercises\)](#)

[16.7: Direct Calculation of Electrical Quantities from Charge Distributions \(Answers\)](#)

16: Direct Calculation of Electrical Quantities from Charge Distributions is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

16.1: Introduction

This chapter covers some the more detailed calculations of electric quantities for various distributions of charge:

- Section 2 provides results related to the electric dipole.
- Section 3 describes how to calculate the electric field for a variety of continuous charge distributions using a method of direct integration.
- Section 4 describes how to calculate the electric potential for a variety of continuous charge distributions using a method of direct integration.

The method of direct integration can be applied to any distribution of charge but the integrals can only sometimes be evaluated analytically to give a function. If they cannot be calculated analytically, then they have to be evaluated numerically. Methods for numerical integration are beyond the scope of this text.

In some cases of high symmetry, it can be easier to calculate the electric field of some continuous distributions using Gauss's Law. This approach, when it can be used, can be quite elegant. This approach is the subject of the subsequent chapter [Gauss's Law for Calculation of Electrical Field from Charge Distributions](#).

16.1: Introduction is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Ronald Kumon.

16.2: Electric Dipoles

Learning Objectives

By the end of this section, you will be able to:

- Define an electric dipole.
- Distinguish a permanent dipole from an induced dipole.
- Define and calculate an electric dipole moment.
- Explain the physical meaning of the dipole moment.
- Calculate the torque on a dipole in a uniform electric field.
- Define and calculate the electric potential of a dipole.

Earlier we discussed, and calculated, the electric field of a dipole: two equal and opposite charges that are “close” to each other. (In this context, “close” means that the distance d between the two charges is much, much less than the distance of the field point P , the location where you are calculating the field.) Let’s now consider what happens to a dipole when it is placed in an external field \vec{E} . We assume that the dipole is a **permanent dipole**; it exists without the field, and does not break apart in the external field.

Rotation of a Dipole due to an Electric Field

For now, we deal with only the simplest case: The external field is uniform in space. Suppose we have the situation depicted in Figure 16.2.1, where we denote the distance between the charges as the vector \vec{d} , pointing from the negative charge to the positive charge.

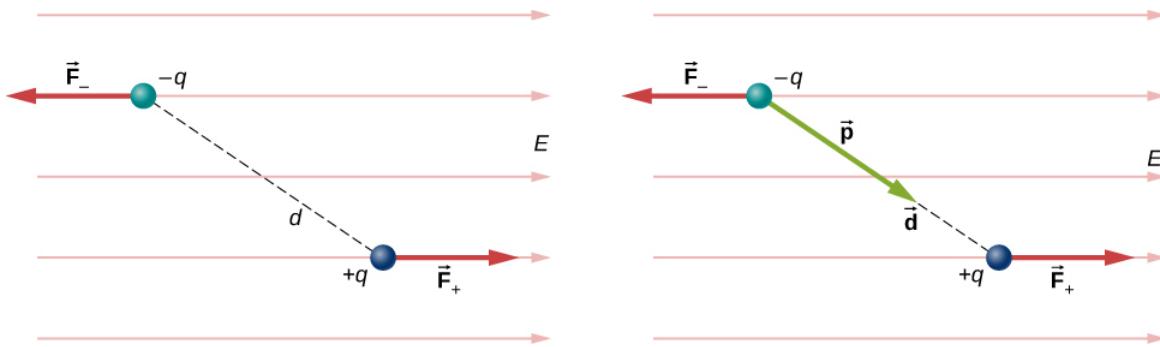


Figure 16.2.1: A dipole in an external electric field. (a) The net force on the dipole is zero, but the net torque is not. As a result, the dipole rotates, becoming aligned with the external field. (b) The dipole moment is a convenient way to characterize this effect. The \vec{d} points in the same direction as \vec{p} .

The forces on the two charges are equal and opposite, so there is no net force on the dipole. However, there is a torque:

$$\vec{r} = \left(\frac{\vec{d}}{2} \times \vec{F}_+ \right) + \left(-\frac{\vec{d}}{2} \times \vec{F}_- \right) \quad (16.2.1)$$

$$= \left[\left(\frac{\vec{d}}{2} \right) \times (+q\vec{E}) + \left(-\frac{\vec{d}}{2} \right) \times (-q\vec{E}) \right] \quad (16.2.2)$$

$$= q\vec{d} \times \vec{E}. \quad (16.2.3)$$

The quantity qd (the magnitude of each charge multiplied by the vector distance between them) is a property of the dipole; its value, as you can see, determines the torque that the dipole experiences in the external field. It is useful, therefore, to define this product as the so-called **dipole moment** of the dipole:

$$\vec{p} \equiv q\vec{d}. \quad (16.2.4)$$

We can therefore write

$$\vec{\tau} = \vec{p} \times \vec{E}. \quad (16.2.5)$$

Recall that a torque changes the angular velocity of an object, the dipole, in this case. In this situation, the effect is to rotate the dipole (that is, align the direction of \vec{p}) so that it is parallel to the direction of the external field.

Induced Dipoles

Neutral atoms are, by definition, electrically neutral; they have equal amounts of positive and negative charge. Furthermore, since they are spherically symmetrical, they do not have a “built-in” dipole moment the way most asymmetrical molecules do. They obtain one, however, when placed in an external electric field, because the external field causes oppositely directed forces on the positive nucleus of the atom versus the negative electrons that surround the nucleus. The result is a new charge distribution of the atom, and therefore, an **induced dipole** moment (Figure 16.2.2).

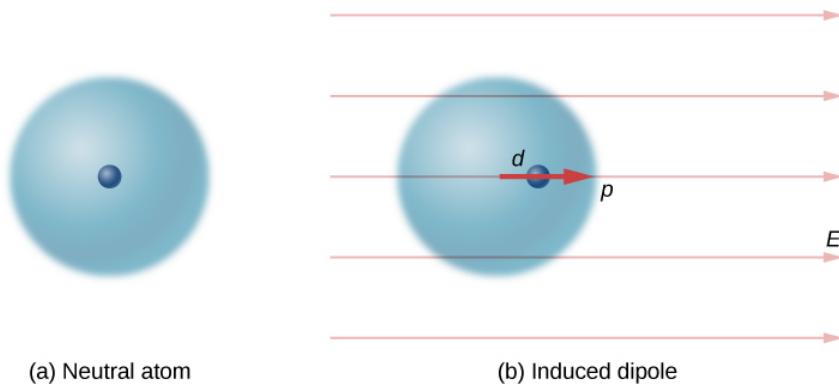


Figure 16.2.2: A dipole is induced in a neutral atom by an external electric field. The induced dipole moment is aligned with the external field.

An important fact here is that, just as for a rotated polar molecule, the result is that the dipole moment ends up aligned parallel to the external electric field. Generally, the magnitude of an induced dipole is much smaller than that of an inherent dipole. For both kinds of dipoles, notice that once the alignment of the dipole (rotated or induced) is complete, the net effect is to decrease the total electric field

$$\vec{E}_{\text{total}} = \vec{E}_{\text{external}} + \vec{E}_{\text{dipole}} \quad (16.2.6)$$

in the regions outside the dipole charges (Figure 16.2.3). By “outside” we mean further from the charges than they are from each other. This effect is crucial for capacitors, as you will see in [Capacitance](#).

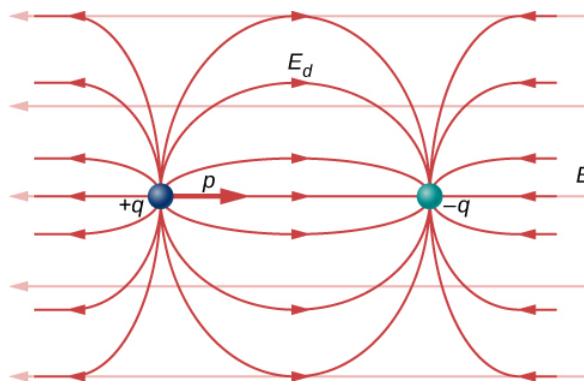


Figure 16.2.3: The net electric field is the vector sum of the field of the dipole plus the external field.

Recall that we found the [electric field of a dipole](#). If we rewrite it in terms of the dipole moment we get:

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0} \frac{\vec{p}}{z^3}. \quad (16.2.7)$$

The form of this field is shown in Figure 16.2.3. Notice that along the plane perpendicular to the axis of the dipole and midway between the charges, the direction of the electric field is opposite that of the dipole and gets weaker the further from the axis one goes. Similarly, on the axis of the dipole (but outside it), the field points in the same direction as the dipole, again getting weaker the further one gets from the charges.

Electric Potential of Dipole

An **electric dipole** is a system of two equal but opposite charges a fixed distance apart. This system is used to model many real-world systems, including atomic and molecular interactions. One of these systems is the water molecule, under certain circumstances. These circumstances are met inside a microwave oven, where electric fields with alternating directions make the water molecules change orientation. This vibration is the same as heat at the molecular level.

✓ Example 16.2.3: Electric Potential of a Dipole

Consider the dipole in Figure 16.2.3 with the charge magnitude of $q = 3.0 \mu\text{C}$ and separation distance $d = 4.0 \text{ cm}$. What is the potential at the following locations in space? (a) $(0, 0, 1.0 \text{ cm})$; (b) $(0, 0, -5.0 \text{ cm})$; (c) $(3.0 \text{ cm}, 0, 2.0 \text{ cm})$.

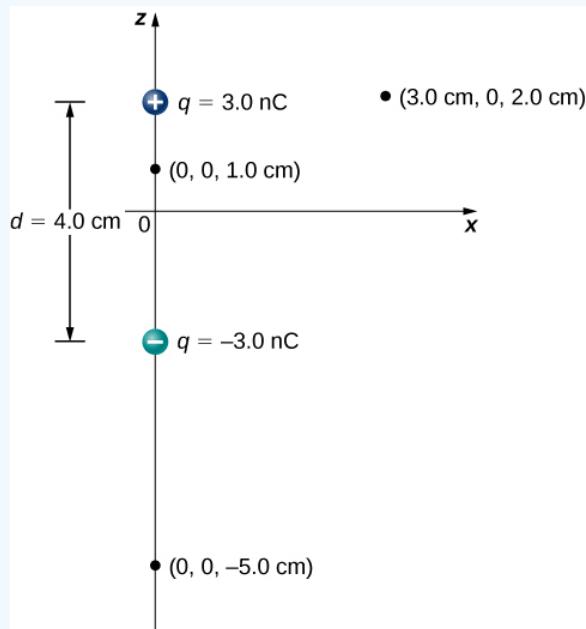


Figure 16.2.3: A general diagram of an electric dipole, and the notation for the distances from the individual charges to a point P in space.

Strategy

Apply $V_p = k \sum_1^N \frac{q_i}{r_i}$ to each of these three points.

Solution

$$\text{a. } V_p = k \sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2) \left(\frac{3.0 \text{ nC}}{0.010 \text{ m}} - \frac{3.0 \text{ nC}}{0.030 \text{ m}} \right) = 1.8 \times 10^3 \text{ V}$$

$$\text{b. } V_p = k \sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2) \left(\frac{3.0 \text{ nC}}{0.070 \text{ m}} - \frac{3.0 \text{ nC}}{0.030 \text{ m}} \right) = -5.1 \times 10^2 \text{ V}$$

$$\text{c. } V_p = k \sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2) \left(\frac{3.0 \text{ nC}}{0.030 \text{ m}} - \frac{3.0 \text{ nC}}{0.050 \text{ m}} \right) = 3.6 \times 10^2 \text{ V}$$

Significance

Note that evaluating potential is significantly simpler than electric field, due to potential being a scalar instead of a vector.

? Exercise 16.2.1

What is the potential on the x -axis? The z -axis?

Answer

The x -axis the potential is zero, due to the equal and opposite charges the same distance from it. On the z -axis, we may superimpose the two potentials; we will find that for $z \gg d$, again the potential goes to zero due to cancellation.

Now let us consider the special case when the distance of the point P from the dipole is much greater than the distance between the charges in the dipole, $r \gg d$; for example, when we are interested in the electric potential due to a polarized molecule such as a water molecule. This is not so far (infinity) that we can simply treat the potential as zero, but the distance is great enough that we can simplify our calculations relative to the previous example.

We start by noting that in Figure 16.2.4 the potential is given by

$$V_p = V_+ + V_- = k \left(\frac{q}{r_+} - \frac{q}{r_-} \right) \quad (16.2.8)$$

where

$$r_{\pm} = \sqrt{x^2 + \left(z \pm \frac{d}{2} \right)^2}. \quad (16.2.9)$$

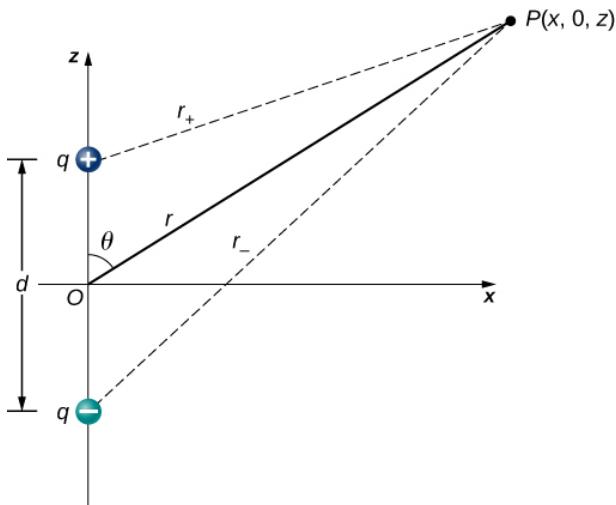


Figure 16.2.4: A general diagram of an electric dipole, and the notation for the distances from the individual charges to a point P in space.

This is still the exact formula. To take advantage of the fact that $r \gg d$, we rewrite the radii in terms of polar coordinates, with $x = r \sin \theta$ and $z = r \cos \theta$. This gives us

$$r_{\pm} = \sqrt{r^2 \sin^2 \theta + \left(r \cos \theta \pm \frac{d}{2} \right)^2}. \quad (16.2.10)$$

We can simplify this expression by pulling r out of the root,

$$r_{\pm} = \sqrt{\sin^2 \theta + \left(r \cos \theta \pm \frac{d}{2} \right)^2} \quad (16.2.11)$$

and then multiplying out the parentheses

$$r_{\pm} = r \sqrt{\sin^2 \theta + \cos^2 \theta \pm \cos \theta \frac{d}{r} + \left(\frac{d}{2r} \right)^2} = r \sqrt{1 \pm \cos \theta \frac{d}{r} + \left(\frac{d}{2r} \right)^2}. \quad (16.2.12)$$

The last term in the root is small enough to be negligible (remember $r \gg d$, and hence $(d/r)^2$ is extremely small, effectively zero to the level we will probably be measuring), leaving us with

$$r_{\pm} = r \sqrt{1 \pm \cos \theta \frac{d}{r}}. \quad (16.2.13)$$

Using the **binomial approximation** (a standard result from the mathematics of series, when a is small)

$$\frac{1}{\sqrt{1 \pm a}} \approx 1 \pm \frac{a}{2} \quad (16.2.14)$$

and substituting this into our formula for V_p , we get

$$V_p = k \left[\frac{q}{r} \left(1 + \frac{d \cos \theta}{2r} \right) - \frac{q}{r} \left(1 - \frac{d \cos \theta}{2r} \right) \right] = k \frac{qd \cos \theta}{r^2}. \quad (16.2.15)$$

This may be written more conveniently if we define a new quantity, the **electric dipole moment**,

$$\vec{p} = q\vec{d}, \quad (16.2.16)$$

where these vectors point from the negative to the positive charge. Note that this has magnitude qd . This quantity allows us to write the potential at point P due to a dipole at the origin as

$$V_p = k \frac{\vec{p} \cdot \hat{r}}{r^2}. \quad (16.2.17)$$

A diagram of the application of this formula is shown in Figure 16.2.5.

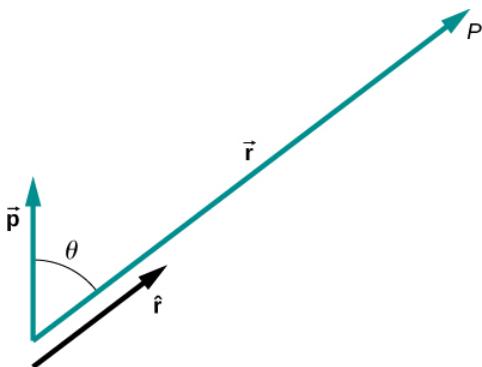


Figure 16.2.5: The geometry for the application of the potential of a dipole.

There are also higher-order moments for **quadrupoles**, **octupoles**, and so on.

This page titled [16.2: Electric Dipoles](#) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [5.8: Electric Dipoles](#) by [OpenStax](#) is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.
- [7.4: Calculations of Electric Potential](#) by [OpenStax](#) is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

16.3: Calculating Electric Fields of Charge Distributions

LEARNING OBJECTIVES

By the end of this section, you will be able to:

- Explain what a continuous source charge distribution is and how it is related to the concept of quantization of charge
- Describe line charges, surface charges, and volume charges
- Calculate the field of a continuous source charge distribution of either sign

The charge distributions we have seen so far have been discrete: made up of individual point particles. This is in contrast with a **continuous charge distribution**, which has at least one nonzero dimension. If a charge distribution is continuous rather than discrete, we can generalize the definition of the electric field. We simply divide the charge into infinitesimal pieces and treat each piece as a point charge.

Note that because charge is quantized, there is no such thing as a “truly” continuous charge distribution. However, in most practical cases, the total charge creating the field involves such a huge number of discrete charges that we can safely ignore the discrete nature of the charge and consider it to be continuous. This is exactly the kind of approximation we make when we deal with a bucket of water as a continuous fluid, rather than a collection of H₂O molecules.

Our first step is to define a charge density for a charge distribution along a line, across a surface, or within a volume, as shown in Figure 16.3.1.

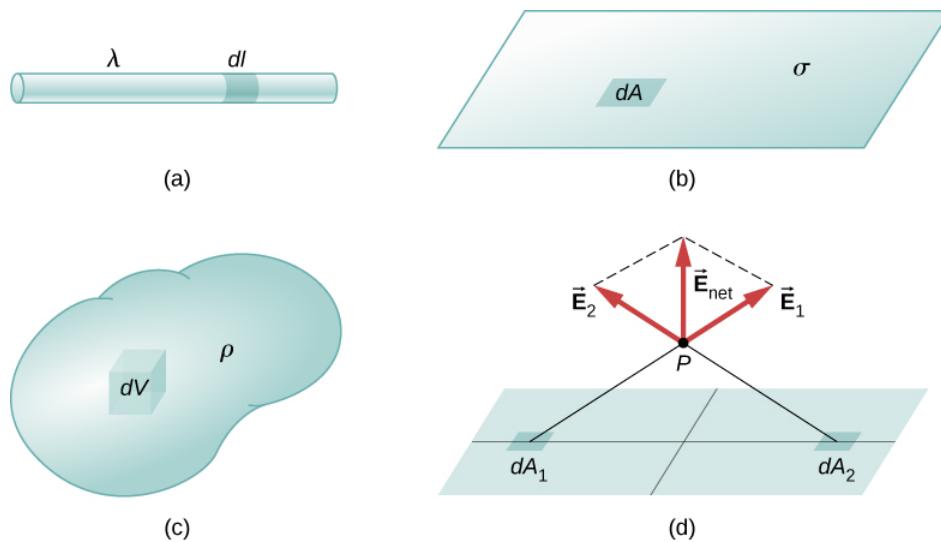


Figure 16.3.1: The configuration of charge differential elements for a (a) line charge, (b) sheet of charge, and (c) a volume of charge. Also note that (d) some of the components of the total electric field cancel out, with the remainder resulting in a net electric field.

Definitions: Charge Densities

Definitions of charge density:

- **linear charge density:** $\lambda \equiv$ charge per unit length (Figure 16.3.1a); units are coulombs per meter (C/m)
- **surface charge density:** $\sigma \equiv$ charge per unit area (Figure 16.3.1b); units are coulombs per square meter (C/m²)
- **volume charge density:** $\rho \equiv$ charge per unit volume (Figure 16.3.1c); units are coulombs per cubic meter (C/m³)

For a line charge, a surface charge, and a volume charge, the summation in the definition of an Electric field discussed previously becomes an integral and q_i is replaced by $dq = \lambda dl$, σdA , or ρdV , respectively:

$$\vec{E}(P) = \underbrace{\frac{1}{4\pi\epsilon_0} \sum_{i=1}^N \left(\frac{q_i}{r^2} \right) \hat{r}}_{\text{Point charges}} \quad (16.3.1)$$

$$\vec{E}(P) = \underbrace{\frac{1}{4\pi\epsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2} \right) \hat{r}}_{\text{Line charge}} \quad (16.3.2)$$

$$\vec{E}(P) = \underbrace{\frac{1}{4\pi\epsilon_0} \int_{\text{surface}} \left(\frac{\sigma dA}{r^2} \right) \hat{r}}_{\text{Surface charge}} \quad (16.3.3)$$

$$\vec{E}(P) = \underbrace{\frac{1}{4\pi\epsilon_0} \int_{\text{volume}} \left(\frac{\rho dV}{r^2} \right) \hat{r}}_{\text{Volume charge}} \quad (16.3.4)$$

The integrals in Equations 16.3.1–16.3.4 are generalizations of the expression for the field of a point charge. They implicitly include and assume the principle of superposition. The “trick” to using them is almost always in coming up with correct expressions for dl , dA , or dV , as the case may be, expressed in terms of \mathbf{r} , and also expressing the charge density function appropriately. It may be constant; it might be dependent on location.

Note carefully the meaning of r in these equations: It is the distance from the charge element (q_i , λdl , σdA , ρdV) to the location of interest, $P(x, y, z)$ (the point in space where you want to determine the field). However, don’t confuse this with the meaning of \hat{r} ; we are using it and the vector notation \vec{E} to write three integrals at once. That is, Equation 16.3.2 is actually

$$E_x(P) = \frac{1}{4\pi\epsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2} \right)_x, \quad (16.3.5)$$

$$E_y(P) = \frac{1}{4\pi\epsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2} \right)_y, \quad (16.3.6)$$

$$E_z(P) = \frac{1}{4\pi\epsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2} \right)_z \quad (16.3.7)$$

✓ Example 16.3.1: Electric Field of a Line Segment

Find the electric field a distance z above the midpoint of a straight line segment of length L that carries a uniform line charge density λ .

Strategy

Since this is a continuous charge distribution, we conceptually break the wire segment into differential pieces of length dl , each of which carries a differential amount of charge

$$dq = \lambda dl.$$

Then, we calculate the differential field created by two symmetrically placed pieces of the wire, using the symmetry of the setup to simplify the calculation (Figure 16.3.2). Finally, we integrate this differential field expression over the length of the wire (half of it, actually, as we explain below) to obtain the complete electric field expression.

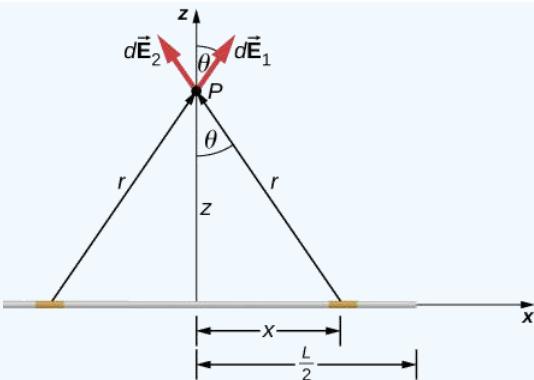


Figure 16.3.2: A uniformly charged segment of wire. The electric field at point P can be found by applying the superposition principle to symmetrically placed charge elements and integrating.

Solution

Before we jump into it, what do we expect the field to “look like” from far away? Since it is a finite line segment, from far away, it should look like a point charge. We will check the expression we get to see if it meets this expectation.

The electric field for a line charge is given by the general expression

$$\vec{E}(P) = \frac{1}{4\pi\epsilon_0} \int_{\text{line}} \frac{\lambda dl}{r^2} \hat{r}.$$

The symmetry of the situation (our choice of the two identical differential pieces of charge) implies the horizontal (x)-components of the field cancel, so that the net field points in the z -direction. Let’s check this formally.

The total field $\vec{E}(P)$ is the vector sum of the fields from each of the two charge elements (call them \vec{E}_1 and \vec{E}_2 , for now):

$$\begin{aligned}\vec{E}(P) &= \vec{E}_1 + \vec{E}_2 \\ &= E_{1x}\hat{i} + E_{1z}\hat{k} + E_{2x}(-\hat{i}) + E_{2z}\hat{k}.\end{aligned}$$

Because the two charge elements are identical and are the same distance away from the point P where we want to calculate the field, $E_{1x} = E_{2x}$, so those components cancel. This leaves

$$\begin{aligned}\vec{E}(P) &= E_{1z}\hat{k} + E_{2z}\hat{k} \\ &= E_1 \cos \theta \hat{k} + E_2 \cos \theta \hat{k}.\end{aligned}$$

These components are also equal, so we have

$$\begin{aligned}\vec{E}(P) &= \frac{1}{4\pi\epsilon_0} \int \frac{\lambda dl}{r^2} \cos \theta \hat{k} + \frac{1}{4\pi\epsilon_0} \int \frac{\lambda dl}{r^2} \cos \theta \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} \int_0^{L/2} \frac{2\lambda dx}{r^2} \cos \theta \hat{k}\end{aligned}$$

where our differential line element dl is dx , in this example, since we are integrating along a line of charge that lies on the x -axis. (The limits of integration are 0 to $\frac{L}{2}$, not $-\frac{L}{2}$ to $+\frac{L}{2}$, because we have constructed the net field from two differential pieces of charge dq . If we integrated along the entire length, we would pick up an erroneous factor of 2.)

In principle, this is complete. However, to actually calculate this integral, we need to eliminate all the variables that are not given. In this case, both r and θ change as we integrate outward to the end of the line charge, so those are the variables to get rid of. We can do that the same way we did for the two point charges: by noticing that

$$r = (z^2 + x^2)^{1/2}$$

and

$$\cos \theta = \frac{z}{r} = \frac{z}{(z^2 + x^2)^{1/2}}.$$

Substituting, we obtain

$$\begin{aligned}\vec{E}(P) &= \frac{1}{4\pi\epsilon_0} \int_0^{L/2} \frac{2\lambda dx}{(z^2 + x^2)} \frac{z}{(z^2 + x^2)^{1/2}} \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} \int_0^{L/2} \frac{2\lambda z}{(z^2 + x^2)^{3/2}} dx \hat{k} \\ &= \frac{2\lambda z}{4\pi\epsilon_0} \left[\frac{x}{z^2\sqrt{z^2 + x^2}} \right]_0^{L/2} \hat{k}.\end{aligned}$$

which simplifies to

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0} \frac{\lambda L}{z\sqrt{z^2 + \frac{L^2}{4}}} \hat{k}. \quad (16.3.8)$$

Significance

Notice, once again, the use of symmetry to simplify the problem. This is a very common strategy for calculating electric fields. The fields of nonsymmetrical charge distributions have to be handled with multiple integrals and may need to be calculated numerically by a computer.

Exercise 16.3.1

How would the strategy used above change to calculate the electric field at a point a distance z above one end of the finite line segment?

Answer

We will no longer be able to take advantage of symmetry. Instead, we will need to calculate each of the two components of the electric field with their own integral.

Example 16.3.2: Electric Field of an Infinite Line of Charge

Find the electric field a distance z above the midpoint of an infinite line of charge that carries a uniform line charge density λ .

Strategy

This is exactly like the preceding example, except the limits of integration will be $-\infty$ to $+\infty$.

Solution

Again, the horizontal components cancel out, so we wind up with

$$\vec{E}(P) = \frac{1}{4\pi\epsilon_0} \int_{-\infty}^{\infty} \frac{\lambda dx}{r^2} \cos \theta \hat{k}$$

where our differential line element dl is dx , in this example, since we are integrating along a line of charge that lies on the x -axis. Again,

$$\begin{aligned}\cos \theta &= \frac{z}{r} \\ &= \frac{z}{(z^2 + x^2)^{1/2}}.\end{aligned}$$

Substituting, we obtain

$$\begin{aligned}\vec{E}(P) &= \frac{1}{4\pi\epsilon_0} \int_{-\infty}^{\infty} \frac{\lambda dx}{(z^2 + x^2)} \frac{z}{(z^2 + x^2)^{1/2}} \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} \int_{-\infty}^{\infty} \frac{\lambda z}{(z^2 + x^2)^{3/2}} dx \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} \left[\frac{x}{z^2 \sqrt{z^2 + x^2}} \right]_{-\infty}^{\infty} \hat{k}\end{aligned}$$

which simplifies to

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{z} \hat{k}.$$

Significance

Our strategy for working with continuous charge distributions also gives useful results for charges with infinite dimension.

In the case of a finite line of charge, note that for $z \gg L$, z^2 dominates the L in the denominator, so that Equation 16.3.8 simplifies to

$$\vec{E} \approx \frac{1}{4\pi\epsilon_0} \frac{\lambda L}{z^2} \hat{k}.$$

If you recall that $\lambda L = q$ the total charge on the wire, we have retrieved the expression for the field of a point charge, as expected.

In the limit $L \rightarrow \infty$ on the other hand, we get the field of an **infinite straight wire**, which is a straight wire whose length is much, much greater than either of its other dimensions, and also much, much greater than the distance at which the field is to be calculated:

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{z} \hat{k}. \quad (16.3.9)$$

An interesting artifact of this infinite limit is that we have lost the usual $1/r^2$ dependence that we are used to. This will become even more intriguing in the case of an infinite plane.

✓ Example 16.3.3A: Electric Field due to a Ring of Charge

A ring has a uniform charge density λ , with units of coulomb per unit meter of arc. Find the electric field at a point on the axis passing through the center of the ring.

Strategy

We use the same procedure as for the charged wire. The difference here is that the charge is distributed on a circle. We divide the circle into infinitesimal elements shaped as arcs on the circle and use polar coordinates shown in Figure 16.3.3.

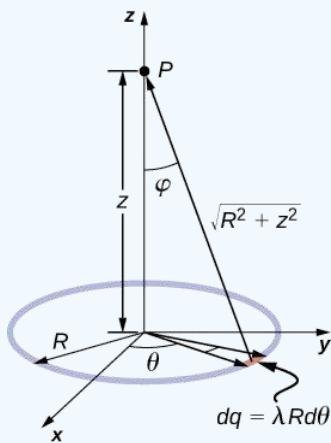


Figure 16.3.3: The system and variable for calculating the electric field due to a ring of charge.

Solution

The electric field for a line charge is given by the general expression

$$\vec{E}(P) = \frac{1}{4\pi\epsilon_0} \int_{\text{line}} \frac{\lambda dl}{r^2} \hat{r}.$$

A general element of the arc between θ and $\theta + d\theta$ is of length $R d\theta$ and therefore contains a charge equal to $\lambda R d\theta$. The element is at a distance of $r = \sqrt{z^2 + R^2}$ from P , the angle is $\cos \phi = \frac{z}{\sqrt{z^2 + R^2}}$ and therefore the electric field is

$$\begin{aligned}\vec{E}(P) &= \frac{1}{4\pi\epsilon_0} \int_{\text{line}} \frac{\lambda dl}{r^2} \hat{r} = \frac{1}{4\pi\epsilon_0} \int_0^{2\pi} \frac{\lambda R d\theta}{z^2 + R^2} \frac{z}{\sqrt{z^2 + R^2}} \hat{z} \\ &= \frac{1}{4\pi\epsilon_0} \frac{\lambda R z}{(z^2 + R^2)^{3/2}} \hat{z} \int_0^{2\pi} d\theta \\ &= \frac{1}{4\pi\epsilon_0} \frac{2\pi\lambda R z}{(z^2 + R^2)^{3/2}} \hat{z} \\ &= \frac{1}{4\pi\epsilon_0} \frac{q_{\text{tot}} z}{(z^2 + R^2)^{3/2}} \hat{z}.\end{aligned}$$

Significance

As usual, symmetry simplified this problem, in this particular case resulting in a trivial integral. Also, when we take the limit of $z \gg R$, we find that

$$\vec{E} \approx \frac{1}{4\pi\epsilon_0} \frac{q_{\text{tot}} z}{z^2} \hat{z},$$

as we expect.

✓ Example 16.3.3B: The Field of a Disk

Find the electric field of a circular thin disk of radius R and uniform charge density at a distance z above the center of the disk (Figure 16.3.4)

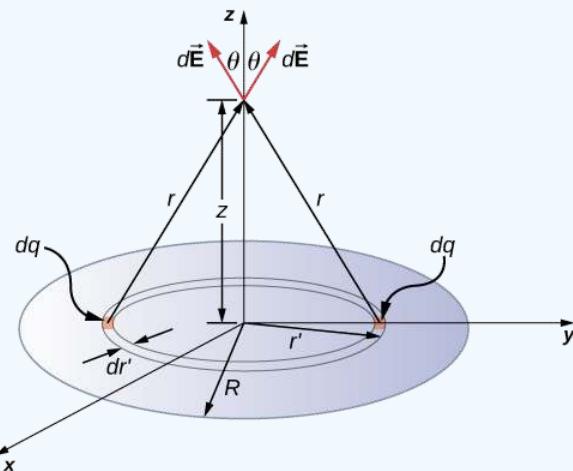


Figure 16.3.4: A uniformly charged disk. As in the line charge example, the field above the center of this disk can be calculated by taking advantage of the symmetry of the charge distribution.

Strategy

The electric field for a surface charge is given by

$$\vec{E}(P) = \frac{1}{4\pi\epsilon_0} \int_{\text{surface}} \frac{\sigma dA}{r^2} \hat{r}.$$

To solve surface charge problems, we break the surface into symmetrical differential “stripes” that match the shape of the surface; here, we’ll use rings, as shown in the figure. Again, by symmetry, the horizontal components cancel and the field is entirely in the vertical (\hat{k}) direction. The vertical component of the electric field is extracted by multiplying by θ , so

$$\vec{E}(P) = \frac{1}{4\pi\epsilon_0} \int_{surface} \frac{\sigma dA}{r^2} \cos\theta \hat{k}.$$

As before, we need to rewrite the unknown factors in the integrand in terms of the given quantities. In this case,

$$dA = 2\pi r' dr'$$

$$r^2 = r'^2 + z^2$$

$$\cos\theta = \frac{z}{(r'^2 + z^2)^{1/2}}.$$

(Please take note of the two different “ r ’s” here; r is the distance from the differential ring of charge to the point P where we wish to determine the field, whereas r' is the distance from the center of the disk to the differential ring of charge.) Also, we already performed the polar angle integral in writing down dA .

Solution

Substituting all this in, we get

$$\begin{aligned}\vec{E}(P) &= \vec{E}(z) \\ &= \frac{1}{4\pi\epsilon_0} \int_0^R \frac{\sigma(2\pi r' dr')z}{(r'^2 + z^2)^{3/2}} \hat{k} \\ &= \frac{1}{4\pi\epsilon_0} (2\pi\sigma z) \left(\frac{1}{z} - \frac{1}{\sqrt{R^2 + z^2}} \right) \hat{k}\end{aligned}$$

or, more simply,

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0} \left(2\pi\sigma - \frac{2\pi\sigma z}{\sqrt{R^2 + z^2}} \right) \hat{k}. \quad (16.3.10)$$

Significance

Again, it can be shown (via a Taylor expansion) that when $z \gg R$, this reduces to

$$\vec{E}(z) \approx \frac{1}{4\pi\epsilon_0} \frac{\sigma\pi R^2}{z^2} \hat{k},$$

which is the expression for a point charge $Q = \sigma\pi R^2$.

Exercise 16.3.3

How would the above limit change with a uniformly charged rectangle instead of a disk?

Answer

The point charge would be $Q = \sigma ab$ where a and b are the sides of the rectangle but otherwise identical.

As $R \rightarrow \infty$, Equation 16.3.10 reduces to the field of an infinite plane, which is a flat sheet whose area is much, much greater than its thickness, and also much, much greater than the distance at which the field is to be calculated:

$$\vec{E} = \lim_{R \rightarrow \infty} \frac{1}{4\pi\epsilon_0} \left(2\pi\sigma - \frac{2\pi\sigma z}{\sqrt{R^2 + z^2}} \right) \hat{k} \quad (16.3.11)$$

$$= \frac{\sigma}{2\epsilon_0} \hat{k}. \quad (16.3.12)$$

Note that this field is constant. This surprising result is, again, an artifact of our limit, although one that we will make use of repeatedly in the future. To understand why this happens, imagine being placed above an infinite plane of constant charge. Does the plane look any different if you vary your altitude? No—you still see the plane going off to infinity, no matter how far you are from it. It is important to note that Equation 16.3.12 is because we are above the plane. If we were below, the field would point in the $-\hat{k}$ direction.

✓ Example 16.3.4: The Field of Two Infinite Planes

Find the electric field everywhere resulting from two infinite planes with equal but opposite charge densities (Figure 16.3.5).

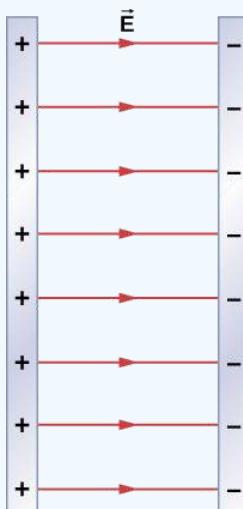


Figure 16.3.5: Two charged infinite planes. Note the direction of the electric field.

Strategy

We already know the electric field resulting from a single infinite plane, so we may use the principle of superposition to find the field from two.

Solution

The electric field points away from the positively charged plane and toward the negatively charged plane. Since the σ are equal and opposite, this means that in the region outside of the two planes, the electric fields cancel each other out to zero. However, in the region between the planes, the electric fields add, and we get

$$\vec{E} = \frac{\sigma}{\epsilon_0} \hat{i}$$

for the electric field. The \hat{i} is because in the figure, the field is pointing in the $+x$ -direction.

Significance

Systems that may be approximated as two infinite planes of this sort provide a useful means of creating uniform electric fields.

? Exercise 16.3.1

What would the electric field look like in a system with two parallel positively charged planes with equal charge densities?

Answer

The electric field would be zero in between, and have magnitude $\frac{\sigma}{\epsilon_0}$ everywhere else.

This page titled [16.3: Calculating Electric Fields of Charge Distributions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [5.6: Calculating Electric Fields of Charge Distributions](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

16.4: Calculating Electric Potential of Charge Distributions

Learning Objectives

By the end of this section, you will be able to:

- Calculate the potential of various continuous charge distributions

How do you calculate the electric potential of continuous charge distributions? Recall that for multiple point charges,

$$V_p = k \sum \frac{q_i}{r_i}. \quad (16.4.1)$$

We may treat a continuous charge distribution as a collection of infinitesimally separated individual points. This yields the integral

$$V_p = \int \frac{dq}{r} \quad (16.4.2)$$

for the potential at a point P . Note that r is the distance from each individual point in the charge distribution to the point P . As we saw in [Electric Charges and Fields](#), the infinitesimal charges are given by

$$\underbrace{dq = \lambda dl}_{\text{one dimension}} \quad (16.4.3)$$

$$\underbrace{dq = \sigma dA}_{\text{two dimensions}} \quad (16.4.4)$$

$$\underbrace{dq = \rho dV}_{\text{three dimensions}} \quad (16.4.5)$$

where λ is linear charge density, σ is the charge per unit area, and ρ is the charge per unit volume.

✓ Example 16.4.1: Potential of a Line of Charge

Find the electric potential of a uniformly charged, nonconducting wire with linear density λ (coulomb/meter) and length L at a point that lies on a line that divides the wire into two equal parts.

Strategy

To set up the problem, we choose Cartesian coordinates in such a way as to exploit the symmetry in the problem as much as possible. We place the origin at the center of the wire and orient the y -axis along the wire so that the ends of the wire are at $y = \pm L/2$. The field point P is in the xy -plane and since the choice of axes is up to us, we choose the x -axis to pass through the field point P , as shown in Figure 16.4.1.

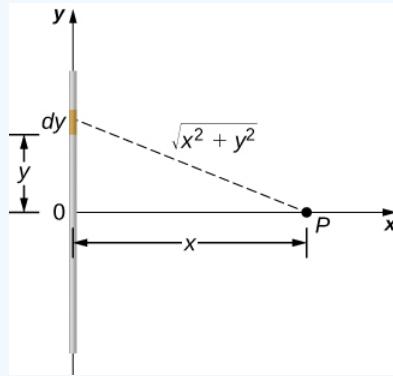


Figure 16.4.1: We want to calculate the electric potential due to a line of charge.

Solution

Consider a small element of the charge distribution between y and $y + dy$. The charge in this cell is $dq = \lambda dy$ and the distance from the cell to the field point P is $\sqrt{x^2 + y^2}$. Therefore, the potential becomes

$$\begin{aligned}
 V_p &= k \int \frac{dq}{r} \\
 &= k \int_{-L/2}^{L/2} \frac{\lambda dy}{\sqrt{x^2 + y^2}} \\
 &= k\lambda \left[\ln \left(y + \sqrt{y^2 + x^2} \right) \right]_{-L/2}^{L/2} \\
 &= k\lambda \left[\ln \left(\left(\frac{L}{2} \right) + \sqrt{\left(\frac{L}{2} \right)^2 + x^2} \right) - \ln \left(\left(-\frac{L}{2} \right) + \sqrt{\left(-\frac{L}{2} \right)^2 + x^2} \right) \right] \\
 &= k\lambda \ln \left[\frac{L + \sqrt{L^2 + 4x^2}}{-L + \sqrt{L^2 + 4x^2}} \right].
 \end{aligned}$$

Significance

Note that this was simpler than the equivalent problem for electric field, due to the use of scalar quantities. Recall that we expect the zero level of the potential to be at infinity, when we have a finite charge. To examine this, we take the limit of the above potential as x approaches infinity; in this case, the terms inside the natural log approach one, and hence the potential approaches zero in this limit. Note that we could have done this problem equivalently in cylindrical coordinates; the only effect would be to substitute r for x and z for y .

✓ Example 16.4.2: Potential Due to a Ring of Charge

A ring has a uniform charge density λ , with units of coulomb per unit meter of arc. Find the electric potential at a point on the axis passing through the center of the ring.

Strategy

We use the same procedure as for the charged wire. The difference here is that the charge is distributed on a circle. We divide the circle into infinitesimal elements shaped as arcs on the circle and use cylindrical coordinates shown in Figure 16.4.2.

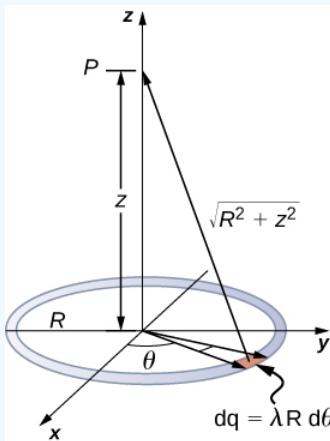


Figure 16.4.2: We want to calculate the electric potential due to a ring of charge.

Solution

A general element of the arc between θ and $\theta + d\theta$ is of length $Rd\theta$ and therefore contains a charge equal to $\lambda R d\theta$. The element is at a distance of $\sqrt{z^2 + R^2}$ from P , and therefore the potential is

$$\begin{aligned}
 V_p &= k \int \frac{dq}{r} \\
 &= k \int_0^{2\pi} \frac{\lambda R d\theta}{\sqrt{z^2 + R^2}} \\
 &= \frac{k\lambda R}{\sqrt{z^2 + R^2}} \int_0^{2\pi} d\theta \\
 &= \frac{2\pi k\lambda R}{\sqrt{z^2 + R^2}} \\
 &= k \frac{q_{tot}}{\sqrt{z^2 + R^2}}.
 \end{aligned}$$

Significance

This result is expected because every element of the ring is at the same distance from point P . The net potential at P is that of the total charge placed at the common distance, $\sqrt{z^2 + R^2}$.

✓ Example 16.4.3: Potential Due to a Uniform Disk of Charge

A disk of radius R has a uniform charge density σ with units of coulomb meter squared. Find the electric potential at any point on the axis passing through the center of the disk.

Strategy

We divide the disk into ring-shaped cells, and make use of the result for a ring worked out in the previous example, then integrate over r in addition to θ . This is shown in Figure 16.4.3.

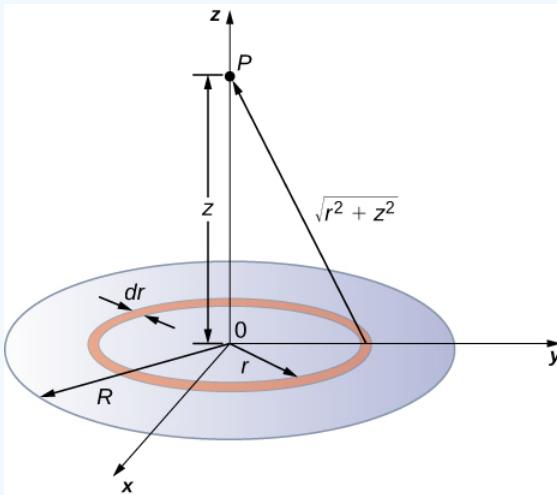


Figure 16.4.3: We want to calculate the electric potential due to a disk of charge.

Solution

An infinitesimal width cell between cylindrical coordinates r and $r + dr$ shown in Figure 16.4.3 will be a ring of charges whose electric potential dV_p at the field point has the following expression

$$dV_p = k \frac{dq}{\sqrt{z^2 + r^2}} \quad (16.4.6)$$

where

$$dq = \sigma 2\pi r dr. \quad (16.4.7)$$

The superposition of potential of all the infinitesimal rings that make up the disk gives the net potential at point P . This is accomplished by integrating from $r = 0$ to $r = R$:

$$V_p = \int dV_p = k2\pi\sigma \int_0^R \frac{r dr}{\sqrt{z^2 + r^2}}, \\ = k2\pi\sigma(\sqrt{z^2 + R^2} - \sqrt{z^2}).$$

Significance

The basic procedure for a disk is to first integrate around θ and then over r . This has been demonstrated for uniform (constant) charge density. Often, the charge density will vary with r , and then the last integral will give different results.

✓ Example 16.4.4: Potential Due to an Infinite Charged Wire

Find the electric potential due to an infinitely long uniformly charged wire.

Strategy

Since we have already worked out the potential of a finite wire of length L in Example 16.4.1, we might wonder if taking $L \rightarrow \infty$ in our previous result will work:

$$V_p = \lim_{L \rightarrow \infty} k\lambda \ln \left(\frac{L + \sqrt{L^2 + 4x^2}}{-L + \sqrt{L^2 + 4x^2}} \right). \quad (16.4.8)$$

However, this limit does not exist because the argument of the logarithm becomes [2/0] as $L \rightarrow \infty$, so this way of finding V of an infinite wire does not work. The reason for this problem may be traced to the fact that the charges are not localized in some space but continue to infinity in the direction of the wire. Hence, our (unspoken) assumption that zero potential must be at an infinite distance from the wire is no longer valid.

To avoid this difficulty in calculating limits, let us use the definition of potential by integrating over the electric field from the previous section, and the value of the electric field from this charge configuration from the previous chapter.

Solution

We use the integral

$$V_p = - \int_R^p \vec{E} \cdot d\vec{l} \quad (16.4.9)$$

where R is a finite distance from the line of charge, as shown in Figure 16.4.4.

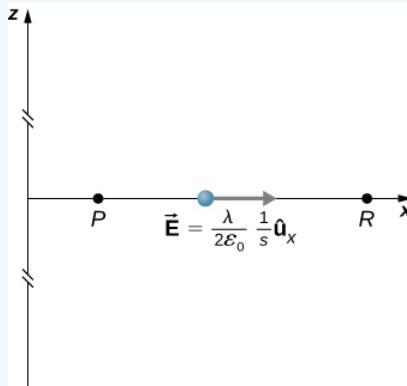


Figure 16.4.4: Points of interest for calculating the potential of an infinite line of charge.

With this setup, we use $\vec{E}_p = 2k\lambda \frac{1}{s} \hat{s}$ and $d\vec{l} = d\vec{s}$ to obtain

$$V_p - V_R = - \int_R^p 2k\lambda \frac{1}{s} ds \\ = -2k\lambda \ln \frac{s_p}{s_R}.$$

Now, if we define the reference potential $V_R = 0$ at $s_R = 1 \text{ m}$, this simplifies to

$$V_p = -2k\lambda \ln s_p. \quad (16.4.10)$$

Note that this form of the potential is quite usable; it is 0 at 1 m and is undefined at infinity, which is why we could not use the latter as a reference.

Significance

Although calculating potential directly can be quite convenient, we just found a system for which this strategy does not work well. In such cases, going back to the definition of potential in terms of the electric field may offer a way forward.

? Exercise 16.4.1

What is the potential on the axis of a nonuniform ring of charge, where the charge density is $\lambda(\theta) = \lambda \cos \theta$?

Solution

It will be zero, as at all points on the axis, there are equal and opposite charges equidistant from the point of interest. Note that this distribution will, in fact, have a dipole moment.

16.4: Calculating Electric Potential of Charge Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Ronald Kumon & OpenStax.

- 7.4: Calculations of Electric Potential by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

16.5: Direct Calculation of Electrical Quantities from Charge Distributions (Summary)

Key Terms

continuous charge distribution	total source charge composed of so large a number of elementary charges that it must be treated as continuous, rather than discrete
infinite straight wire	straight wire whose length is much, much greater than either of its other dimensions, and also much, much greater than the distance at which the field is to be calculated
linear charge density	amount of charge in an element of a charge distribution that is essentially one-dimensional (the width and height are much, much smaller than its length); its units are C/m
surface charge density	amount of charge in an element of a two-dimensional charge distribution (the thickness is small); its units are C/m^2
volume charge density	amount of charge in an element of a three-dimensional charge distribution; its units are C/m^3

Key Equations

Coulomb's law	$\vec{F}_{12}(r) = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{r}_{12}$
Superposition of electric forces	$\vec{F}(r) = \frac{1}{4\pi\epsilon_0} Q \sum_{i=1}^N \frac{q_i}{r_i^2} \hat{r}_i$
Electric force due to an electric field	$\vec{F} = Q\vec{E}$
Electric field at point P	$\vec{E}(P) \equiv \frac{1}{4\pi\epsilon_0} \sum_{i=1}^N \frac{q_i}{r_i^2} \hat{r}_i$
Field of an infinite wire	$\vec{E}(z) = \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{z} \hat{k}$
Field of an infinite plane	$\vec{E} = \frac{\sigma}{2\epsilon_0} \hat{k}$
Dipole moment	$\vec{p} \equiv q\vec{d}$
Torque on dipole in external E-field	$\vec{\tau} = \vec{p} \times \vec{E}$
Electric field of a continuous charge distribution	$\vec{E} = k \int \frac{dq\hat{r}}{r}$
Electric potential of a continuous charge distribution	$V_P = k \int \frac{dq}{r}$

Summary

Calculating Electric Fields of Charge Distributions

- A very large number of charges can be treated as a continuous charge distribution, where the calculation of the field requires integration. Common cases are:
 - one-dimensional (like a wire); uses a line charge density λ
 - two-dimensional (metal plate); uses surface charge density σ
 - three-dimensional (metal sphere); uses volume charge density ρ
- The “source charge” is a differential amount of charge dq . Calculating dq depends on the type of source charge distribution:

$$dq = \lambda dl; dq = \sigma dA; dq = \rho dV.$$

- The field of continuous charge distributions may be calculated with $\vec{E} = k \int \frac{dq\hat{r}}{r}$.

- Symmetry of the charge distribution is usually key.
- Important special cases are the field of an “infinite” wire and the field of an “infinite” plane.

Electric Dipoles

- If a permanent dipole is placed in an external electric field, it results in a torque that aligns it with the external field.
- If a nonpolar atom (or molecule) is placed in an external field, it gains an induced dipole that is aligned with the external field.
- The net field is the vector sum of the external field plus the field of the dipole (physical or induced).
- The strength of the polarization is described by the dipole moment of the dipole, $\vec{p} = q\vec{d}$.

Calculating Electric Potential of Charge Distributions

- The potential of continuous charge distributions may be calculated with $V_P = k \int \frac{dq}{r}$.

Contributors and Attributions

Samuel J. Ling (Truman State University), Jeff Sanny (Loyola Marymount University), and Bill Moebs with many contributing authors. This work is licensed by OpenStax University Physics under a [Creative Commons Attribution License \(by 4.0\)](#).

16.5: Direct Calculation of Electrical Quantities from Charge Distributions ([Summary](#)) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Ronald Kumon & OpenStax.

- [5.9: Electric Charges and Fields \(Summary\)](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.
- [7.8: Electric Potential \(Summary\)](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

16.6: Direct Calculation of Electrical Quantities from Charge Distributions (Exercises)

Conceptual Questions

Electric Dipoles

- 36.** What are the stable orientation(s) for a dipole in an external electric field? What happens if the dipole is slightly perturbed from these orientations?

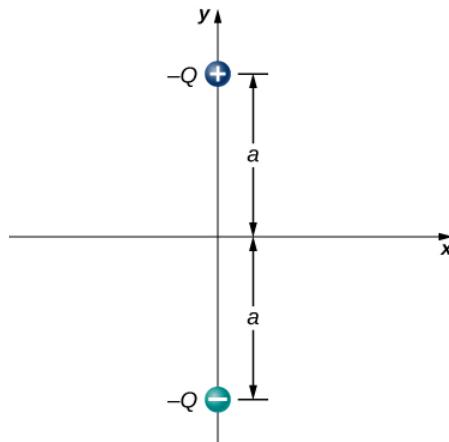
Calculating Electric Potential of Charge Distributions

- 13.** Compare the electric dipole moments of charges $\pm Q$ separated by a distance d and charges $\pm Q/2$ separated by a distance $d/2$.
- 15.** In what region of space is the potential due to a uniformly charged sphere the same as that of a point charge? In what region does it differ from that of a point charge?
- 16.** Can the potential of a nonuniformly charged sphere be the same as that of a point charge? Explain.

Problems

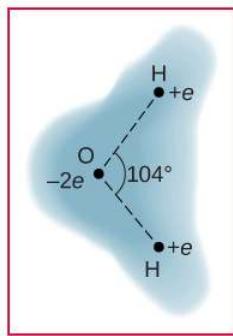
Electric Dipoles

- 105.** Consider the equal and opposite charges shown below. (a) Show that at all points on the x -axis for which $|x| \gg a$, $E \approx Qa/2\pi\epsilon_0 x^3$. (b) Show that at all points on the y -axis for which $|y| \gg a$, $E \approx Qa/\pi\epsilon_0 y^3$.



- 106.** (a) What is the dipole moment of the configuration shown above? If $Q=4.0\mu\text{C}$,
- (b) what is the torque on this dipole with an electric field of $4.0 \times 10^5 \text{ N/C} \hat{i}$?
 - (c) What is the torque on this dipole with an electric field of $-4.0 \times 10^5 \text{ N/C} \hat{i}$?
 - (d) What is the torque on this dipole with an electric field of $\pm 4.0 \times 10^5 \text{ N/C} \hat{j}$?

- 107.** A water molecule consists of two hydrogen atoms bonded with one oxygen atom. The bond angle between the two hydrogen atoms is **104°** (see below). Calculate the net dipole moment of a hypothetical water molecule where the charge at the oxygen molecule is $-2e$ and at each hydrogen atom is $+e$. The net dipole moment of the molecule is the vector sum of the individual dipole moment between the two O-Hs. The separation O-H is 0.9578 angstroms.

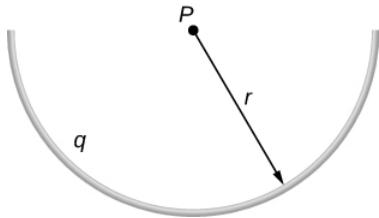


Calculating Electric Fields of Charge Distributions

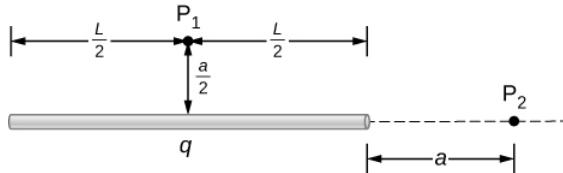
83. The charge per unit length on the thin rod shown below is λ . What is the electric field at the point P ? (Hint: Solve this problem by first considering the electric field $d\vec{E}$ at P due to a small segment dx of the rod, which contains charge $dq = \lambda dx$. Then find the net field by integrating $d\vec{E}$ over the length of the rod.)



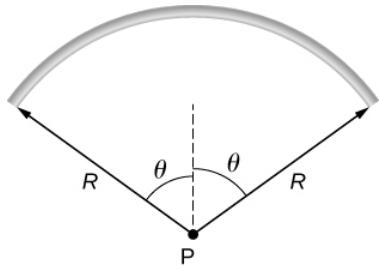
84. The charge per unit length on the thin semicircular wire shown below is λ . What is the electric field at the point P ?



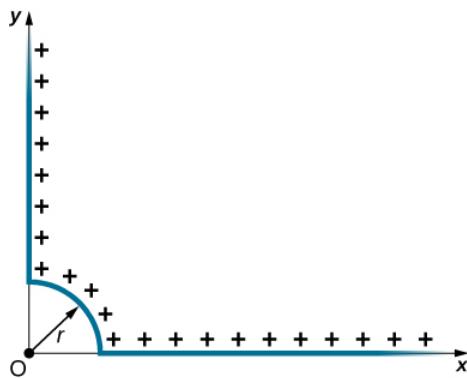
87. A total charge q is distributed uniformly along a thin, straight rod of length L (see below). What is the electric field at P_1 ? At P_2 ?



90. A rod bent into the arc of a circle subtends an angle 2θ at the center P of the circle (see below). If the rod is charged uniformly with a total charge Q , what is the electric field at P ?



97. Positive charge is distributed with a uniform density λ along the positive x -axis from r to ∞ , along the positive y -axis from r to ∞ , and along a 90° arc of a circle of radius r , as shown below. What is the electric field at O ?

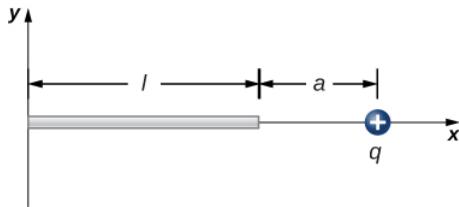


121. Charge is distributed uniformly along the entire y -axis with a density y_λ and along the positive x -axis from $x = a$ to $x = b$ with a density λ_x . What is the force between the two distributions?

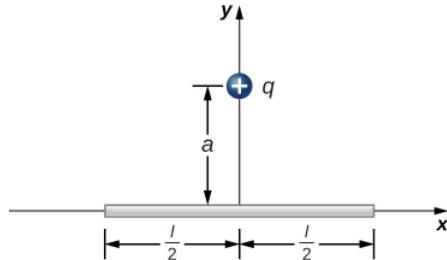
122. The circular arc shown below carries a charge per unit length $\lambda = \lambda_0 \cos\theta$, where θ is measured from the x -axis. What is the electric field at the origin?

123. Calculate the electric field due to a uniformly charged rod of length L , aligned with the x -axis with one end at the origin; at a point P on the z -axis.

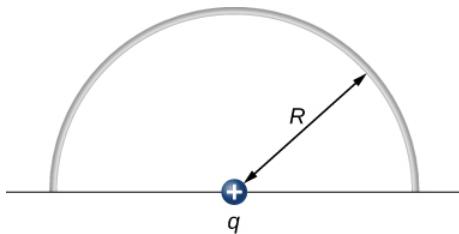
124. The charge per unit length on the thin rod shown below is λ . What is the electric force on the point charge q ? Solve this problem by first considering the electric force $d\vec{F}$ on q due to a small segment dx of the rod, which contains charge λdx . Then, find the net force by integrating $d\vec{F}$ over the length of the rod.



125. The charge per unit length on the thin rod shown here is λ . What is the electric force on the point charge q ? (See the preceding problem.)



126. The charge per unit length on the thin semicircular wire shown below is λ . What is the electric force on the point charge q ? (See the preceding problems.)



16.6: Direct Calculation of Electrical Quantities from Charge Distributions (Exercises) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- 5.10: Electric Charges and Fields (Exercises) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.
- 7.9: Electric Potential (Exercises) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

16.7: Direct Calculation of Electrical Quantities from Charge Distributions (Answers)

Note: Answers are provided for only the odd-numbered questions.

Conceptual Questions

Calculating Electric Potential of Charge Distributions

13. The second has 1/4 the dipole moment of the first.

15. The region outside of the sphere will have a potential indistinguishable from a point charge; the interior of the sphere will have a different potential.

Problems

Electric Dipoles

$$105. E_x = 0, E_y = \frac{1}{4\pi\epsilon_0} \left[\frac{2q}{(x^2 + a^2)} \right] \frac{a}{\sqrt{(x^2 + a^2)}} \Rightarrow x \gg a \Rightarrow \frac{1}{2\pi\epsilon_0} \frac{qa}{x^3}$$

$$E_y = \frac{q}{4\pi\epsilon_0} \left[\frac{2ya + 2ya}{(y-a)^2(y+a)^2} \right] \Rightarrow y \gg a \Rightarrow \frac{1}{\pi\epsilon_0} \frac{qa}{y^3}$$

107. The net dipole moment of the molecule is the vector sum of the individual dipole moments between the two O-H. The separation O-H is 0.9578 angstroms:

$$\vec{p} = 1.889 \times 10^{-29} Cm\hat{i}$$

Calculating Electric Fields of Charge Distributions

$$83. dE = \frac{1}{4\pi\epsilon_0} \frac{\lambda dx}{(x+a)^2}, E = \frac{\lambda}{4\pi\epsilon_0} \left[\frac{1}{l+a} - \frac{1}{a} \right]$$

$$87. \text{At } P_1 : \vec{E}(y) = \frac{1}{4\pi\epsilon_0} \frac{\lambda L}{y\sqrt{y^2 + \frac{L^2}{4}}} \hat{j} \Rightarrow \frac{1}{4\pi\epsilon_0} \frac{q}{\frac{a}{2}\sqrt{(\frac{a}{2})^2 + \frac{L^2}{4}}} \hat{j} = \frac{1}{\pi\epsilon_0} \frac{q}{a\sqrt{a^2 + L^2}} \hat{j}$$

At P_2 : Put the origin at the end of L.

$$dE = \frac{1}{4\pi\epsilon_0} \frac{\lambda dx}{(x+a)^2}, \vec{E} = -\frac{q}{4\pi\epsilon_0 l} \left[\frac{1}{l+a} - \frac{1}{a} \right] \hat{i}$$

$$97. \text{circular arc } dE_x(-\hat{i}) = \frac{1}{4\pi\epsilon_0} \frac{\lambda ds}{r^2} \cos\theta(-\hat{i}),$$

$$\vec{E}_x = \frac{\lambda}{4\pi\epsilon_0 r} (-\hat{i}),$$

$$dE_y(-\hat{i}) = \frac{1}{4\pi\epsilon_0} \frac{\lambda ds}{r^2} \sin\theta(-\hat{j}),$$

$$\vec{E}_y = \frac{\lambda}{4\pi\epsilon_0 r} (-\hat{j});$$

$$\text{y-axis: } \vec{E}_x = \frac{\lambda}{4\pi\epsilon_0 r} (-\hat{i});$$

$$\text{x-axis: } \vec{E}_y = \frac{\lambda}{4\pi\epsilon_0 r} (-\hat{j}),$$

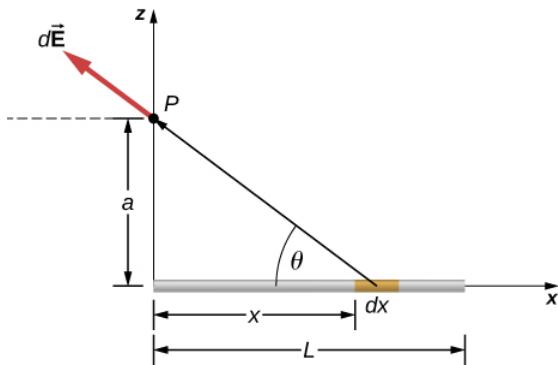
$$\vec{E} = \frac{\lambda}{2\pi\epsilon_0 r} (-\hat{i}) + \frac{\lambda}{2\pi\epsilon_0 r} (-\hat{j})$$

Additional Problems

$$121. \text{Electric field of wire at x: } \vec{E}(x) = \frac{1}{4\pi\epsilon_0} \frac{2\lambda_y}{x} \hat{i},$$

$$dF = \frac{\lambda_y \lambda_x}{2\pi\epsilon_0} (\ln b - \ln a)$$

123.



$$dEx = \frac{1}{4\pi\epsilon_0} \frac{\lambda dx}{(x^2 + a^2)} \frac{x}{\sqrt{x^2 + a^2}},$$

$$\vec{E}_x = \frac{\lambda}{4\pi\epsilon_0} \left[\frac{1}{\sqrt{L^2 + a^2}} - \frac{1}{a} \right] \hat{i},$$

$$dE_z = \frac{1}{4\pi\epsilon_0} \frac{\lambda dx}{(x^2 + a^2)} \frac{a}{\sqrt{x^2 + a^2}},$$

$$\vec{E}_z = \frac{\lambda}{4\pi\epsilon_0 a} \frac{L}{\sqrt{L^2 + a^2}} \hat{k},$$

Substituting z for a, we have:

$$\vec{E}(z) = \frac{\lambda}{4\pi\epsilon_0} \left[\frac{1}{\sqrt{L^2 + z^2}} - \frac{1}{z} \right] \hat{i} + \frac{\lambda}{4\pi\epsilon_0 z} \frac{L}{\sqrt{L^2 + z^2}} \hat{k}$$

125. There is a net force only in the y-direction. Let θ be the angle the vector from dx to q makes with the x-axis. The components along the x-axis cancel due to symmetry, leaving the y-component of the force.

$$dF_y = \frac{1}{4\pi\epsilon_0} \frac{aq\lambda dx}{(x^2 + a^2)^{3/2}},$$

$$F_y = \frac{1}{2\pi\epsilon_0} \frac{q\lambda}{a} \left[\frac{l/2}{((l/2)^2 + a^2)^{1/2}} \right]$$

16.7: Direct Calculation of Electrical Quantities from Charge Distributions (Answers) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- 7.10: Electric Potential (Answer) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.
- 5.11: Electric Charges and Fields (Answer) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

CHAPTER OVERVIEW

17: Gauss's Law for Calculation of Electrical Field from Charge Distributions

- [17.1: Introduction to Gauss's Law](#)
- [17.2: Electric Flux](#)
- [17.3: Gauss's Law](#)
- [17.4: Calculating Electric Field Using Gauss's Law](#)
- [17.5: Conductors in Electrostatic Equilibrium via Gauss's Law](#)
- [17.6: Gauss's Law \(Summary\)](#)
- [17.7: Gauss's Law \(Exercises\)](#)
- [17.8: Gauss's Law \(Answers\)](#)

17: Gauss's Law for Calculation of Electrical Field from Charge Distributions is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

17.1: Introduction to Gauss's Law

Flux is a general and broadly applicable concept in physics. However, in this chapter, we concentrate on the flux of the electric field. This allows us to introduce Gauss's law, which is particularly useful for finding the electric fields of charge distributions exhibiting spatial symmetry. The main topics discussed here are

1. **Electric flux.** We define electric flux for both open and closed surfaces.
2. **Gauss's law.** We derive Gauss's law for an arbitrary charge distribution and examine the role of electric flux in Gauss's law.
3. **Calculating electric fields with Gauss's law.** The main focus of this chapter is to explain how to use Gauss's law to find the electric fields of spatially symmetrical charge distributions. We discuss the importance of choosing a Gaussian surface and provide examples involving the applications of Gauss's law.
4. **Electric fields in conductors.** Gauss's law provides useful insight into the absence of electric fields in conducting materials.



Figure 17.1.1: This chapter introduces the concept of flux, which relates a physical quantity and the area through which it is flowing. Although we introduce this concept with the electric field, the concept may be used for many other quantities, such as fluid flow. (credit: modification of work by "Alessandro"/Flickr)

So far, we have found that the electrostatic field begins and ends at point charges and that the field of a point charge varies inversely with the square of the distance from that charge. These characteristics of the electrostatic field lead to an important mathematical relationship known as Gauss's law. This law is named in honor of the extraordinary German mathematician and scientist Karl Friedrich Gauss (Figure 17.1.2). Gauss's law gives us an elegantly simple way of finding the electric field, and, as you will see, it can be much easier to use than the integration method described in the previous chapter. However, there is a catch—Gauss's law has a limitation in that, while always true, it can be readily applied only for charge distributions with certain symmetries.



Figure 17.1.2: Karl Friedrich Gauss (1777–1855) was a legendary mathematician of the nineteenth century. Although his major contributions were to the field of mathematics, he also did important work in physics and astronomy.

This page titled [17.1: Introduction to Gauss's Law](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.1: Prelude to Gauss's Law](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

17.2: Electric Flux

Learning Objectives

By the end of this section, you will be able to:

- Define the concept of flux
- Describe electric flux
- Calculate electric flux for a given situation

The concept of **flux** describes how much of something goes through a given area. More formally, it is the dot product of a vector field (in this chapter, the electric field) with an area. You may conceptualize the flux of an electric field as a measure of the number of electric field lines passing through an area (Figure 17.2.1). The larger the area, the more field lines go through it and, hence, the greater the flux; similarly, the stronger the electric field is (represented by a greater density of lines), the greater the flux. On the other hand, if the area rotated so that the plane is aligned with the field lines, none will pass through and there will be no flux.

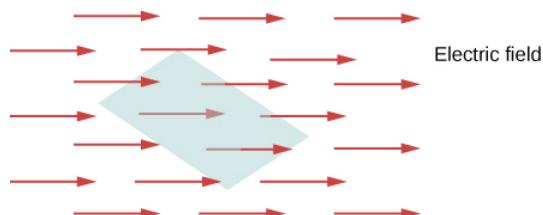


Figure 17.2.1: The flux of an electric field through the shaded area captures information about the “number” of electric field lines passing through the area. The numerical value of the electric flux depends on the magnitudes of the electric field and the area, as well as the relative orientation of the area with respect to the direction of the electric field.

A macroscopic analogy that might help you imagine this is to put a hula hoop in a flowing river. As you change the angle of the hoop relative to the direction of the current, more or less of the flow will go through the hoop. Similarly, the amount of flow through the hoop depends on the strength of the current and the size of the hoop. Again, flux is a general concept; we can also use it to describe the amount of sunlight hitting a solar panel or the amount of energy a telescope receives from a distant star, for example.

To quantify this idea, Figure 17.2.1a shows a planar surface S_1 of area A_1 that is perpendicular to the uniform electric field $\vec{E} = E\hat{j}$. If N field lines pass through S_1 , then we know from the definition of electric field lines ([Electric Charges and Fields](#)) that $N/A \propto E$, or $N \propto EA_1$.

The quantity EA_1 is the **electric flux** through S_1 . We represent the electric flux through an open surface like S_1 by the symbol Φ . Electric flux is a scalar quantity and has an SI unit of newton-meters squared per coulomb ($N \cdot m^2/C$). Notice that $N \propto EA_1$ may also be written as $N \propto \Phi$, demonstrating that **electric flux is a measure of the number of field lines crossing a surface**.

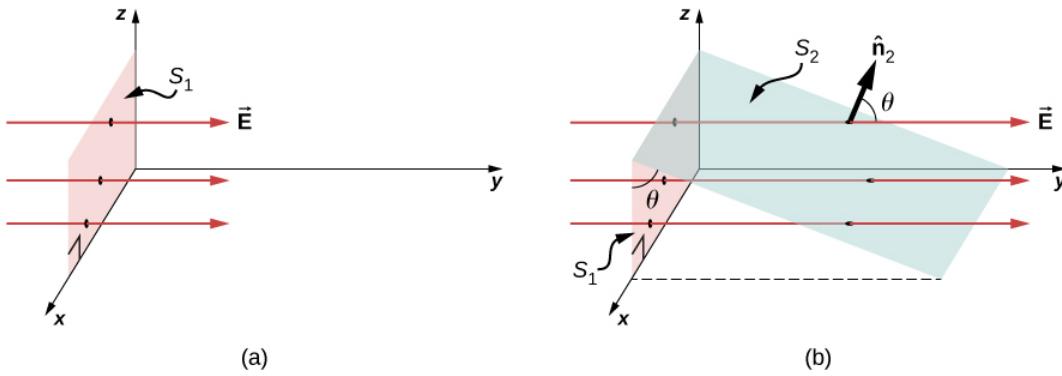


Figure 17.2.2: (a) A planar surface S_1 of area A_1 is perpendicular to the electric field $E\hat{j}$. N field lines cross surface S_1 . (b) A surface S_2 of area A_2 whose projection onto the xz -plane is S_1 . The same number of field lines cross each surface.

Now consider a planar surface that is not perpendicular to the field. How would we represent the electric flux? Figure 17.2.2b shows a surface S_2 of area A_2 that is inclined at an angle θ to the xz -plane and whose projection in that plane is S_1 (area A_1). The areas are

related by $A_2 \cos \theta = A_1$. Because the same number of field lines crosses both S_1 and S_2 , the fluxes through both surfaces must be the same. The flux through S_2 is therefore $\Phi = EA_1 = EA_2 \cos \theta$. Designating \hat{n}_2 as a unit vector normal to S_2 (see Figure 17.2.2b), we obtain

$$\Phi = \vec{E} \cdot \hat{n}_2 A_2. \quad (17.2.1)$$

Note

Check out this [video](#) to observe what happens to the flux as the area changes in size and angle, or the electric field changes in strength.

Area Vector

For discussing the flux of a vector field, it is helpful to introduce an area vector \vec{A} . This allows us to write the last equation in a more compact form. What should the magnitude of the area vector be? What should the direction of the area vector be? What are the implications of how you answer the previous question?

The **area vector** of a flat surface of area A has the following magnitude and direction:

- Magnitude is equal to area (A)
- Direction is along the normal to the surface (\hat{n}); that is, perpendicular to the surface.

Since the normal to a flat surface can point in either direction from the surface, the direction of the area vector of an open surface needs to be chosen, as shown in Figure 17.2.3.

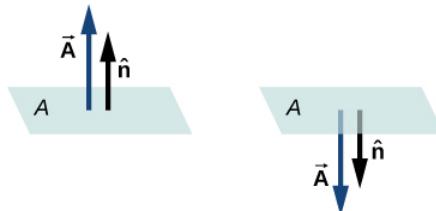


Figure 17.2.3: The direction of the area vector of an open surface needs to be chosen; it could be either of the two cases displayed here. The area vector of a part of a closed surface is defined to point from the inside of the closed space to the outside. This rule gives a unique direction.

Since \hat{n} is a unit normal to a surface, it has two possible directions at every point on that surface (Figure 17.2.1a). For an open surface, we can use either direction, as long as we are consistent over the entire surface. 17.2.1c of the figure shows several cases.

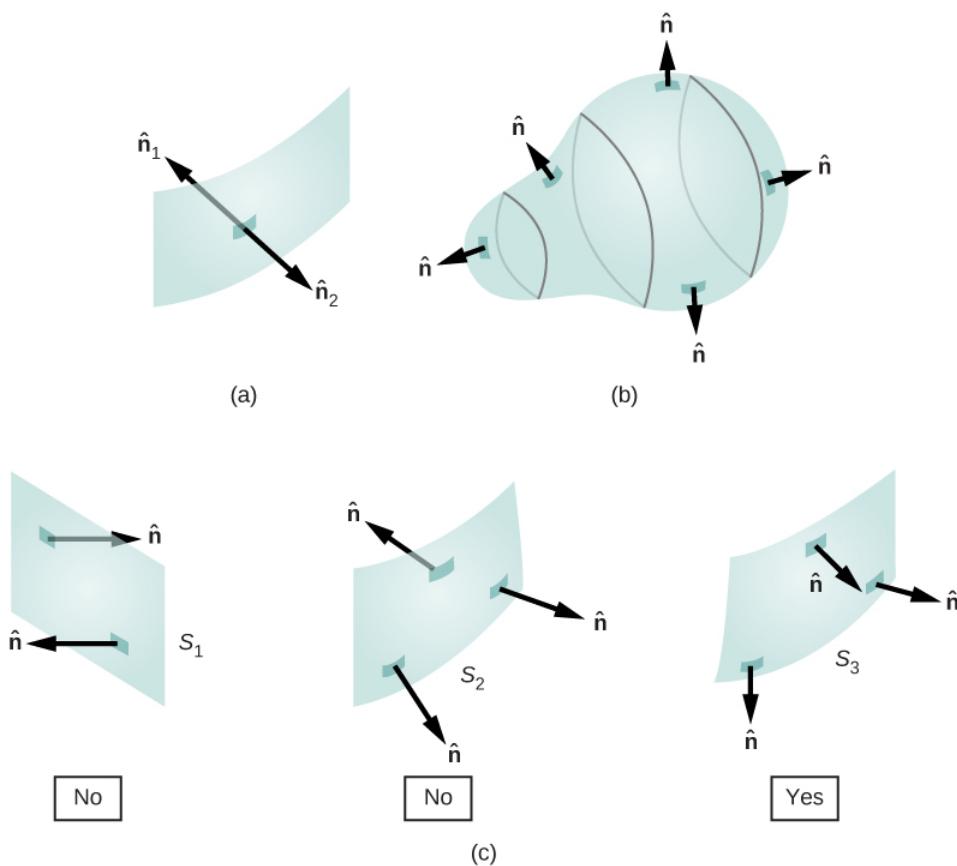


Figure 17.2.4: (a) Two potential normal vectors arise at every point on a surface. (b) The outward normal is used to calculate the flux through a closed surface. (c) Only S_3 has been given a consistent set of normal vectors that allows us to define the flux through the surface.

However, if a surface is closed, then the surface encloses a volume. In that case, the direction of the **normal vector** at any point on the surface points from the inside to the outside. On a **closed surface** such as that of Figure 17.2.1b, \hat{n} is chosen to be the **outward normal** at every point, to be consistent with the sign convention for electric charge.

Electric Flux

Now that we have defined the area vector of a surface, we can define the electric flux of a uniform electric field through a flat area as the scalar product of the electric field and the area vector:

$$\Phi = \vec{E} \cdot \vec{A} \text{ (uniform } \vec{E}, \text{ flat surface).} \quad (17.2.2)$$

Figure 17.2.5 shows the electric field of an oppositely charged, parallel-plate system and an imaginary box between the plates. The electric field between the plates is uniform and points from the positive plate toward the negative plate. A calculation of the flux of this field through various faces of the box shows that the net flux through the box is zero. Why does the flux cancel out here?

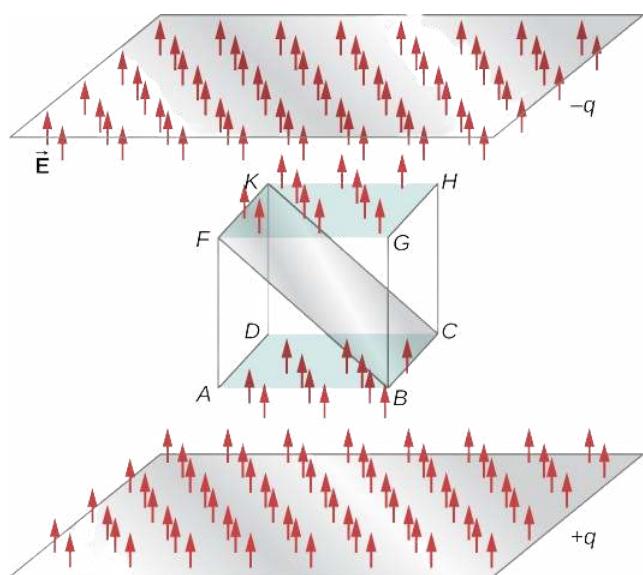


Figure 17.2.5: Electric flux through a cube, placed between two charged plates. Electric flux through the bottom face ($ABCD$) is negative, because \vec{E} is in the opposite direction to the normal to the surface. The electric flux through the top face ($FGHK$) is positive, because the electric field and the normal are in the same direction. The electric flux through the other faces is zero, since the electric field is perpendicular to the normal vectors of those faces. The net electric flux through the cube is the sum of fluxes through the six faces. Here, the net flux through the cube is equal to zero. The magnitude of the flux through rectangle $BCKF$ is equal to the magnitudes of the flux through both the top and bottom faces.

The reason is that the sources of the electric field are outside the box. Therefore, if any electric field line enters the volume of the box, it must also exit somewhere on the surface because there is no charge inside for the lines to land on. Therefore, quite generally, electric flux through a closed surface is zero if there are no sources of electric field, whether positive or negative charges, inside the enclosed volume. In general, when field lines leave (or “flow out of”) a closed surface, Φ is positive; when they enter (or “flow into”) the surface, Φ is negative.

Any smooth, non-flat surface can be replaced by a collection of tiny, approximately flat surfaces, as shown in Figure 17.2.6. If we divide a surface S into small patches, then we notice that, as the patches become smaller, they can be approximated by flat surfaces. This is similar to the way we treat the surface of Earth as locally flat, even though we know that globally, it is approximately spherical.

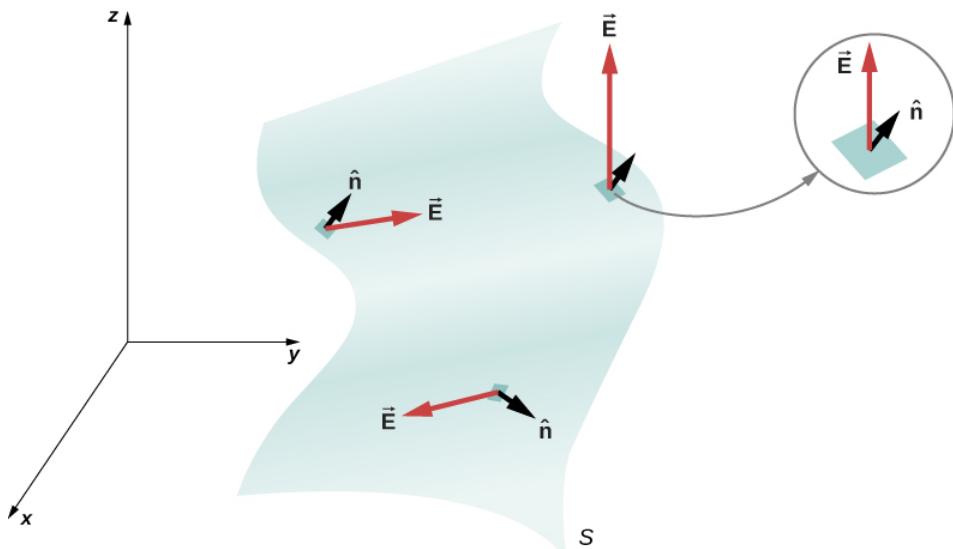


Figure 17.2.6: A surface is divided into patches to find the flux.

To keep track of the patches, we can number them from 1 through N . Now, we define the area vector for each patch as the area of the patch pointed in the direction of the normal. Let us denote the area vector for the i th patch by $\delta \vec{A}_i$. (We have used the symbol δ

to remind us that the area is of an arbitrarily small patch.) With sufficiently small patches, we may approximate the electric field over any given patch as uniform. Let us denote the average electric field at the location of the i th patch by \vec{E}_i .

$$\vec{E}_i = \text{average electric field over the } i\text{th patch.} \quad (17.2.3)$$

Therefore, we can write the electric flux Φ through the area of the i th patch as

$$\Phi_i = \vec{E}_i \cdot \delta \vec{A}_i \text{ (} i\text{th patch).} \quad (17.2.4)$$

The flux through each of the individual patches can be constructed in this manner and then added to give us an estimate of the net flux through the entire surface S , which we denote simply as Φ .

$$\Phi = \sum_{i=1}^N \Phi_i = \sum_{i=1}^N \vec{E}_i \cdot \delta \vec{A}_i \text{ (} N \text{ patch estimate).} \quad (17.2.5)$$

This estimate of the flux gets better as we decrease the size of the patches. However, when you use smaller patches, you need more of them to cover the same surface. In the limit of infinitesimally small patches, they may be considered to have area dA and unit normal \hat{n} . Since the elements are infinitesimal, they may be assumed to be planar, and \vec{E}_i may be taken as constant over any element. Then the flux $d\Phi$ through an area dA is given by $d\Phi = \vec{E} \cdot \hat{n} dA$. It is positive when the angle between \vec{E}_i and \hat{n} is less than 90° and negative when the angle is greater than 90° . The net flux is the sum of the infinitesimal flux elements over the entire surface. With infinitesimally small patches, you need infinitely many patches, and the limit of the sum becomes a surface integral. With \int_S representing the integral over S ,

$$\Phi = \int_S \vec{E} \cdot \hat{n} dA = \int_S \vec{E} \cdot d\vec{A} \text{ (open surface).} \quad (17.2.6)$$

In practical terms, surface integrals are computed by taking the antiderivatives of both dimensions defining the area, with the edges of the surface in question being the bounds of the integral.

To distinguish between the flux through an open surface like that of Figure 17.2.2 and the flux through a closed surface (one that completely bounds some volume), we represent flux through a closed surface by

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \oint_S \vec{E} \cdot d\vec{A} \text{ (closed surface)} \quad (17.2.7)$$

where the circle through the integral symbol simply means that the surface is closed, and we are integrating over the entire thing. If you only integrate over a portion of a closed surface, that means you are treating a subset of it as an open surface.

✓ Example 17.2.1: Flux of a Uniform Electric Field

A constant electric field of magnitude E_0 points in the direction of the positive z -axis (Figure 17.2.7). What is the electric flux through a rectangle with sides a and b in the (a) xy -plane and in the (b) xz -plane?

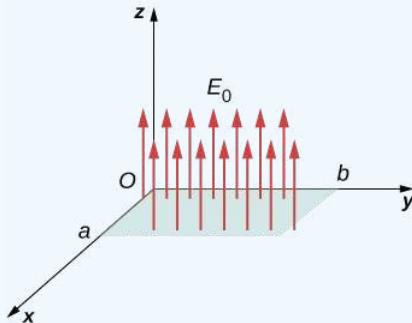


Figure 17.2.7: Calculating the flux of E_0 through a rectangular surface.

Strategy

Apply the definition of flux: $\Phi = \vec{E} \cdot \vec{A}$ (*uniform* \vec{E}), where the definition of dot product is crucial.

Solution

1. In this case, $\Phi = \vec{E}_0 \cdot \vec{A} = E_0 A = E_0 ab$.
2. Here, the direction of the area vector is either along the positive y -axis or toward the negative y -axis. Therefore, the scalar product of the electric field with the area vector is zero, giving zero flux.

Significance

The relative directions of the electric field and area can cause the flux through the area to be zero.

✓ Flux of a Uniform Electric Field through a Closed Surface

A constant electric field of magnitude E_0 points in the direction of the positive z -axis (Figure 17.2.8). What is the net electric flux through a cube?

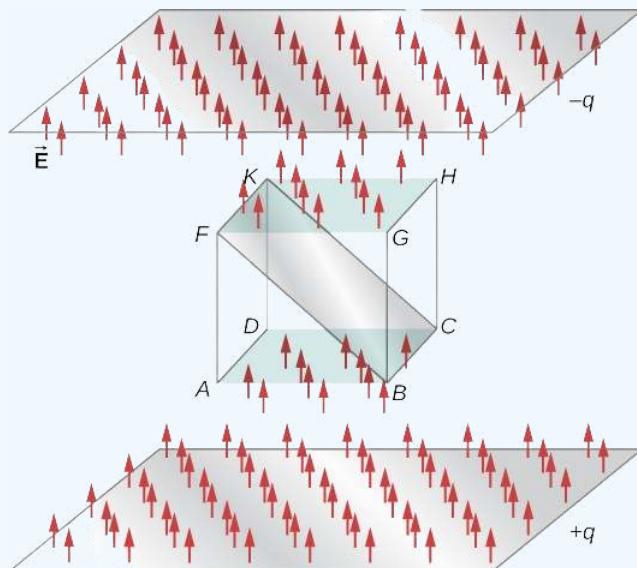


Figure 17.2.8: Calculating the flux of E_0 through a closed cubic surface.

Strategy

Apply the definition of flux: $\Phi = \vec{E} \cdot \vec{A}$ (*uniform* \vec{E}), noting that a closed surface eliminates the ambiguity in the direction of the area vector.

Solution

Through the top face of the cube $\Phi = \vec{E}_0 \cdot \vec{A} = E_0 A$.

Through the bottom face of the cube, $\Phi = \vec{E}_0 \cdot \vec{A} = -E_0 A$, because the area vector here points downward.

Along the other four sides, the direction of the area vector is perpendicular to the direction of the electric field. Therefore, the scalar product of the electric field with the area vector is zero, giving zero flux.

The net flux is $\Phi_{net} = E_0 A - E_0 A + 0 + 0 + 0 + 0 = 0$.

Significance

The net flux of a uniform electric field through a closed surface is zero.

✓ Example 17.2.3: Electric Flux through a Plane, Integral Method

A uniform electric field \vec{E} of magnitude 10 N/C is directed parallel to the yz -plane at 30° above the xy -plane, as shown in Figure 17.2.9. What is the electric flux through the plane surface of area 6.0 m^2 located in the xz -plane? Assume that \hat{n} points in the positive y -direction.

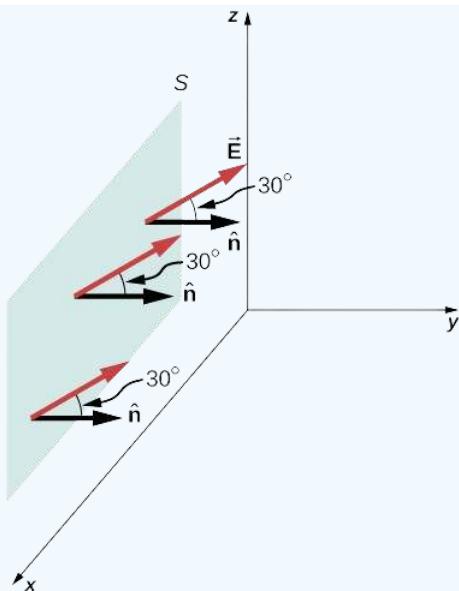


Figure 17.2.9: The electric field produces a net electric flux through the surface S .

Strategy

Apply $\Phi = \int_S \vec{E} \cdot \hat{n} dA$, where the direction and magnitude of the electric field are constant.

Solution

The angle between the uniform electric field \vec{E} and the unit normal \hat{n} to the planar surface is 30° . Since both the direction and magnitude are constant, E comes outside the integral. All that is left is a surface integral over dA , which is A . Therefore, using the open-surface equation, we find that the electric flux through the surface is

$$\Phi = \int_S \vec{E} \cdot \hat{n} dA = EA \cos \theta \quad (17.2.8)$$

$$= (10 \text{ N/C})(6.0 \text{ m}^2)(\cos 30^\circ) = 52 \text{ N} \cdot \text{m}^2/\text{C}. \quad (17.2.9)$$

Significance

Again, the relative directions of the field and the area matter, and the general equation with the integral will simplify to the simple dot product of area and electric field.

Exercise 17.2.1

What angle should there be between the electric field and the surface shown in Figure 17.2.9 in the previous example so that no electric flux passes through the surface?

Solution

Place it so that its unit normal is perpendicular to \vec{E} .

✓ Example 17.2.4 : Inhomogeneous Electric Field

What is the total flux of the electric field $\vec{E} = cy^2 \hat{k}$ through the rectangular surface shown in Figure 17.2.10?

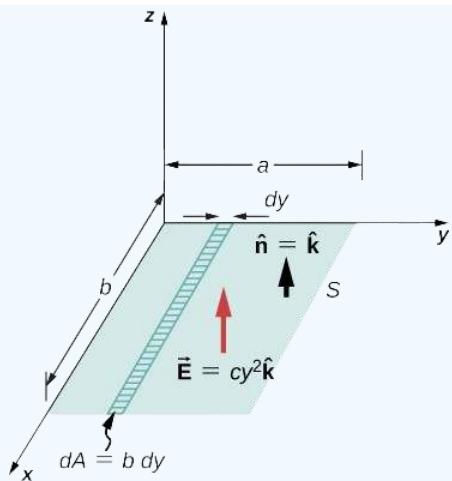


Figure 17.2.10: Since the electric field is not constant over the surface, an integration is necessary to determine the flux.

Strategy

Apply $\Phi = \int_S \vec{E} \cdot \hat{n} dA$. We assume that the unit normal \hat{n} to the given surface points in the positive z -direction, so $\hat{n} = \hat{k}$. Since the electric field is not uniform over the surface, it is necessary to divide the surface into infinitesimal strips along which \vec{E} is essentially constant. As shown in Figure 17.2.10, these strips are parallel to the x -axis, and each strip has an area $dA = b dy$.

Solution

From the open surface integral, we find that the net flux through the rectangular surface is

$$\begin{aligned}\Phi &= \int_S \vec{E} \cdot \hat{n} dA = \int_0^a (cy^2 \hat{k}) \cdot \hat{k} (b dy) \\ &= cb \int_0^a y^2 dy = \frac{1}{3} a^3 bc.\end{aligned}$$

Significance

For a non-constant electric field, the integral method is required.

Exercise 17.2.1

If the electric field in Example 17.2.4 is $\vec{E} = mx\hat{k}$, what is the flux through the rectangular area?

Answer

$$mab^2/2$$

This page titled [17.2: Electric Flux](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.2: Electric Flux](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

17.3: Gauss's Law

Learning Objectives

By the end of this section, you will be able to:

- State Gauss's law
- Explain the conditions under which Gauss's law may be used
- Apply Gauss's law in appropriate systems

We can now determine the electric flux through an arbitrary closed surface due to an arbitrary charge distribution. We found that if a closed surface does not have any charge inside where an electric field line can terminate, then any electric field line entering the surface at one point must necessarily exit at some other point of the surface. Therefore, if a closed surface does not have any charges inside the enclosed volume, then the electric flux through the surface is zero. Now, what happens to the electric flux if there are some charges inside the enclosed volume? Gauss's law gives a quantitative answer to this question.

To get a feel for what to expect, let's calculate the electric flux through a spherical surface around a positive point charge q , since we already know the electric field in such a situation. Recall that when we place the point charge at the origin of a coordinate system, the electric field at a point P that is at a distance r from the charge at the origin is given by

$$\vec{E}_P = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r}, \quad (17.3.1)$$

where \hat{r} is the radial vector from the charge at the origin to the point P . We can use this electric field to find the flux through the spherical surface of radius r , as shown in Figure 17.3.1.

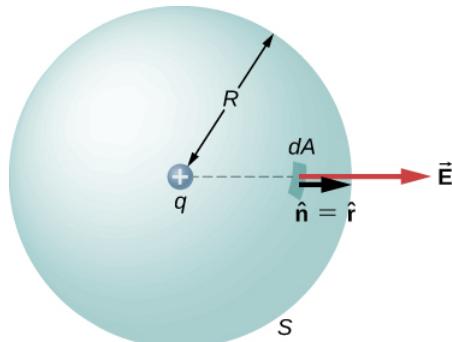


Figure 17.3.1: A closed spherical surface surrounding a point charge q .

Then we apply $\Phi = \int_S \vec{E} \cdot \hat{n} dA$ to this system and substitute known values. On the sphere, \hat{n} and $r = R$ so for an infinitesimal area dA ,

$$\begin{aligned} d\Phi &= \vec{E} \cdot \hat{n} dA \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} \hat{r} \cdot \hat{r} dA \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} dA. \end{aligned}$$

We now find the net flux by integrating this flux over the surface of the sphere:

$$\Phi = \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} \oint_S dA = \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} (4\pi R^2) = \frac{q}{\epsilon_0}. \quad (17.3.2)$$

where the total surface area of the spherical surface is $4\pi R^2$. This gives the flux through the closed spherical surface at radius r as

$$\Phi = \frac{q}{\epsilon_0}. \quad (17.3.3)$$

A remarkable fact about this equation is that the flux is independent of the size of the spherical surface. This can be directly attributed to the fact that the electric field of a point charge decreases as $1/r^2$ with distance, which just cancels the r^2 rate of increase of the surface area.

Electric Field Lines Picture

An alternative way to see why the flux through a closed spherical surface is independent of the radius of the surface is to look at the electric field lines. Note that every field line from q that pierces the surface at radius R_1 also pierces the surface at R_2 (Figure 17.3.2).

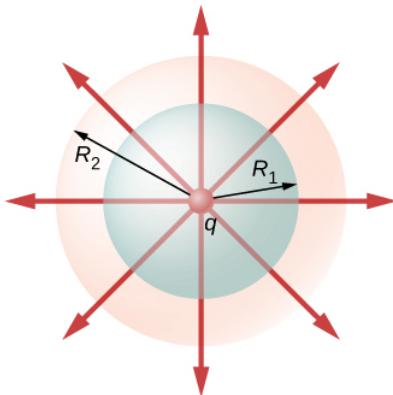


Figure 17.3.2: Flux through spherical surfaces of radii R_1 and R_2 enclosing a charge q are equal, independent of the size of the surface, since all \vec{E} -field lines that pierce one surface from the inside to outside direction also pierce the other surface in the same direction.

Therefore, the net number of electric field lines passing through the two surfaces from the inside to outside direction is equal. This net number of electric field lines, which is obtained by subtracting the number of lines in the direction from outside to inside from the number of lines in the direction from inside to outside gives a visual measure of the electric flux through the surfaces.

You can see that if no charges are included within a closed surface, then the electric flux through it must be zero. A typical field line enters the surface at dA_1 and leaves at dA_2 . Every line that enters the surface must also leave that surface. Hence the net “flow” of the field lines into or out of the surface is zero (Figure 17.3.3a). The same thing happens if charges of equal and opposite sign are included inside the closed surface, so that the total charge included is zero (Figure 17.3.3b). A surface that includes the same amount of charge has the same number of field lines crossing it, regardless of the shape or size of the surface, as long as the surface encloses the same amount of charge (Figure 17.3.3c).

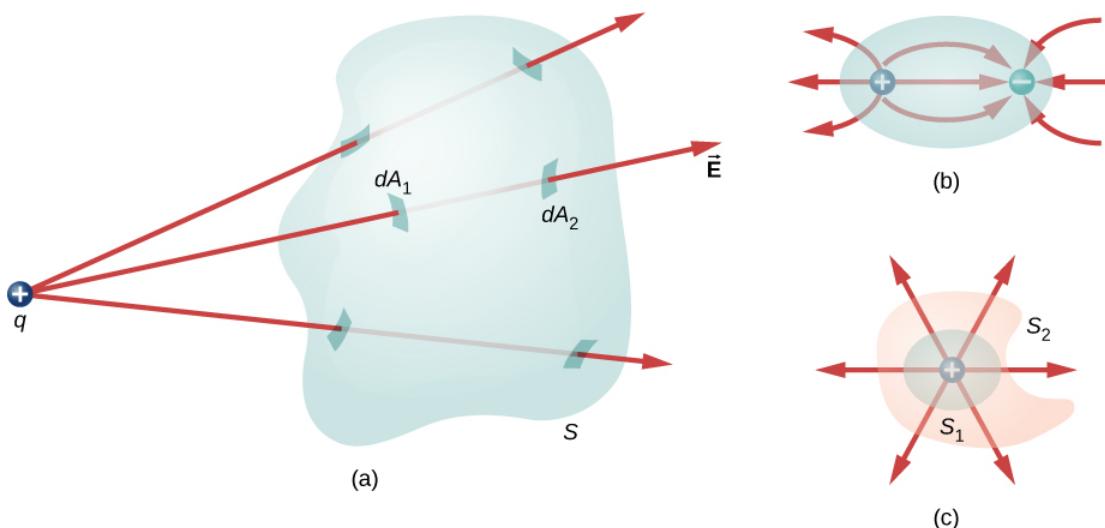


Figure 17.3.3: Understanding the flux in terms of field lines. (a) The electric flux through a closed surface due to a charge outside that surface is zero. (b) Charges are enclosed, but because the net charge included is zero, the net flux through the closed surface is also zero. (c) The shape and size of the surfaces that enclose a charge does not matter because all surfaces enclosing the same charge have the same flux.

Statement of Gauss's Law

Gauss's law generalizes this result to the case of any number of charges and any location of the charges in the space inside the closed surface. According to Gauss's law, the flux of the electric field \vec{E} through any closed surface, also called a **Gaussian surface**, is equal to the net charge enclosed (q_{enc}) divided by the permittivity of free space (ϵ_0):

$$\Phi_{\text{ClosedSurface}} = \frac{q_{enc}}{\epsilon_0}. \quad (17.3.4)$$

This equation holds for **charges of either sign**, because we define the area vector of a closed surface to point outward. If the enclosed charge is negative (Figure 17.3.4b), then the flux through either S or S' is negative.

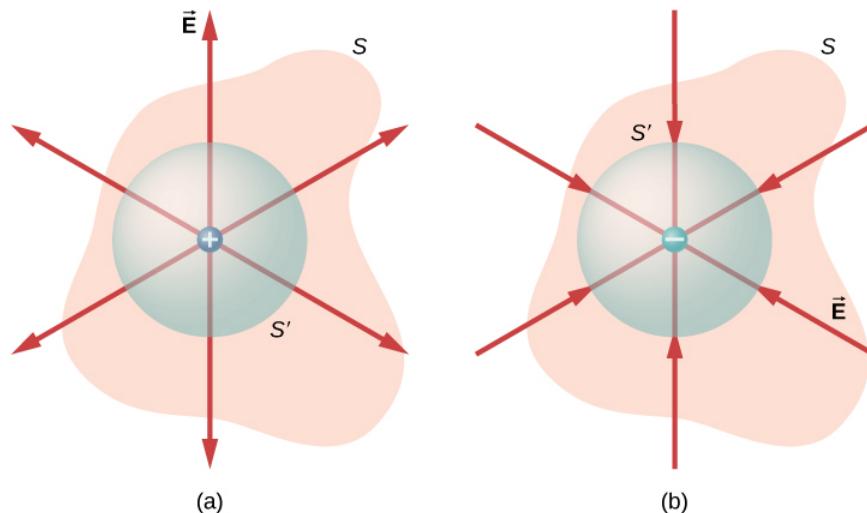


Figure 17.3.4: The electric flux through any closed surface surrounding a point charge q is given by Gauss's law. (a) Enclosed charge is positive. (b) Enclosed charge is negative.

The Gaussian surface does not need to correspond to a real, physical object; indeed, it rarely will. It is a mathematical construct that may be of any shape, provided that it is closed. However, since our goal is to integrate the flux over it, we tend to choose shapes that are highly symmetrical.

If the charges are discrete point charges, then we just add them. If the charge is described by a continuous distribution, then we need to integrate appropriately to find the total charge that resides inside the enclosed volume. For example, the flux through the Gaussian surface S of Figure 17.3.5 is

$$\Phi = (q_1 + q_2 + q_5)/\epsilon_0. \quad (17.3.5)$$

Note that q_{enc} is simply the sum of the point charges. If the charge distribution were continuous, we would need to integrate appropriately to compute the total charge within the Gaussian surface.

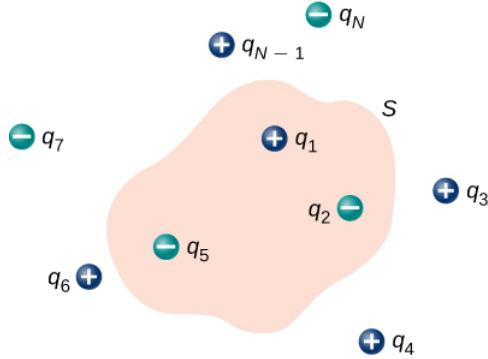


Figure 17.3.5: The flux through the Gaussian surface shown, due to the charge distribution, is $\Phi = (q_1 + q_2 + q_5)/\epsilon_0$.

Recall that the principle of superposition holds for the electric field. Therefore, the total electric field at any point, including those on the chosen Gaussian surface, is the sum of all the electric fields present at this point. This allows us to write Gauss's law in

terms of the total electric field.

Gauss's Law

The flux Φ of the electric field \vec{E} through any closed surface S (a Gaussian surface) is equal to the net charge enclosed (q_{enc}) divided by the permittivity of free space (ϵ_0):

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \frac{q_{enc}}{\epsilon_0}. \quad (17.3.6)$$

To use Gauss's law effectively, you must have a clear understanding of what each term in the equation represents. The field \vec{E} is the **total electric field** at every point on the Gaussian surface. This total field includes contributions from charges both inside and outside the Gaussian surface. However, q_{enc} is just the charge **inside** the Gaussian surface. Finally, the Gaussian surface is any closed surface in space. That surface can coincide with the actual surface of a conductor, or it can be an imaginary geometric surface. The only requirement imposed on a Gaussian surface is that it be closed (Figure 17.3.5).



Figure 17.3.6: A Klein bottle partially filled with a liquid. Could the Klein bottle be used as a Gaussian surface?

Example 17.3.1: Electric Flux through Gaussian Surfaces

Calculate the electric flux through each Gaussian surface shown in Figure 17.3.7.

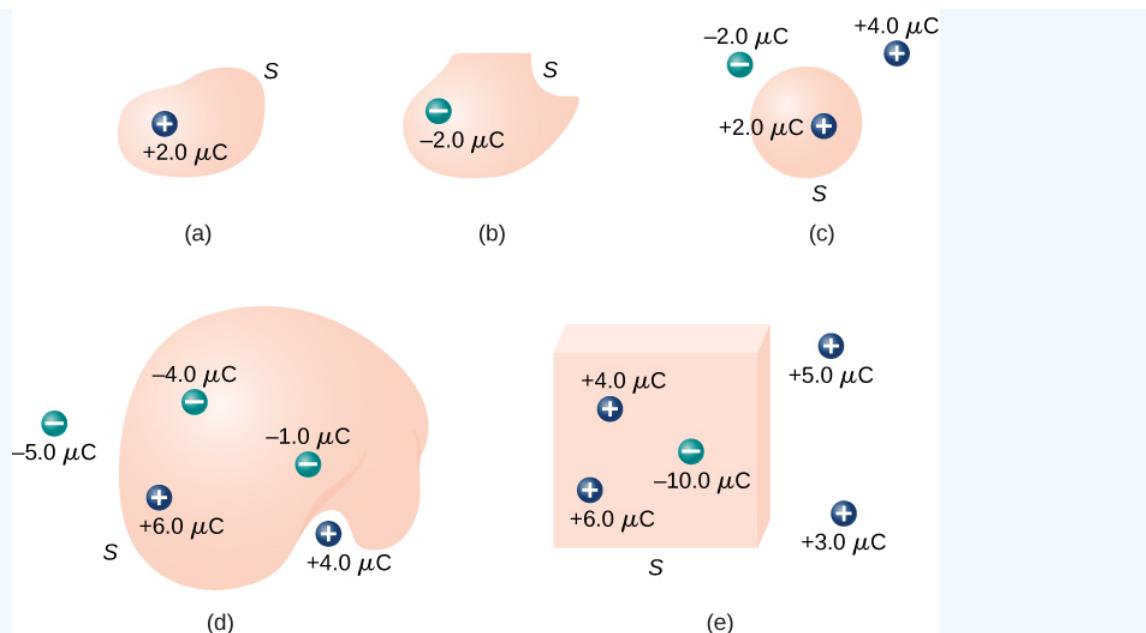


Figure 17.3.7: Various Gaussian surfaces and charges.

Strategy

From Gauss's law, the flux through each surface is given by q_{enc}/ϵ_0 , where q_{enc} is the charge enclosed by that surface.

Solution

For the surfaces and charges shown, we find

a. $\Phi = \frac{2.0 \mu C}{\epsilon_0} = 2.3 \times 10^5 N \cdot m^2/C$.

b. $\Phi = \frac{-2.0 \mu C}{\epsilon_0} = -2.3 \times 10^5 N \cdot m^2/C$.

c. $\Phi = \frac{2.0 \mu C}{\epsilon_0} = 2.3 \times 10^5 N \cdot m^2/C$.

d. $\frac{-4.0 \mu C + 6.0 \mu C - 1.0 \mu C}{\epsilon_0} = 1.1 \times 10^5 N \cdot m^2/C$.

e. $\frac{4.0 \mu C + 6.0 \mu C - 10.0 \mu C}{\epsilon_0} = 0$.

Significance

In the special case of a closed surface, the flux calculations become a sum of charges. In the next section, this will allow us to work with more complex systems.

Exercise 17.3.1

Calculate the electric flux through the closed cubical surface for each charge distribution shown in Figure 17.3.8.

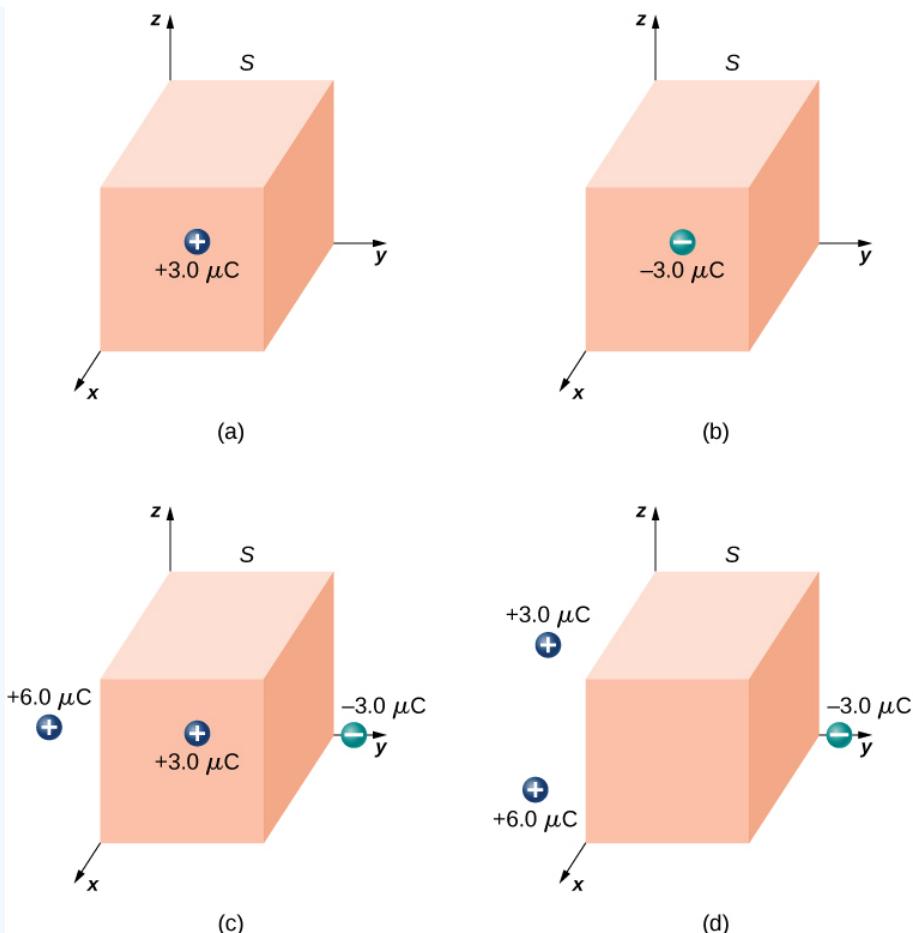


Figure 17.3.8: A cubical Gaussian surface with various charge distributions.

Answer a

$$3.4 \times 10^5 N \cdot m^2/C$$

Answer b

$$-3.4 \times 10^5 N \cdot m^2/C$$

Answer c

$$3.4 \times 10^5 N \cdot m^2/C$$

Answer d

$$0$$

Open Source Physics Simulation

Use this [simulation](#) to adjust the magnitude of the charge and the radius of the Gaussian surface around it. See how this affects the total flux and the magnitude of the electric field at the Gaussian surface.

This page titled [17.3: Gauss's Law](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.3: Explaining Gauss's Law](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

17.4: Calculating Electric Field Using Gauss's Law

Learning Objectives

By the end of this section, you will be able to:

- Explain what spherical, cylindrical, and planar symmetry are
- Recognize whether or not a given system possesses one of these symmetries
- Apply Gauss's law to determine the electric field of a system with one of these symmetries

Gauss's law is very helpful in determining expressions for the electric field, even though the law is not directly about the electric field; it is about the electric flux. It turns out that in situations that have certain symmetries (spherical, cylindrical, or planar) in the charge distribution, we can deduce the electric field based on knowledge of the electric flux. In these systems, we can find a Gaussian surface S over which the electric field has constant magnitude. Furthermore, if \vec{E} is parallel to \hat{n} everywhere on the surface, then $\vec{E} \cdot \hat{n} = E$. (If \vec{E} and \hat{n} are antiparallel everywhere on the surface, $\vec{E} \cdot \hat{n} = -E$.) Gauss's law then simplifies to

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = E \oint_S dA = EA = \frac{q_{enc}}{\epsilon_0}, \quad (17.4.1)$$

where A is the area of the surface. Note that these symmetries lead to the transformation of the flux integral into a product of the magnitude of the electric field and an appropriate area. When you use this flux in the expression for Gauss's law, you obtain an algebraic equation that you can solve for the magnitude of the electric field, which looks like

$$E \approx \frac{q_{enc}}{\epsilon_0 \text{ area}}. \quad (17.4.2)$$

The direction of the electric field at point P is obtained from the symmetry of the charge distribution and the type of charge in the distribution. Therefore, Gauss's law can be used to determine \vec{E} . Here is a summary of the steps we will follow:

Problem-Solving Strategy: Gauss's Law

1. **Identify the spatial symmetry of the charge distribution.** This is an important first step that allows us to choose the appropriate Gaussian surface. As examples, an isolated point charge has spherical symmetry, and an infinite line of charge has cylindrical symmetry.
2. **Choose a Gaussian surface with the same symmetry as the charge distribution and identify its consequences.** With this choice, $\vec{E} \cdot \hat{n}$ is easily determined over the Gaussian surface.
3. **Evaluate the integral $\oint_S \vec{E} \cdot \hat{n} dA$ over the Gaussian surface, that is, calculate the flux through the surface.** The symmetry of the Gaussian surface allows us to factor $\vec{E} \cdot \hat{n}$ outside the integral.
4. **Determine the amount of charge enclosed by the Gaussian surface.** This is an evaluation of the right-hand side of the equation representing Gauss's law. It is often necessary to perform an integration to obtain the net enclosed charge.
5. **Evaluate the electric field of the charge distribution.** The field may now be found using the results of steps 3 and 4.

Basically, there are only three types of symmetry that allow Gauss's law to be used to deduce the electric field. They are

- A charge distribution with spherical symmetry
- A charge distribution with cylindrical symmetry
- A charge distribution with planar symmetry

To exploit the symmetry, we perform the calculations in appropriate coordinate systems and use the right kind of Gaussian surface for that symmetry, applying the remaining four steps.

Charge Distribution with Spherical Symmetry

A charge distribution has **spherical symmetry** if the density of charge depends only on the distance from a point in space and not on the direction. In other words, if you rotate the system, it doesn't look different. For instance, if a sphere of radius R is uniformly charged with charge density ρ_0 then the distribution has spherical symmetry (Figure 17.4.1a). On the other hand, if a sphere of

radius R is charged so that the top half of the sphere has uniform charge density ρ_1 and the bottom half has a uniform charge density $\rho_2 \neq \rho_1$ then the sphere does not have spherical symmetry because the charge density depends on the direction (Figure 17.4.1b). Thus, it is not the shape of the object but rather the shape of the charge distribution that determines whether or not a system has spherical symmetry.

Figure 17.4.1c shows a sphere with four different shells, each with its own uniform charge density. Although this is a situation where charge density in the full sphere is not uniform, the charge density function depends only on the distance from the center and not on the direction. Therefore, this charge distribution does have spherical symmetry.

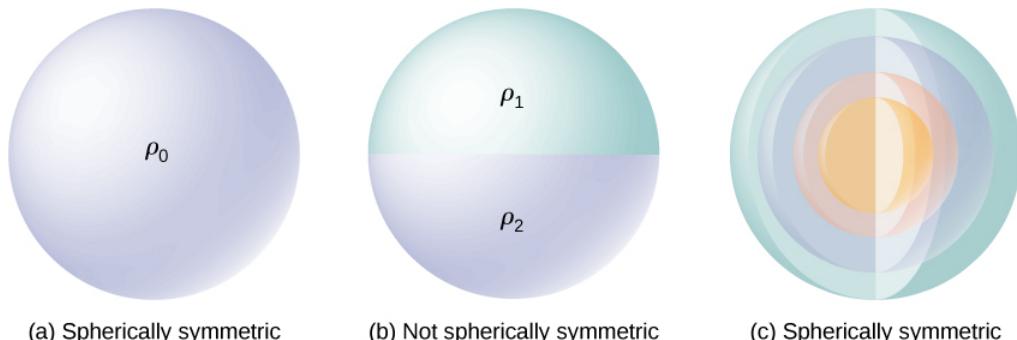


Figure 17.4.1: Illustrations of spherically symmetrical and nonsymmetrical systems. Different shadings indicate different charge densities. Charges on spherically shaped objects do not necessarily mean the charges are distributed with spherical symmetry. The spherical symmetry occurs only when the charge density does not depend on the direction. In (a), charges are distributed uniformly in a sphere. In (b), the upper half of the sphere has a different charge density from the lower half; therefore, (b) does not have spherical symmetry. In (c), the charges are in spherical shells of different charge densities, which means that charge density is only a function of the radial distance from the center; therefore, the system has spherical symmetry.

One good way to determine whether or not your problem has spherical symmetry is to look at the charge density function in spherical coordinates, $\rho(r, \theta, \phi)$. If the charge density is only a function of r , that is $\rho = \rho(r)$, then you have spherical symmetry. If the density depends on θ or ϕ , you could change it by rotation; hence, you would not have spherical symmetry.

Consequences of symmetry

In all spherically symmetrical cases, the electric field at any point must be radially directed, because the charge and, hence, the field must be invariant under rotation. Therefore, using spherical coordinates with their origins at the center of the spherical charge distribution, we can write down the expected form of the electric field at a point P located at a distance r from the center:

$$\text{Spherical symmetry : } \vec{E}_p = E_p(r)\hat{r}, \quad (17.4.3)$$

where \hat{r} is the unit vector pointed in the direction from the origin to the field point P . The radial component E_p of the electric field can be positive or negative. When $E_p > 0$, the electric field at P points away from the origin, and when $E_p < 0$, the electric field at P points toward the origin.

Gaussian surface and flux calculations

We can now use this form of the electric field to obtain the flux of the electric field through the Gaussian surface. For spherical symmetry, the Gaussian surface is a closed spherical surface that has the same center as the center of the charge distribution. Thus, the direction of the area vector of an area element on the Gaussian surface at any point is parallel to the direction of the electric field at that point, since they are both radially directed outward (Figure 17.4.2).

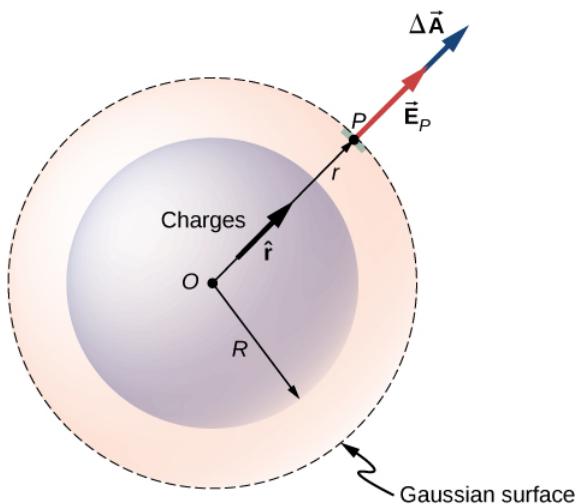


Figure 17.4.2: The electric field at any point of the spherical Gaussian surface for a spherically symmetrical charge distribution is parallel to the area element vector at that point, giving flux as the product of the magnitude of electric field and the value of the area. Note that the radius R of the charge distribution and the radius r of the Gaussian surface are different quantities.

The magnitude of the electric field \vec{E} must be the same everywhere on a spherical Gaussian surface concentric with the distribution. For a spherical surface of radius r :

$$\Phi = \oint_S \vec{E}_p \cdot \hat{n} dA = E_p \oint_S dA = E_p 4\pi r^2. \quad (17.4.4)$$

Using Gauss's law

According to Gauss's law, the flux through a closed surface is equal to the total charge enclosed within the closed surface divided by the permittivity of vacuum ϵ_0 . Let q_{enc} be the total charge enclosed inside the distance r from the origin, which is the space inside the Gaussian spherical surface of radius r . This gives the following relation for Gauss's law:

$$4\pi r^2 E = \frac{q_{enc}}{\epsilon_0}. \quad (17.4.5)$$

Hence, the electric field at point P that is a distance r from the center of a spherically symmetrical charge distribution has the following magnitude and direction:

$$\text{Magnitude : } E(r) = \frac{1}{4\pi\epsilon_0} \frac{q_{enc}}{r^2} \quad (17.4.6)$$

Direction: radial from O to P or from P to O .

The direction of the field at point P depends on whether the charge in the sphere is positive or negative. For a net positive charge enclosed within the Gaussian surface, the direction is from O to P , and for a net negative charge, the direction is from P to O . This is all we need for a point charge, and you will notice that the result above is identical to that for a point charge. However, Gauss's law becomes truly useful in cases where the charge occupies a finite volume.

Computing Enclosed Charge

The more interesting case is when a spherical charge distribution occupies a volume, and asking what the electric field inside the charge distribution is thus becomes relevant. In this case, the charge enclosed depends on the distance r of the field point relative to the radius of the charge distribution R , such as that shown in Figure 17.4.3.

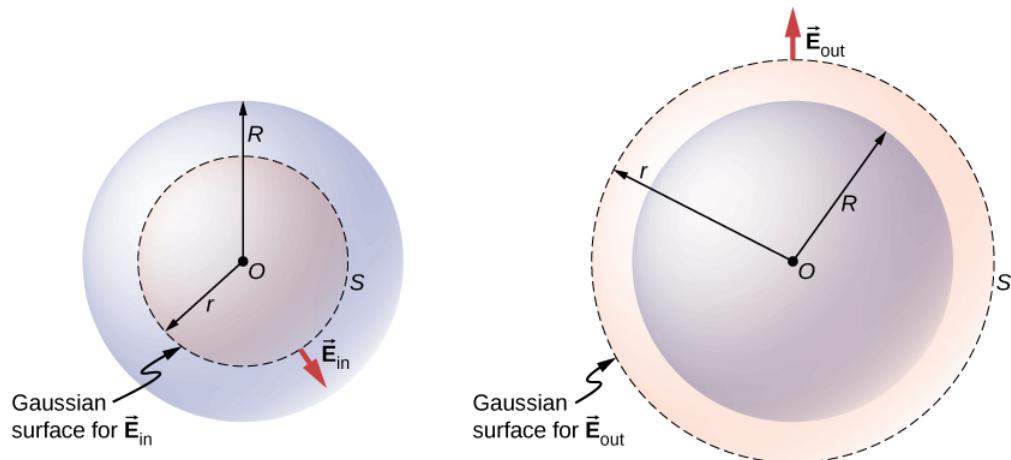


Figure 17.4.3: A spherically symmetrical charge distribution and the Gaussian surface used for finding the field (a) inside and (b) outside the distribution.

If point P is located outside the charge distribution—that is, if $r \geq R$ —then the Gaussian surface containing P encloses all charges in the sphere. In this case, q_{enc} equals the total charge in the sphere. On the other hand, if point P is within the spherical charge distribution, that is, if $r < R$, then the Gaussian surface encloses a smaller sphere than the sphere of charge distribution. In this case, q_{enc} is less than the total charge present in the sphere. Referring to Figure 17.4.3, we can write q_{enc} as

$$q_{enc} = q_{tot}(\text{total charge}) \text{ if } r \geq R \quad (17.4.7)$$

$$q_{enc} = q_{\text{within } r < R}(\text{only charge within } r < R) \text{ if } r < R \quad (17.4.8)$$

The field at a point outside the charge distribution is also called \vec{E}_{out} , and the field at a point inside the charge distribution is called \vec{E}_{in} . Focusing on the two types of field points, either inside or outside the charge distribution, we can now write the magnitude of the electric field as

$$\text{P outside sphere } E_{out} = \frac{1}{4\pi\epsilon_0} \frac{q_{tot}}{r^2} \quad (17.4.9)$$

$$\text{P inside sphere } E_{in} = \frac{1}{4\pi\epsilon_0} \frac{q_{\text{within } r < R}}{r^2}. \quad (17.4.10)$$

Note that the electric field outside a spherically symmetrical charge distribution is identical to that of a point charge at the center that has a charge equal to the total charge of the spherical charge distribution. This is remarkable since the charges are not located at the center only. We now work out specific examples of spherical charge distributions, starting with the case of a uniformly charged sphere.

✓ Uniformly Charged Sphere

A sphere of radius R , such as that shown in Figure 17.4.3, has a uniform volume charge density ρ_0 . Find the electric field at a point outside the sphere and at a point inside the sphere.

Strategy

Apply the Gauss's law problem-solving strategy, where we have already worked out the flux calculation.

Solution

The charge enclosed by the Gaussian surface is given by

$$q_{enc} = \int \rho_0 dV = \int_0^r \rho_0 4\pi r'^2 dr' = \rho_0 \left(\frac{4}{3}\pi r^3 \right). \quad (17.4.11)$$

The answer for electric field amplitude can then be written down immediately for a point outside the sphere, labeled E_{out} and a point inside the sphere, labeled E_{in} .

$$E_{out} = \frac{1}{4\pi\epsilon_0} \frac{q_{tot}}{r^2}, q_{tot} = \frac{4}{3}\pi R^3 \rho_0, \quad (17.4.12)$$

$$E_{in} = \frac{q_{enc}}{4\pi\epsilon_0 r^2} = \frac{\rho_0 r}{3\epsilon_0}, \text{ since } q_{enc} = \frac{4}{3}\pi r^3 \rho_0. \quad (17.4.13)$$

It is interesting to note that the magnitude of the electric field increases inside the material as you go out, since the amount of charge enclosed by the Gaussian surface increases with the volume. Specifically, the charge enclosed grows $\propto r^3$, whereas the field from each infinitesimal element of charge drops off $\propto 1/r^2$ with the net result that the electric field within the distribution increases in strength linearly with the radius. The magnitude of the electric field outside the sphere decreases as you go away from the charges, because the included charge remains the same but the distance increases. Figure 17.4.4 displays the variation of the magnitude of the electric field with distance from the center of a uniformly charged sphere.

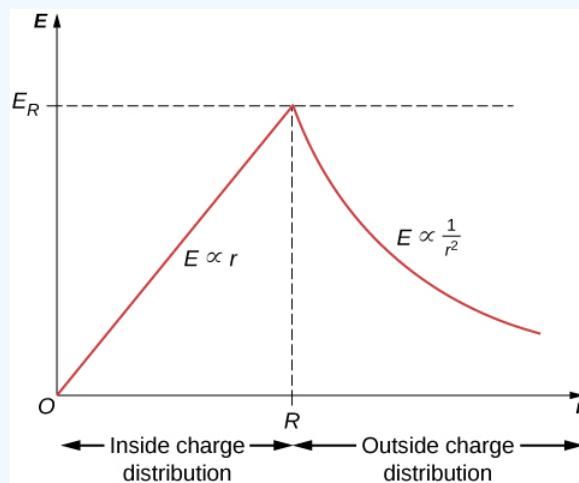


Figure 17.4.4: Electric field of a uniformly charged, non-conducting sphere increases inside the sphere to a maximum at the surface and then decreases as $1/r^2$. Here, $E_R = \frac{\rho_0 R}{3\epsilon_0}$. The electric field is due to a spherical charge distribution of uniform charge density and total charge Q as a function of distance from the center of the distribution.

The direction of the electric field at any point P is radially outward from the origin if ρ_0 is positive, and inward (i.e., toward the center) if ρ_0 is negative. The electric field at some representative space points are displayed in Figure 17.4.5 whose radial coordinates r are $r = R/2$, $r = R$, and $r = 2R$.

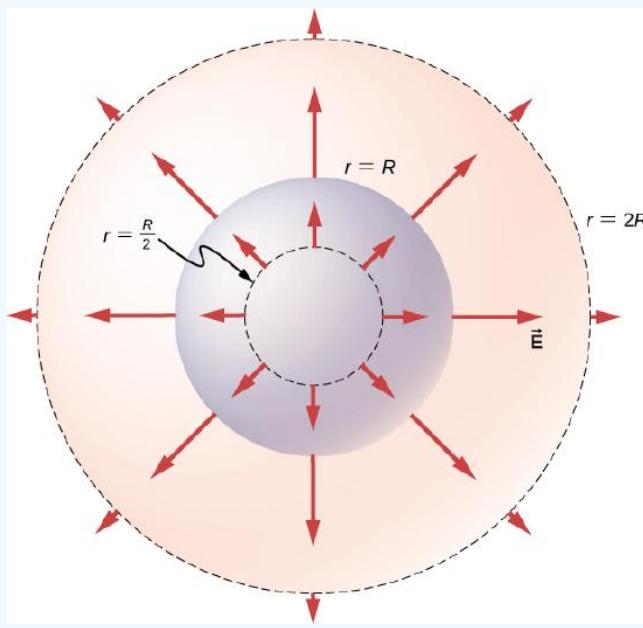


Figure 17.4.5: Electric field vectors inside and outside a uniformly charged sphere.

Significance

Notice that E_{out} has the same form as the equation of the electric field of an isolated point charge. In determining the electric field of a uniform spherical charge distribution, we can therefore assume that all of the charge inside the appropriate spherical Gaussian surface is located at the center of the distribution.

✓ Non-Uniformly Charged Sphere

A non-conducting sphere of radius R has a non-uniform charge density that varies with the distance from its center as given by

$$\rho(r) = ar^n \quad (r \leq R; n \geq 0),$$

where a is a constant. We require $n \geq 0$ so that the charge density is not undefined at $r = 0$. Find the electric field at a point outside the sphere and at a point inside the sphere.

Strategy

Apply the Gauss's law strategy given above, where we work out the enclosed charge integrals separately for cases inside and outside the sphere.

Solution

Since the given charge density function has only a radial dependence and no dependence on direction, we have a spherically symmetrical situation. Therefore, the magnitude of the electric field at any point is given above and the direction is radial. We just need to find the enclosed charge q_{enc} , which depends on the location of the field point.

A note about symbols: We use r' for locating charges in the charge distribution and r for locating the field point(s) at the Gaussian surface(s). The letter R is used for the radius of the charge distribution.

As charge density is not constant here, we need to integrate the charge density function over the volume enclosed by the Gaussian surface. Therefore, we set up the problem for charges in one spherical shell, say between r' and $r' + dr'$ as shown in Figure 17.4.6. The volume of charges in the shell of infinitesimal width is equal to the product of the area of surface $4\pi r'^2$ and the thickness dr' . Multiplying the volume with the density at this location, which is ar'^n , gives the charge in the shell:

$$dq = ar'^n 4\pi r'^2 dr'.$$

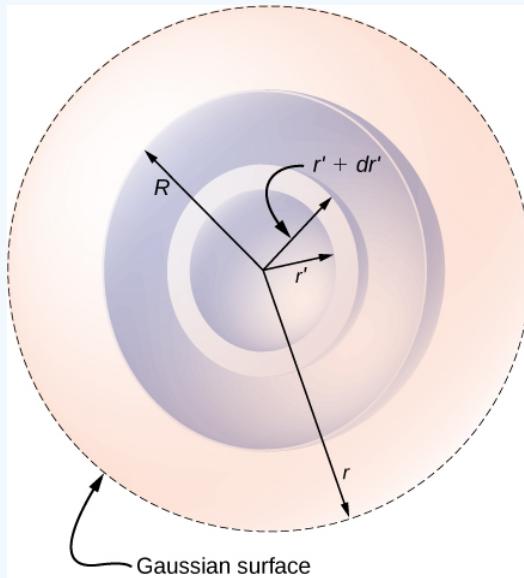


Figure 17.4.6: Spherical symmetry with non-uniform charge distribution. In this type of problem, we need four radii: R is the radius of the charge distribution, r is the radius of the Gaussian surface, r' is the inner radius of the spherical shell, and $r' + dr'$ is the outer radius of the spherical shell. The spherical shell is used to calculate the charge enclosed within the Gaussian surface. The range for r' is from 0 to r for the field at a point inside the charge distribution and from 0 to R for the field at a point outside the charge distribution. If $r > R$, then the Gaussian surface encloses more volume than the charge distribution, but the additional volume does not contribute to q_{enc} .

(a) **Field at a point outside the charge distribution.** In this case, the Gaussian surface, which contains the field point P , has a radius r that is greater than the radius R of the charge distribution, $r > R$. Therefore, all charges of the charge distribution are enclosed within the Gaussian surface. Note that the space between $r' = R$ and $r' = r$ is empty of charges and therefore does not contribute to the integral over the volume enclosed by the Gaussian surface:

$$q_{enc} = \int dq = \int_0^R ar'^n 4\pi r'^2 dr' = \frac{4\pi a}{n+3} R^{n+3}.$$

This is used in the general result for \mathbf{E}_{out} above to obtain the electric field at a point outside the charge distribution as

$$\vec{E}_{out} = \left[\frac{aR^{n+3}}{\epsilon_0(n+3)} \right] \frac{1}{r^2} \hat{r},$$

where \hat{r} is a unit vector in the direction from the origin to the field point at the Gaussian surface.

(b) **Field at a point inside the charge distribution.** The Gaussian surface is now buried inside the charge distribution, with $r < R$. Therefore, only those charges in the distribution that are within a distance r of the center of the spherical charge distribution count in q_{enc} :

$$q_{enc} = \int_0^r ar'^n 4\pi r'^2 dr' = \frac{4\pi a}{n+3} r^{n+3}.$$

Now, using the general result above for \vec{E}_{in} , we find the electric field at a point that is a distance r from the center and lies within the charge distribution as

$$\vec{E}_{in} = \left[\frac{a}{\epsilon_0(n+3)} \right] r^{n+1} \hat{r},$$

where the direction information is included by using the unit radial vector.

? Exercise 17.4.1

Check that the electric fields for the sphere reduce to the correct values for a point charge.

Answer

In this case, there is only \vec{E}_{out} . So, yes.

Charge Distribution with Cylindrical Symmetry

A charge distribution has **cylindrical symmetry** if the charge density depends only upon the distance r from the axis of a cylinder and must not vary along the axis or with direction about the axis. In other words, if your system varies if you rotate it around the axis, or shift it along the axis, you do not have cylindrical symmetry.

Figure 17.4.7 shows four situations in which charges are distributed in a cylinder. A uniform charge density ρ_0 in an infinite straight wire has a cylindrical symmetry, and so does an infinitely long cylinder with constant charge density ρ_0 . An infinitely long cylinder that has different charge densities along its length, such as a charge density ρ_1 for $z > 0$ and $\rho_2 \neq \rho_1$ for $z < 0$, does not have a usable cylindrical symmetry for this course. Neither does a cylinder in which charge density varies with the direction, such as a charge density ρ_1 for $0 \leq \theta < \pi$ and $\rho_2 \neq \rho_1$ for $\pi \leq \theta < 2\pi$. A system with concentric cylindrical shells, each with uniform charge densities, albeit different in different shells, as in Figure 17.4.7d, does have cylindrical symmetry if they are infinitely long. The infinite length requirement is due to the charge density changing along the axis of a finite cylinder. In real systems, we don't have infinite cylinders; however, if the cylindrical object is considerably longer than the radius from it that we are interested in, then the approximation of an infinite cylinder becomes useful.

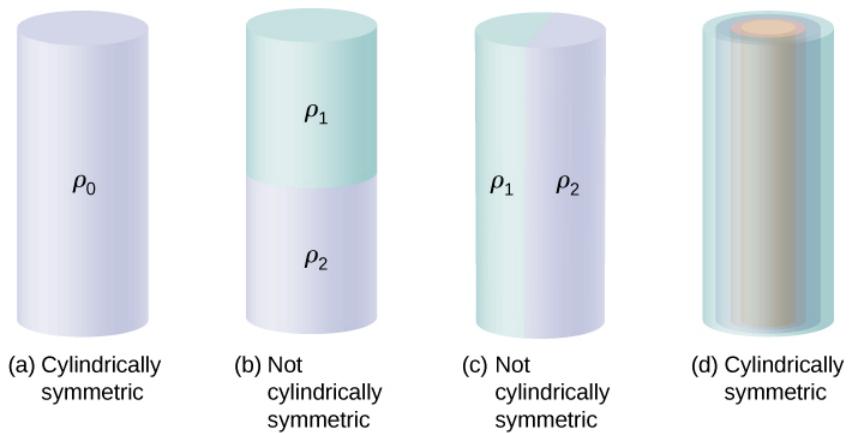


Figure 17.4.7: To determine whether a given charge distribution has cylindrical symmetry, look at the cross-section of an “infinitely long” cylinder. If the charge density does not depend on the polar angle of the cross-section or along the axis, then you have cylindrical symmetry. (a) Charge density is constant in the cylinder; (b) upper half of the cylinder has a different charge density from the lower half; (c) left half of the cylinder has a different charge density from the right half; (d) charges are constant in different cylindrical rings, but the density does not depend on the polar angle. Cases (a) and (d) have cylindrical symmetry, whereas (b) and (c) do not.

Consequences of SAymmetry

In all cylindrically symmetrical cases, the electric field \vec{E}_p at any point P must also display cylindrical symmetry.

Cylindrical symmetry: $\vec{E}_p = E_p(r)\hat{r}$, where r is the distance from the axis and \hat{r} is a unit vector directed perpendicularly away from the axis (Figure 17.4.8).

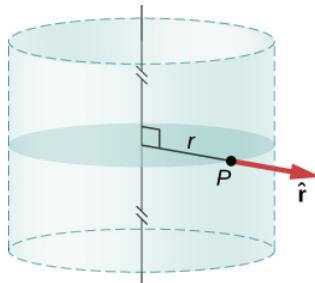


Figure 17.4.8: The electric field in a cylindrically symmetrical situation depends only on the distance from the axis. The direction of the electric field is pointed away from the axis for positive charges and toward the axis for negative charges.

Gaussian surface and flux calculation

To make use of the direction and functional dependence of the electric field, we choose a closed Gaussian surface in the shape of a cylinder with the same axis as the axis of the charge distribution. The flux through this surface of radius s and height L is easy to compute if we divide our task into two parts: (a) a flux through the flat ends and (b) a flux through the curved surface (Figure 17.4.9).

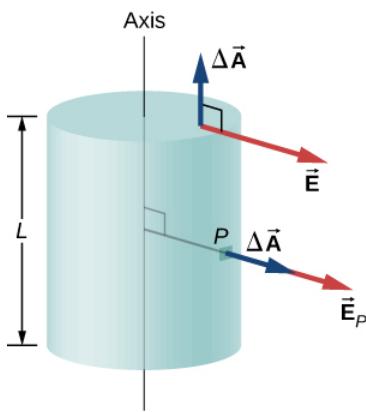


Figure 17.4.9: The Gaussian surface in the case of cylindrical symmetry. The electric field at a patch is either parallel or perpendicular to the normal to the patch of the Gaussian surface.

The electric field is perpendicular to the cylindrical side and parallel to the planar end caps of the surface. The flux through the cylindrical part is

$$\int_S \vec{E} \cdot \hat{n} dA = E \int_S dA = E(2\pi r L),$$

whereas the flux through the end caps is zero because $\vec{E} \cdot \hat{n} = 0$ there. Thus, the flux is

$$\int_S \vec{E} \cdot \hat{n} dA = E(2\pi r L) + 0 + 0 = 2\pi r L E.$$

Using Gauss's law

According to Gauss's law, the flux must equal the amount of charge within the volume enclosed by this surface, divided by the permittivity of free space. When you do the calculation for a cylinder of length L , you find that q_{enc} of Gauss's law is directly proportional to L . Let us write it as charge per unit length (λ_{enc}) times length L :

$$q_{enc} = \lambda_{enc} L. \quad (17.4.14)$$

Hence, Gauss's law for any cylindrically symmetrical charge distribution yields the following magnitude of the electric field a distance s away from the axis:

$$\text{Magnitude : } E(r) = \frac{\lambda_{enc}}{2\pi\epsilon_0} \frac{1}{r}. \quad (17.4.15)$$

The charge per unit length λ_{enc} depends on whether the field point is inside or outside the cylinder of charge distribution, just as we have seen for the spherical distribution.

Computing enclosed charge

Let R be the radius of the cylinder within which charges are distributed in a cylindrically symmetrical way. Let the field point P be at a distance s from the axis. (The side of the Gaussian surface includes the field point P .) When $r > R$ (that is, when P is outside the charge distribution), the Gaussian surface includes all the charge in the cylinder of radius R and length L . When $r < R$ (P is located inside the charge distribution), then only the charge within a cylinder of radius s and length L is enclosed by the Gaussian surface:

$$\lambda_{enc} = (\text{total charge}) \text{ if } r \geq R$$

$$\lambda_{enc} = (\text{only charge within } r < R) \text{ if } r < R$$

✓ Uniformly Charged Cylindrical Shell

A very long non-conducting cylindrical shell of radius R has a uniform surface charge density σ_0 . Find the electric field (a) at a point outside the shell and (b) at a point inside the shell.

Strategy

Apply the Gauss's law strategy given earlier, where we treat the cases inside and outside the shell separately.

Solution

a. Electric field at a point outside the shell. For a point outside the cylindrical shell, the Gaussian surface is the surface of a cylinder of radius $r > R$ and length L , as shown in Figure 17.4.10. The charge enclosed by the Gaussian cylinder is equal to the charge on the cylindrical shell of length L . Therefore, λ_{enc} is given by

$$\lambda_{enc} = \frac{\sigma_0 2\pi R L}{L} = 2\pi R \sigma_0. \quad (17.4.16)$$

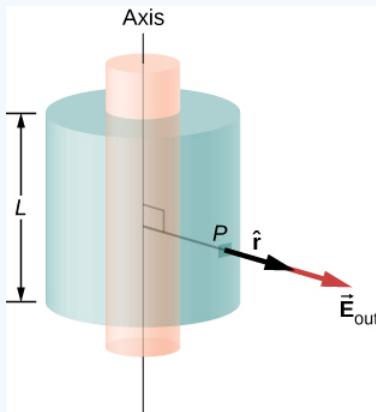


Figure 17.4.10: A Gaussian surface surrounding a cylindrical shell.

Hence, the electric field at a point P outside the shell at a distance r away from the axis is

$$\vec{E} = \frac{2\pi R \sigma_0}{2\pi\epsilon_0} \frac{1}{r} \hat{r} = \frac{R \sigma_0}{\epsilon_0} \frac{1}{r} \hat{r} (r > R) \quad (17.4.17)$$

where \hat{r} is a unit vector, perpendicular to the axis and pointing away from it, as shown in the figure. The electric field at P points in the direction of \hat{r} given in Figure 17.4.10 if $\sigma_0 > 0$ and in the opposite direction to \hat{r} if $\sigma_0 < 0$.

b. Electric field at a point inside the shell. For a point inside the cylindrical shell, the Gaussian surface is a cylinder whose radius r is less than R (Figure 17.4.11). This means no charges are included inside the Gaussian surface:

$$\lambda_{enc} = 0. \quad (17.4.18)$$

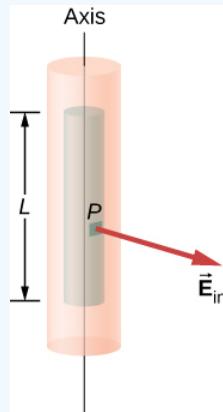


Figure 17.4.11: A Gaussian surface within a cylindrical shell.

This gives the following equation for the magnitude of the electric field E_{in} at a point whose r is less than R of the shell of charges.

$$E_{in} 2\pi r L = 0 (r < R),$$

This gives us

$$E_{in} = 0 (r < R).$$

Significance

Notice that the result inside the shell is exactly what we should expect: No enclosed charge means zero electric field. Outside the shell, the result becomes identical to a wire with uniform charge $R\sigma$.

Exercise 17.4.1

A thin straight wire has a uniform linear charge density λ_0 . Find the electric field at a distance d from the wire, where d is much less than the length of the wire.

Answer

$\vec{E} = \frac{\lambda_0}{2\pi\epsilon_0 d} \hat{r}$; This agrees with the calculation of [Calculating Electric Fields of Charge Distributions](#) where we found the electric field by integrating over the charged wire. Notice how much simpler the calculation of this electric field is with Gauss's law.

Charge Distribution with Planar Symmetry

A planar symmetry of charge density is obtained when charges are uniformly spread over a large flat surface. In planar symmetry, all points in a plane parallel to the plane of charge are identical with respect to the charges.

Consequences of symmetry

We take the plane of the charge distribution to be the xy -plane and we find the electric field at a space point P with coordinates (x, y, z) . Since the charge density is the same at all (x, y) -coordinates in the $z = 0$ plane, by symmetry, the electric field at P cannot depend on the x - or y -coordinates of point P , as shown in Figure 17.4.12. Therefore, the electric field at P can only depend on the distance from the plane and has a direction either toward the plane or away from the plane. That is, the electric field at P has only a nonzero z -component.

Uniform charges in xy plane: $\vec{E} = E(z)\hat{z}$ where z is the distance from the plane and \hat{z} is the unit vector normal to the plane. Note that in this system, $E(z) = E(-z)$, although of course they point in opposite directions.

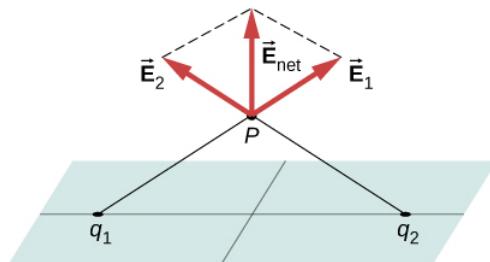


Figure 17.4.12: The components of the electric field parallel to a plane of charges cancel out the two charges located symmetrically from the field point P . Therefore, the field at any point is pointed vertically from the plane of charges. For any point P and charge q_1 , we can always find a q_2 with this effect.

Gaussian surface and flux calculation

In the present case, a convenient Gaussian surface is a box, since the expected electric field points in one direction only. To keep the Gaussian box symmetrical about the plane of charges, we take it to straddle the plane of the charges, such that one face containing the field point P is taken parallel to the plane of the charges. In Figure 17.4.13, sides I and II of the Gaussian surface (the box) that are parallel to the infinite plane have been shaded. They are the only surfaces that give rise to nonzero flux because the electric field and the area vectors of the other faces are perpendicular to each other.

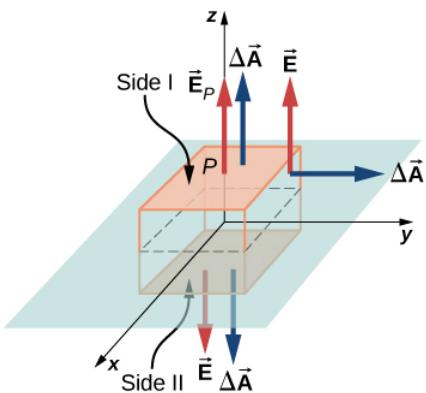


Figure 17.4.13: A thin charged sheet and the Gaussian box for finding the electric field at the field point P . The normal to each face of the box is from inside the box to outside. On two faces of the box, the electric fields are parallel to the area vectors, and on the other four faces, the electric fields are perpendicular to the area vectors.

Let A be the area of the shaded surface on each side of the plane and E_p be the magnitude of the electric field at point P . Since sides I and II are at the same distance from the plane, the electric field has the same magnitude at points in these planes, although the directions of the electric field at these points in the two planes are opposite to each other.

Magnitude at I or II: $E(z) = E_p$.

If the charge on the plane is positive, then the direction of the electric field and the area vectors are as shown in Figure 17.4.13. Therefore, we find for the flux of electric field through the box

$$\Phi = \int_S \vec{E}_p \cdot \hat{n} dA = E_p A + E_p A + 0 + 0 + 0 + 0 = 2E_p A \quad (17.4.19)$$

where the zeros are for the flux through the other sides of the box. Note that if the charge on the plane is negative, the directions of electric field and area vectors for planes I and II are opposite to each other, and we get a negative sign for the flux. According to Gauss's law, the flux must equal q_{enc}/ϵ_0 . From Figure 17.4.13, we see that the charges inside the volume enclosed by the Gaussian box reside on an area A of the xy -plane. Hence,

$$q_{enc} = \sigma_0 A.$$

Using the equations for the flux and enclosed charge in Gauss's law, we can immediately determine the electric field at a point at height z from a uniformly charged plane in the xy -plane:

$$\vec{E}_p = \frac{\sigma_0}{2\epsilon_0} \hat{n}.$$

The direction of the field depends on the sign of the charge on the plane and the side of the plane where the field point P is located. Note that above the plane, $\hat{n} = +\hat{z}$, while below the plane, $\hat{n} = -\hat{z}$.

You may be surprised to note that the electric field does not actually depend on the distance from the plane; this is an effect of the assumption that the plane is infinite. In practical terms, the result given above is still a useful approximation for finite planes near the center.

This page titled [17.4: Calculating Electric Field Using Gauss's Law](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.4: Applying Gauss's Law](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

17.5: Conductors in Electrostatic Equilibrium via Gauss's Law

Learning Objectives

By the end of this section, you will be able to:

- Describe the electric field within a conductor at equilibrium
- Describe the electric field immediately outside the surface of a charged conductor at equilibrium
- Explain why if the field is not as described in the first two objectives, the conductor is not at equilibrium

So far, we have generally been working with charges occupying a volume within an insulator. We now study what happens when free charges are placed on a conductor. Generally, in the presence of a (generally external) electric field, the free charge in a conductor redistributes and very quickly reaches electrostatic equilibrium. The resulting charge distribution and its electric field have many interesting properties, which we can investigate with the help of Gauss's law and the concept of electric potential.

The Electric Field inside a Conductor Vanishes

If an electric field is present inside a conductor, it exerts forces on the **free electrons** (also called conduction electrons), which are electrons in the material that are not bound to an atom. These free electrons then accelerate. However, moving charges by definition means nonstatic conditions, contrary to our assumption. Therefore, when electrostatic equilibrium is reached, the charge is distributed in such a way that the electric field inside the conductor vanishes.

If you place a piece of a metal near a positive charge, the free electrons in the metal are attracted to the external positive charge and migrate freely toward that region. The region the electrons move to then has an excess of electrons over the protons in the atoms and the region from where the electrons have migrated has more protons than electrons. Consequently, the metal develops a negative region near the charge and a positive region at the far end (Figure 17.5.1). As we saw in the preceding chapter, this separation of equal magnitude and opposite type of electric charge is called **polarization**. If you remove the external charge, the electrons migrate back and neutralize the positive region.

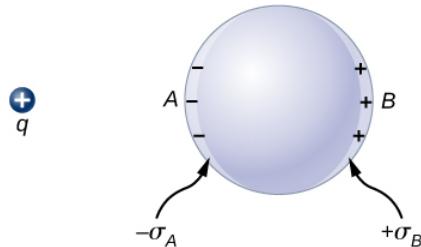


Figure 17.5.1: Polarization of a metallic sphere by an external point charge $+q$. The near side of the metal has an opposite surface charge compared to the far side of the metal. The sphere is said to be polarized. When you remove the external charge, the polarization of the metal also disappears.

The polarization of the metal happens only in the presence of external charges. You can think of this in terms of electric fields. The external charge creates an external electric field. When the metal is placed in the region of this electric field, the electrons and protons of the metal experience electric forces due to this external electric field, but only the conduction electrons are free to move in the metal over macroscopic distances. The movement of the conduction electrons leads to the polarization, which creates an induced electric field in addition to the external electric field (Figure 17.5.2). The net electric field is a vector sum of the fields of $+q$ and the surface charge densities $-\sigma_A$ and $+\sigma_B$. This means that the net field inside the conductor is different from the field outside the conductor.

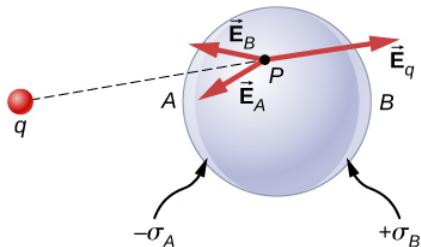


Figure 17.5.2: In the presence of an external charge q , the charges in a metal redistribute. The electric field at any point has three contributions, from $+q$ and the induced charges $-\sigma_A$ and $+\sigma_B$. Note that the surface charge distribution will not be uniform in this case.

The redistribution of charges is such that the sum of the three contributions at any point P inside the conductor is

$$\vec{E}_p = \vec{E}_q + \vec{E}_B + \vec{E}_A = \vec{0}.$$

Now, thanks to [Gauss's law](#), we know that there is no net charge enclosed by a Gaussian surface that is solely within the volume of the conductor at equilibrium. That is, $q_{\text{enc}} = 0$ and hence

$$\vec{E}_{\text{net}} = \vec{0} \text{ (at points inside a conductor).} \quad (17.5.1)$$

Charge on a Conductor

An interesting property of a conductor in static equilibrium is that extra charges on the conductor end up on the outer surface of the conductor, regardless of where they originate. Figure 17.5.3 illustrates a system in which we bring an external positive charge inside the cavity of a metal and then touch it to the inside surface. Initially, the inside surface of the cavity is negatively charged and the outside surface of the conductor is positively charged. When we touch the inside surface of the cavity, the induced charge is neutralized, leaving the outside surface and the whole metal charged with a net positive charge.

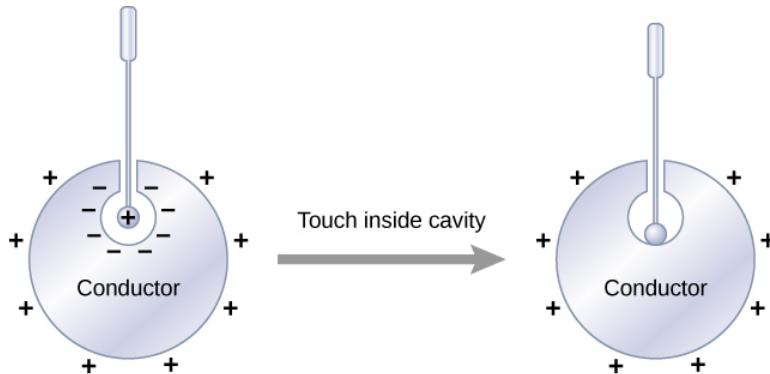


Figure 17.5.3: Electric charges on a conductor migrate to the outside surface no matter where you put them initially.

To see why this happens, note that the Gaussian surface in Figure 17.5.4 (the dashed line) follows the contour of the actual surface of the conductor and is located an infinitesimal distance **within** it. Since $\vec{E} = \vec{0}$ everywhere inside a conductor,

$$\oint \vec{E} \cdot \hat{n} dA = 0. \quad (17.5.2)$$

Thus, from Gauss' law, there is no net charge inside the Gaussian surface. But the Gaussian surface lies just below the actual surface of the conductor; consequently, there is no net charge inside the conductor. Any excess charge must lie on its surface.

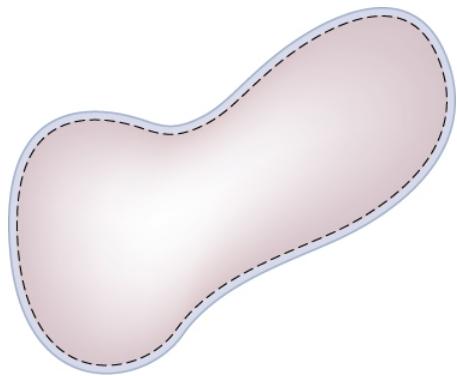


Figure 17.5.4: The dashed line represents a Gaussian surface that is just beneath the actual surface of the conductor.

This particular property of conductors is the basis for an extremely accurate method developed by Plimpton and Lawton in 1936 to verify Gauss's law and, correspondingly, Coulomb's law. A sketch of their apparatus is shown in Figure 17.5.5. Two spherical shells are connected to one another through an electrometer E , a device that can detect a very slight amount of charge flowing from one shell to the other. When switch S is thrown to the left, charge is placed on the outer shell by the battery B . Will charge flow through the electrometer to the inner shell?

No. Doing so would mean a violation of Gauss's law. Plimpton and Lawton did not detect any flow and, knowing the sensitivity of their electrometer, concluded that if the radial dependence in Coulomb's law were $1/r^{2+\delta}$, δ would be less than 2×10^{-9} ¹. More recent measurements place δ at less than 3×10^{-18} ², a number so small that the validity of Coulomb's law seems indisputable.

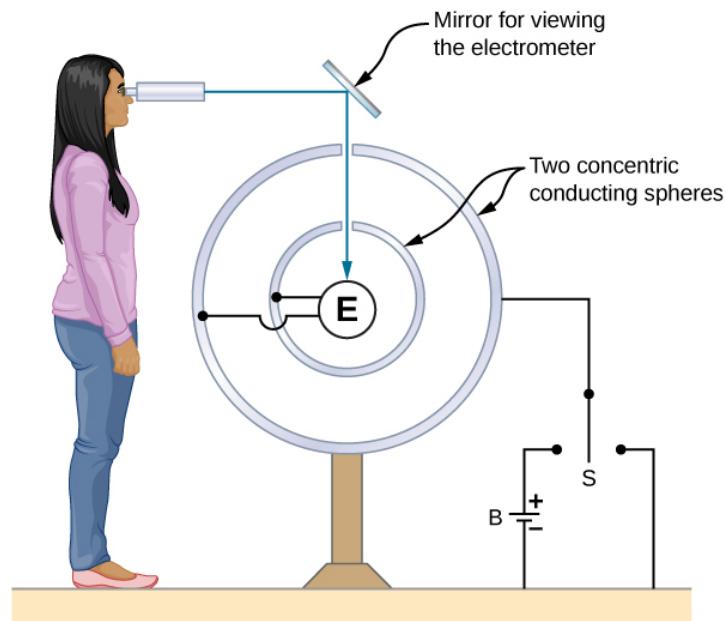


Figure 17.5.5: A representation of the apparatus used by Plimpton and Lawton. Any transfer of charge between the spheres is detected by the electrometer E .

The Electric Field at the Surface of a Conductor

If the electric field had a component parallel to the surface of a conductor, free charges on the surface would move, a situation contrary to the assumption of electrostatic equilibrium. Therefore, the electric field is always perpendicular to the surface of a conductor.

At any point just above the surface of a conductor, the surface charge density σ and the magnitude of the electric field \vec{E} are related by

$$E = \frac{\sigma}{\epsilon_0}. \quad (17.5.3)$$

To see this, consider an infinitesimally small Gaussian cylinder that surrounds a point on the surface of the conductor, as in Figure 17.5.6. The cylinder has one end face inside and one end face outside the surface. The height and cross-sectional area of the cylinder are δ and ΔA , respectively. The cylinder's sides are perpendicular to the surface of the conductor, and its end faces are parallel to the surface. Because the cylinder is infinitesimally small, the charge density σ is essentially constant over the surface enclosed, so the total charge inside the Gaussian cylinder is $\sigma\Delta A$. Now \vec{E} is perpendicular to the surface of the conductor outside the conductor and vanishes within it, because otherwise, the charges would accelerate, and we would not be in equilibrium. Electric flux therefore crosses only the outer end face of the Gaussian surface and may be written as $E\Delta A$ since the cylinder is assumed to be small enough that \vec{E} is approximately constant over that area. From Gauss' law,

$$E\Delta A = \frac{\sigma\Delta A}{\epsilon_0}. \quad (17.5.4)$$

Thus

$$E = \frac{\sigma}{\epsilon_0}. \quad (17.5.5)$$

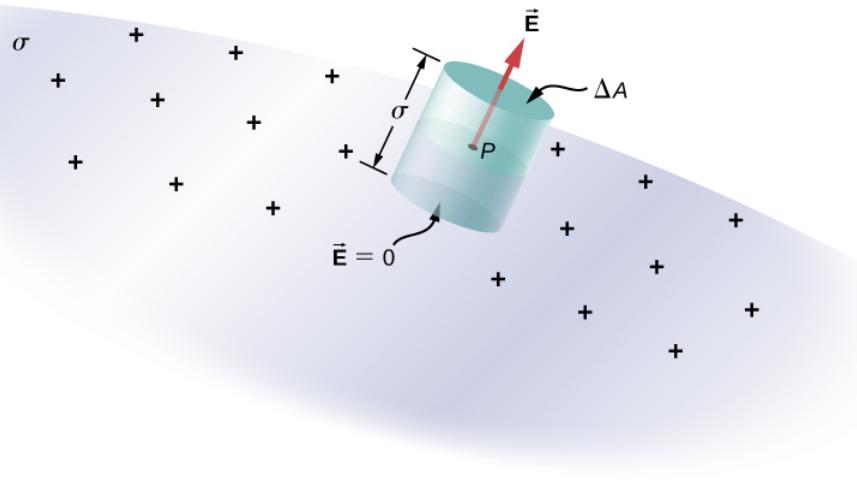


Figure 17.5.6: An infinitesimally small cylindrical Gaussian surface surrounds point **P**, which is on the surface of the conductor. The field \vec{E} is perpendicular to the surface of the conductor outside the conductor and vanishes within it.

✓ Electric Field of a Conducting Plate

The infinite conducting plate in Figure 17.5.7 has a uniform surface charge density σ . Use Gauss' law to find the electric field outside the plate. Compare this result with that previously calculated directly.

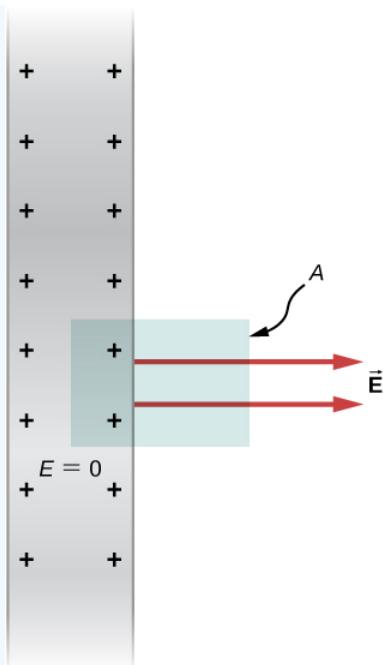


Figure 17.5.7: A side view of an infinite conducting plate and Gaussian cylinder with cross-sectional area A .

Strategy

For this case, we use a cylindrical Gaussian surface, a side view of which is shown.

Solution

The flux calculation is similar to that for an infinite sheet of charge from the previous chapter with one major exception: The left face of the Gaussian surface is inside the conductor where $\vec{E} = \vec{0}$, so the total flux through the Gaussian surface is EA rather than $2EA$. Then from Gauss' law,

$$EA = \frac{\sigma A}{\epsilon_0}$$

and the electric field outside the plate is

$$E = \frac{\sigma}{\epsilon_0}$$

Significance

This result is in agreement with the result from the previous section, and consistent with the rule stated above.

✓ Electric Field between Oppositely Charged Parallel Plates

Two large conducting plates carry equal and opposite charges, with a surface charge density σ of magnitude $6.81 \times 10^{-7} C/m^2$, as shown in Figure 17.5.8. The separation between the plates is $l = 6.50 \text{ mm}$. What is the electric field between the plates?

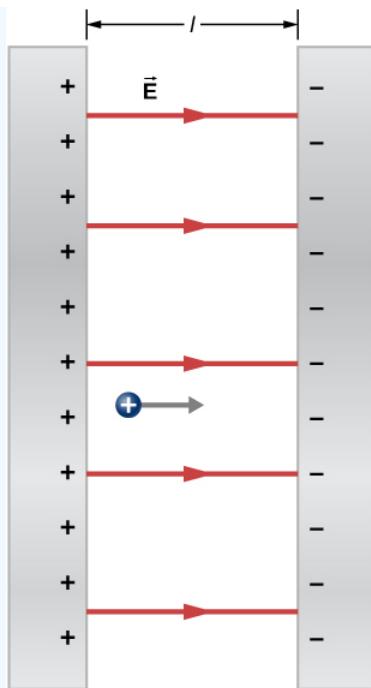


Figure 17.5.8: The electric field between oppositely charged parallel plates. A test charge is released at the positive plate.

Strategy Note that the electric field at the surface of one plate only depends on the charge on that plate. Thus, apply $E = \sigma/\epsilon_0$ with the given values.

Solution The electric field is directed from the positive to the negative plate, as shown in the figure, and its magnitude is given by

$$E = \frac{\sigma}{\epsilon_0} = \frac{6.81 \times 10^{-7} C/m^2}{8.85 \times 10^{-12} C^2/Nm^2} = 7.69 \times 10^4 N/C \quad (17.5.6)$$

Significance

This formula is applicable to more than just a plate. Furthermore, two-plate systems will be important later.

✓ A Conducting Sphere

The isolated conducting sphere (Figure 17.5.9) has a radius R and an excess charge q . What is the electric field both inside and outside the sphere?

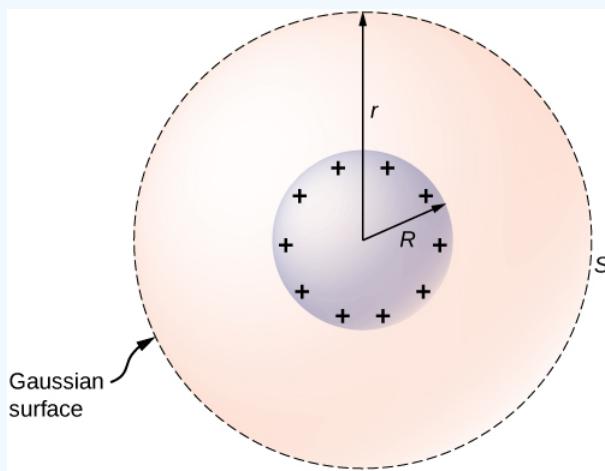


Figure 17.5.9: An isolated conducting sphere.

Strategy The sphere is isolated, so its surface charge distribution and the electric field of that distribution are spherically symmetrical. We can therefore represent the field as $\vec{E} = E(r)\hat{r}$. To calculate $\vec{E}(r)$, we apply Gauss's law over a closed spherical surface S of radius r that is concentric with the conducting sphere.

Solution

Since \mathbf{r} is constant and $\hat{\mathbf{n}} = \hat{r}$ on the sphere,

$$\oint_S \vec{E} \cdot \hat{n} dA = E(r) \oint_S dA = E(r) 4\pi r^2. \quad (17.5.7)$$

For $r < R$, S is within the conductor, so $q_{enc} = 0$, and Gauss's law gives

$$E(r) = 0, \quad (17.5.8)$$

as expected inside a conductor. If $r > R$, S encloses the conductor so $q_{enc} = q$. From Gauss's law,

$$E(r) 4\pi r^2 = \frac{q}{\epsilon_0}. \quad (17.5.9)$$

The electric field of the sphere may therefore be written as

$$\vec{E} = \vec{0} \quad (r < R), \quad (17.5.10)$$

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r} \quad (r \geq R). \quad (17.5.11)$$

Significance Notice that in the region $r \geq R$, the electric field due to a charge q placed on an isolated conducting sphere of radius R is identical to the electric field of a point charge q located at the center of the sphere. The difference between the charged metal and a point charge occurs only at the space points inside the conductor. For a point charge placed at the center of the sphere, the electric field is not zero at points of space occupied by the sphere, but a conductor with the same amount of charge has a zero electric field at those points (Figure 17.5.10). However, there is no distinction at the outside points in space where $r > R$, and we can replace the isolated charged spherical conductor by a point charge at its center with impunity.

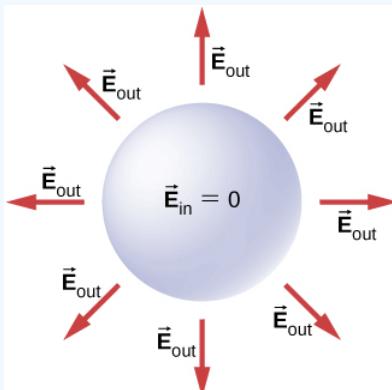


Figure 17.5.10: Electric field of a positively charged metal sphere. The electric field inside is zero, and the electric field outside is same as the electric field of a point charge at the center, although the charge on the metal sphere is at the surface.

Exercise 17.5.1

How will the system above change if there are charged objects external to the sphere?

Answer

If there are other charged objects around, then the charges on the surface of the sphere will not necessarily be spherically symmetrical; there will be more in certain directions than in other directions.

For a conductor with a cavity, if we put a charge $+q$ inside the cavity, then the charge separation takes place in the conductor, with $-q$ amount of charge on the inside surface and a $+q$ amount of charge at the outside surface (Figure 17.5.11a). For the same

conductor with a charge $+q$ outside it, there is no excess charge on the inside surface; both the positive and negative induced charges reside on the outside surface (Figure 17.5.11b).

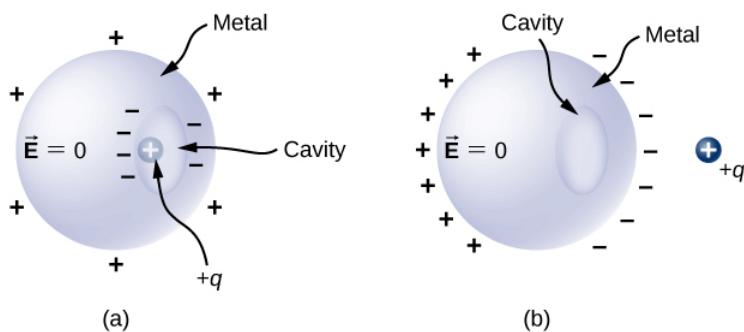


Figure 17.5.11: (a) A charge inside a cavity in a metal. The distribution of charges at the outer surface does not depend on how the charges are distributed at the inner surface, since the E -field inside the body of the metal is zero. That magnitude of the charge on the outer surface does depend on the magnitude of the charge inside, however. (b) A charge outside a conductor containing an inner cavity. The cavity remains free of charge. The polarization of charges on the conductor happens at the surface.

If a conductor has two cavities, one of them having a charge $+q_a$ inside it and the other a charge $-q_b$, the polarization of the conductor results in $-q_a$ on the inside surface of the cavity a , $+q_b$ on the inside surface of the cavity b , and $q_a - q_b$ on the outside surface (Figure 17.5.12). The charges on the surfaces may not be uniformly spread out; their spread depends upon the geometry. The only rule obeyed is that when the equilibrium has been reached, the charge distribution in a conductor is such that the electric field by the charge distribution in the conductor cancels the electric field of the external charges at all space points inside the body of the conductor.

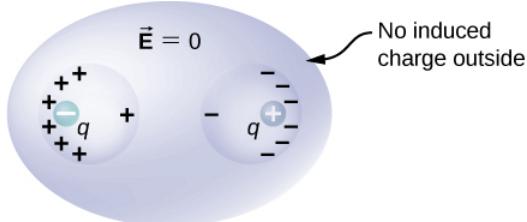


Figure 17.5.12: The charges induced by two equal and opposite charges in two separate cavities of a conductor. If the net charge on the cavity is nonzero, the external surface becomes charged to the amount of the net charge.

This page titled [17.5: Conductors in Electrostatic Equilibrium via Gauss's Law](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via source content that was edited to the style and standards of the LibreTexts platform.

- [6.5: Conductors in Electrostatic Equilibrium](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

17.6: Gauss's Law (Summary)

Key Terms

area vector	vector with magnitude equal to the area of a surface and direction perpendicular to the surface
cylindrical symmetry	system only varies with distance from the axis, not direction
electric flux	dot product of the electric field and the area through which it is passing
flux	quantity of something passing through a given area
free electrons	also called conduction electrons, these are the electrons in a conductor that are not bound to any particular atom, and hence are free to move around
Gaussian surface	any enclosed (usually imaginary) surface
planar symmetry	system only varies with distance from a plane
spherical symmetry	system only varies with the distance from the origin, not in direction

Key Equations

Definition of electric flux, for uniform electric field	$\Phi = \vec{E} \cdot \vec{A} \rightarrow EA \cos \theta$
Electric flux through an open surface	$\Phi = \int_S \vec{E} \cdot \hat{n} dA = \int_S \vec{E} \cdot d\vec{A}$
Electric flux through a closed surface	$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \oint_S \vec{E} \cdot d\vec{A}$
Gauss's law	$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \frac{q_{enc}}{\epsilon_0}$
Gauss's Law for systems with symmetry	$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = E \oint_S dA = EA = \frac{q_{enc}}{\epsilon_0}$
The magnitude of the electric field just outside the surface of a conductor	$E = \frac{\sigma}{\epsilon_0}$

Summary

6.2 Electric Flux

- The electric flux through a surface is proportional to the number of field lines crossing that surface. Note that this means the magnitude is proportional to the portion of the field perpendicular to the area.
- The electric flux is obtained by evaluating the surface integral

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \oint_S \vec{E} \cdot d\vec{A},$$

where the notation used here is for a closed surface S.

6.3 Explaining Gauss's Law

- Gauss's law relates the electric flux through a closed surface to the net charge within that surface,

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \frac{q_{enc}}{\epsilon_0},$$

where q_{enc} is the total charge inside the Gaussian surface S.

- All surfaces that include the same amount of charge have the same number of field lines crossing it, regardless of the shape or size of the surface, as long as the surfaces enclose the same amount of charge.

6.4 Applying Gauss's Law

- For a charge distribution with certain spatial symmetries (spherical, cylindrical, and planar), we can find a Gaussian surface over which $\vec{E} \cdot \hat{n} = E$, where E is constant over the surface. The electric field is then determined with Gauss's law.
- For spherical symmetry, the Gaussian surface is also a sphere, and Gauss's law simplifies to $4\pi r^2 E = \frac{q_{enc}}{\epsilon_0}$.
- For cylindrical symmetry, we use a cylindrical Gaussian surface, and find that Gauss's law simplifies to $2\pi r L E = \frac{q_{enc}}{\epsilon_0}$.
- For planar symmetry, a convenient Gaussian surface is a box penetrating the plane, with two faces parallel to the plane and the remainder perpendicular, resulting in Gauss's law being $2AE = \frac{q_{enc}}{\epsilon_0}$.

6.5 Conductors in Electrostatic Equilibrium

- The electric field inside a conductor vanishes.
- Any excess charge placed on a conductor resides entirely on the surface of the conductor.
- The electric field is perpendicular to the surface of a conductor everywhere on that surface.
- The magnitude of the electric field just above the surface of a conductor is given by $E = \frac{\sigma}{\epsilon_0}$.

This page titled [17.6: Gauss's Law \(Summary\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.6: Gauss's Law \(Summary\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

17.7: Gauss's Law (Exercises)

Conceptual Questions

6.2 Electric Flux

1. Discuss how would orient a planar surface of area A in a uniform electric field of magnitude E_0 to obtain
 - (a) the maximum flux and
 - (b) the minimum flux through the area.
2. What are the maximum and minimum values of the flux in the preceding question?
3. The net electric flux crossing a closed surface is always zero. True or false?
4. The net electric flux crossing an open surface is never zero. True or false?

6.3 Explaining Gauss's Law

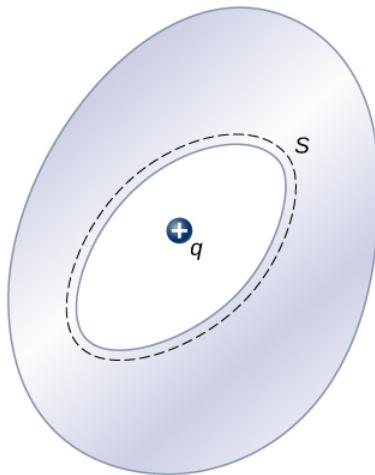
5. Two concentric spherical surfaces enclose a point charge q . The radius of the outer sphere is twice that of the inner one. Compare the electric fluxes crossing the two surfaces.
6. Compare the electric flux through the surface of a cube of side length a that has a charge q at its center to the flux through a spherical surface of radius a with a charge q at its center.
7. (a) If the electric flux through a closed surface is zero, is the electric field necessarily zero at all points on the surface?
(b) What is the net charge inside the surface?
8. Discuss how Gauss's law would be affected if the electric field of a point charge did not vary as $1/r^2$.
9. Discuss the similarities and differences between the gravitational field of a point mass m and the electric field of a point charge q .
10. Discuss whether Gauss's law can be applied to other forces, and if so, which ones.
11. Is the term \vec{E} in Gauss's law the electric field produced by just the charge inside the Gaussian surface?
12. Reformulate Gauss's law by choosing the unit normal of the Gaussian surface to be the one directed inward.

6.4 Applying Gauss's Law

13. Would Gauss's law be helpful for determining the electric field of two equal but opposite charges a fixed distance apart?
14. Discuss the role that symmetry plays in the application of Gauss's law. Give examples of continuous charge distributions in which Gauss's law is useful and not useful in determining the electric field.
15. Discuss the restrictions on the Gaussian surface used to discuss planar symmetry. For example, is its length important? Does the cross-section have to be square? Must the end faces be on opposite sides of the sheet?

6.5 Conductors in Electrostatic Equilibrium

16. Is the electric field inside a metal always zero?
17. Under electrostatic conditions, the excess charge on a conductor resides on its surface. Does this mean that all the conduction electrons in a conductor are on the surface?
18. A charge q is placed in the cavity of a conductor as shown below. Will a charge outside the conductor experience an electric field due to the presence of q ?



- 19.** The conductor in the preceding figure has an excess charge of $-5.0\mu C$. If a $2.0 - \mu C$ point charge is placed in the cavity, what is the net charge on the surface of the cavity and on the outer surface of the conductor?

Problems

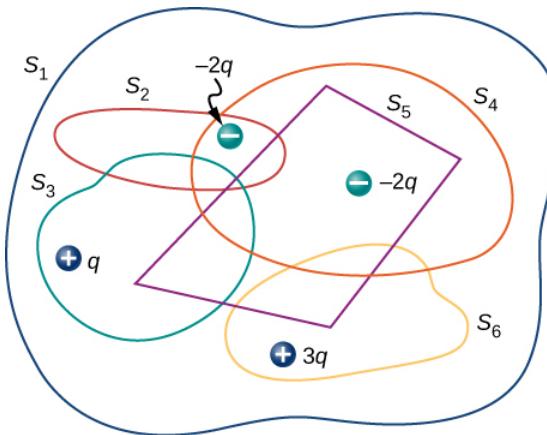
6.2 Electric Flux

- 20.** A uniform electric field of magnitude $1.1 \times 10^4 N/C$ is perpendicular to a square sheet with sides 2.0 m long. What is the electric flux through the sheet?
- 21.** Calculate the flux through the sheet of the previous problem if the plane of the sheet is at an angle of 60° to the field. Find the flux for both directions of the unit normal to the sheet.
- 22.** Find the electric flux through a rectangular area $3\text{cm} \times 2\text{cm}$ between two parallel plates where there is a constant electric field of 30 N/C for the following orientations of the area: (a) parallel to the plates, (b) perpendicular to the plates, and (c) the normal to the area making a 30° angle with the direction of the electric field. Note that this angle can also be given as $180^\circ + 30^\circ$.
- 23.** The electric flux through a square-shaped area of side 5 cm near a large charged sheet is found to be $3 \times 10^{-5} N \cdot m^2/C$. when the area is parallel to the plate. Find the charge density on the sheet.
- 24.** Two large rectangular aluminum plates of area 150cm^2 face each other with a separation of 3 mm between them. The plates are charged with equal amount of opposite charges, $\pm 20\mu C$. The charges on the plates face each other. Find the flux through a circle of radius 3 cm between the plates when the normal to the circle makes an angle of 5° with a line perpendicular to the plates. Note that this angle can also be given as $180^\circ + 5^\circ$.
- 25.** A square surface of area 2cm^2 is in a space of uniform electric field of magnitude $10^3 N/C$. The amount of flux through it depends on how the square is oriented relative to the direction of the electric field. Find the electric flux through the square, when the normal to it makes the following angles with electric field: (a) 30° , (b) 90° , and (c) 0° . Note that these angles can also be given as $180^\circ + \theta$.
- 26.** A vector field is pointed along the z -axis, $\vec{v} = \frac{\alpha}{x^2+y^2}\hat{z}$.
- Find the flux of the vector field through a rectangle in the xy -plane between $a < x < b$ and $c < y < d$.
 - Do the same through a rectangle in the yz -plane between $a < z < b$ and $c < y < d$. (Leave your answer as an integral.)
- 27.** Consider the uniform electric field $\vec{E} = (4.0\hat{j} + 3.0\hat{k}) \times 10^3 N/C$. What is its electric flux through a circular area of radius 2.0 m that lies in the xy -plane?
- 28.** Repeat the previous problem, given that the circular area is (a) in the yz -plane and (b) 45° above the xy -plane.

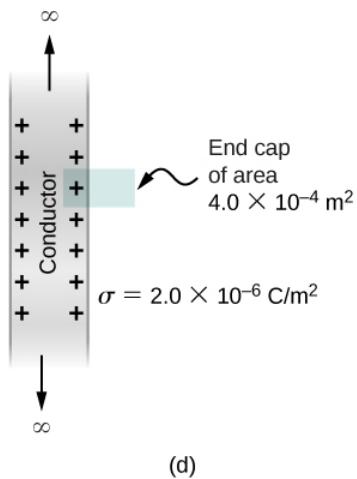
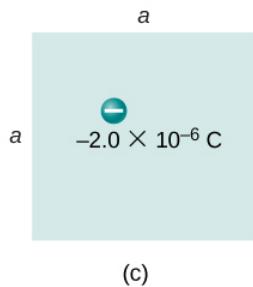
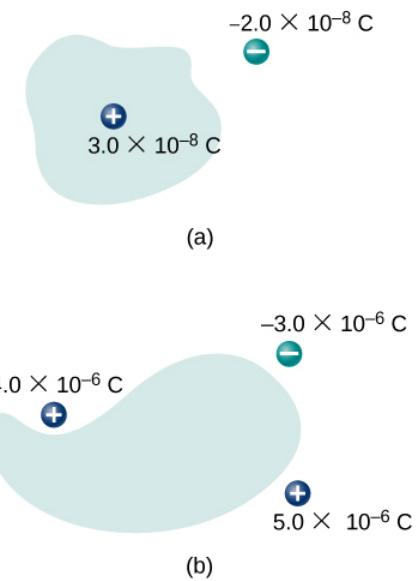
29. An infinite charged wire with charge per unit length λ lies along the central axis of a cylindrical surface of radius r and length l . What is the flux through the surface due to the electric field of the charged wire?

6.3 Explaining Gauss's Law

30. Determine the electric flux through each closed surface whose cross-section inside the surface is shown below.



31. Find the electric flux through the closed surface whose cross-sections are shown below.



32. A point charge \mathbf{q} is located at the center of a cube whose sides are of length \mathbf{a} . If there are no other charges in this system, what is the electric flux through one face of the cube?

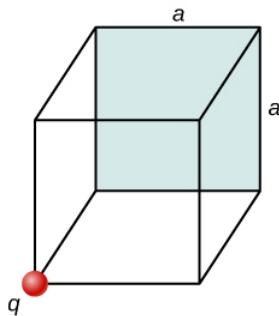
33. A point charge of $10\mu\text{C}$ is at an unspecified location inside a cube of side 2 cm. Find the net electric flux though the surfaces of the cube.

34. A net flux of $1.0 \times 10^4 \text{ N} \cdot \text{m}^2/\text{C}$ passes inward through the surface of a sphere of radius 5 cm.

(a) How much charge is inside the sphere?

(b) How precisely can we determine the location of the charge from this information?

35. A charge q is placed at one of the corners of a cube of side a , as shown below. Find the magnitude of the electric flux through the shaded face due to q . Assume $q > 0$.



36. The electric flux through a cubical box 8.0 cm on a side is $1.2 \times 10^3 N \cdot m^2/C$. What is the total charge enclosed by the box?
37. The electric flux through a spherical surface is $4.0 \times 10^4 N \cdot m^2/C$. What is the net charge enclosed by the surface?
38. A cube whose sides are of length d is placed in a uniform electric field of magnitude $E = 4.0 \times 10^3 N/C$ so that the field is perpendicular to two opposite faces of the cube. What is the net flux through the cube?
39. Repeat the previous problem, assuming that the electric field is directed along a body diagonal of the cube.
40. A total charge $5.0 \times 10^{-6} C$ is distributed uniformly throughout a cubical volume whose edges are 8.0 cm long.

- (a) What is the charge density in the cube?
- (b) What is the electric flux through a cube with 12.0-cm edges that is concentric with the charge distribution?
- (c) Do the same calculation for cubes whose edges are 10.0 cm long and 5.0 cm long.
- (d) What is the electric flux through a spherical surface of radius 3.0 cm that is also concentric with the charge distribution?

6.4 Applying Gauss's Law

41. Recall that in the example of a uniform charged sphere, $\rho_0 = Q/(\frac{4}{3}\pi R^3)$. Rewrite the answers in terms of the total charge Q on the sphere.
42. Suppose that the charge density of the spherical charge distribution shown in Figure 6.23 is $\rho(r) = \rho_0 r/R$ for $r \leq R$ and zero for $r > R$. Obtain expressions for the electric field both inside and outside the distribution.
43. A very long, thin wire has a uniform linear charge density of $50\mu C/m$. What is the electric field at a distance 2.0 cm from the wire?
44. A charge of $-30\mu C$ is distributed uniformly throughout a spherical volume of radius 10.0 cm. Determine the electric field due to this charge at a distance of
 - (a) 2.0 cm,
 - (b) 5.0 cm, and
 - (c) 20.0 cm from the center of the sphere.

45. Repeat your calculations for the preceding problem, given that the charge is distributed uniformly over the surface of a spherical conductor of radius 10.0 cm.
46. A total charge Q is distributed uniformly throughout a spherical shell of inner and outer radii r_1 and r_2 , respectively. Show that the electric field due to the charge is

$$\vec{E} = \vec{0} \quad (r \leq r_1);$$

$$\vec{E} = \frac{Q}{4\pi\epsilon_0 r^2} \left(\frac{r^3 - r_1^3}{r_2^3 - r_1^3} \right) \hat{r} \quad (r_1 \leq r \leq r_2);$$

$$\vec{E} = \frac{Q}{4\pi\epsilon_0 r^2} \hat{r} \quad (r \geq r_2).$$

47. When a charge is placed on a metal sphere, it ends up in equilibrium at the outer surface. Use this information to determine the electric field of $+3.0\mu C$ charge put on a 5.0-cm aluminum spherical ball at the following two points in space:

- (a) a point 1.0 cm from the center of the ball (an inside point) and
- (b) a point 10 cm from the center of the ball (an outside point).

48. A large sheet of charge has a uniform charge density of $10\mu C/m^2$. What is the electric field due to this charge at a point just above the surface of the sheet?

49. Determine if approximate cylindrical symmetry holds for the following situations. State why or why not.

- (a) A 300-cm long copper rod of radius 1 cm is charged with $+500$ nC of charge and we seek electric field at a point 5 cm from the center of the rod.
- (b) A 10-cm long copper rod of radius 1 cm is charged with $+500$ nC of charge and we seek electric field at a point 5 cm from the center of the rod.
- (c) A 150-cm wooden rod is glued to a 150-cm plastic rod to make a 300-cm long rod, which is then painted with a charged paint so that one obtains a uniform charge density. The radius of each rod is 1 cm, and we seek an electric field at a point that is 4 cm from the center of the rod.
- (d) Same rod as (c), but we seek electric field at a point that is 500 cm from the center of the rod.

50. A long silver rod of radius 3 cm has a charge of $-5\mu C/cm$ on its surface.

- (a) Find the electric field at a point 5 cm from the center of the rod (an outside point).
- (b) Find the electric field at a point 2 cm from the center of the rod (an inside point).

51. The electric field at 2 cm from the center of long copper rod of radius 1 cm has a magnitude 3 N/C and directed outward from the axis of the rod.

- (a) How much charge per unit length exists on the copper rod?
- (b) What would be the electric flux through a cube of side 5 cm situated such that the rod passes through opposite sides of the cube perpendicularly?

52. A long copper cylindrical shell of inner radius 2 cm and outer radius 3 cm surrounds concentrically a charged long aluminum rod of radius 1 cm with a charge density of 4 pC/m. All charges on the aluminum rod reside at its surface. The inner surface of the copper shell has exactly opposite charge to that of the aluminum rod while the outer surface of the copper shell has the same charge as the aluminum rod. Find the magnitude and direction of the electric field at points that are at the following distances from the center of the aluminum rod:

- (a) 0.5 cm, (b) 1.5 cm, (c) 2.5 cm, (d) 3.5 cm, and (e) 7 cm.

53. Charge is distributed uniformly with a density ρ throughout an infinitely long cylindrical volume of radius R . Show that the field of this charge distribution is directed radially with respect to the cylinder and that

$$E = \frac{\rho r}{2\epsilon_0} \quad (r \leq R);$$

$$E = \frac{\rho R^2}{2\epsilon_0 r} \quad (r \geq R)$$

54. Charge is distributed throughout a very long cylindrical volume of radius R such that the charge density increases with the distance r from the central axis of the cylinder according to $\rho = \alpha r$, where α is a constant. Show that the field of this charge distribution is directed radially with respect to the cylinder and that

$$E = \frac{\alpha r^2}{3\epsilon_0} \quad (r \leq R);$$

$$E = \frac{\alpha R^3}{3\epsilon_0 r} \quad (r \geq R).$$

55. The electric field 10.0 cm from the surface of a copper ball of radius 5.0 cm is directed toward the ball's center and has magnitude $4.0 \times 10^2 \text{ N/C}$. How much charge is on the surface of the ball?

56. Charge is distributed throughout a spherical shell of inner radius r_1 and outer radius r_2 with a volume density given by $\rho = \rho_0 r_1 / r$, where ρ_0 is a constant. Determine the electric field due to this charge as a function of r , the distance from the center of the shell.

57. Charge is distributed throughout a spherical volume of radius R with a density $\rho = \alpha r^2$, where α is a constant. Determine the electric field due to the charge at points both inside and outside the sphere.

58. Consider a uranium nucleus to be sphere of radius $R = 7.4 \times 10^{-15} \text{ m}$ with a charge of $92e$ distributed uniformly throughout its volume. (a) What is the electric force exerted on an electron when it is $3.0 \times 10^{-15} \text{ m}$ from the center of the nucleus? (b) What is the acceleration of the electron at this point?

59. The volume charge density of a spherical charge distribution is given by $\rho(r) = \rho_0 e^{-\alpha r}$, where ρ_0 and α are constants. What is the electric field produced by this charge distribution?

6.5 Conductors in Electrostatic Equilibrium

60. An uncharged conductor with an internal cavity is shown in the following figure. Use the closed surface S along with Gauss' law to show that when a charge q is placed in the cavity a total charge $-q$ is induced on the inner surface of the conductor. What is the charge on the outer surface of the conductor?

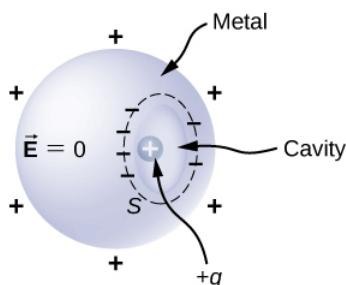
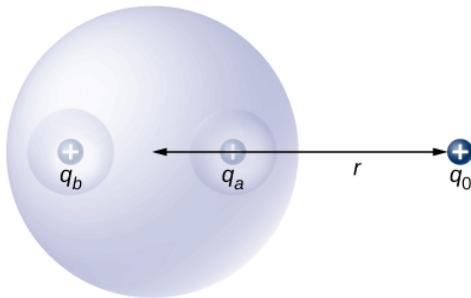


Figure 6.46: A charge inside a cavity of a metal. Charges at the outer surface do not depend on how the charges are distributed at the inner surface since E field inside the body of the metal is zero.

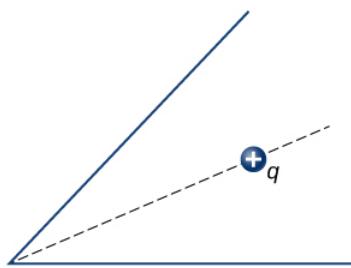
61. An uncharged spherical conductor S of radius R has two spherical cavities **A** and **B** of radii a and b , respectively as shown below. Two point charges $+q_a$ and $+q_b$ are placed at the center of the two cavities by using non-conducting supports. In addition, a point charge $+q_0$ is placed outside at a distance r from the center of the sphere.

(a) Draw approximate charge distributions in the metal although metal sphere has no net charge.

(b) Draw electric field lines. Draw enough lines to represent all distinctly different places.



62. A positive point charge is placed at the angle bisector of two uncharged plane conductors that make an angle of 45° . See below. Draw the electric field lines.



63. A long cylinder of copper of radius 3 cm is charged so that it has a uniform charge per unit length on its surface of 3 C/m .
 (a) Find the electric field inside and outside the cylinder. (b) Draw electric field lines in a plane perpendicular to the rod.

64. An aluminum spherical ball of radius 4 cm is charged with $5\mu\text{C}$ of charge. A copper spherical shell of inner radius 6 cm and outer radius 8 cm surrounds it. A total charge of $-8\mu\text{C}$ is put on the copper shell.

(a) Find the electric field at all points in space, including points inside the aluminum and copper shell when copper shell and aluminum sphere are concentric.

(b) Find the electric field at all points in space, including points inside the aluminum and copper shell when the centers of copper shell and aluminum sphere are 1 cm apart.

65. A long cylinder of aluminum of radius \mathbf{R} meters is charged so that it has a uniform charge per unit length on its surface of λ . (a) Find the electric field inside and outside the cylinder. (b) Plot electric field as a function of distance from the center of the rod.

66. At the surface of any conductor in electrostatic equilibrium, $E = \sigma/\epsilon_0$. Show that this equation is consistent with the fact that $E = kq/r^2$ at the surface of a spherical conductor.

67. Two parallel plates 10 cm on a side are given equal and opposite charges of magnitude $5.0 \times 10^{-9}\text{C}$. The plates are 1.5 mm apart. What is the electric field at the center of the region between the plates?

68. Two parallel conducting plates, each of cross-sectional area 400cm^2 , are 2.0 cm apart and uncharged. If 1.0×10^{12} electrons are transferred from one plate to the other, what are (a) the charge density on each plate? (b) The electric field between the plates?

69. The surface charge density on a long straight metallic pipe is σ . What is the electric field outside and inside the pipe? Assume the pipe has a diameter of $2a$.



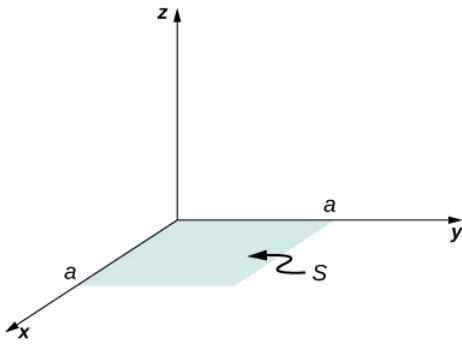
- 70.** A point charge $q = -5.0 \times 10^{-12} C$ is placed at the center of a spherical conducting shell of inner radius 3.5 cm and outer radius 4.0 cm. The electric field just above the surface of the conductor is directed radially outward and has magnitude 8.0 N/C.

- (a) What is the charge density on the inner surface of the shell?
- (b) What is the charge density on the outer surface of the shell?
- (c) What is the net charge on the conductor?

- 71.** A solid cylindrical conductor of radius a is surrounded by a concentric cylindrical shell of inner radius b . The solid cylinder and the shell carry charges $+Q$ and $-Q$, respectively. Assuming that the length L of both conductors is much greater than a or b , determine the electric field as a function of r , the distance from the common central axis of the cylinders, for (a) $r < a$; (b) $a < r < b$; and (c) $r > b$.

Additional Problems

- 72.** A vector field \vec{E} (not necessarily an electric field; note units) is given by $\vec{E} = 3x^2\hat{k}$. Calculate $\int_S \vec{E} \cdot \hat{n} dA$, where S is the area shown below. Assume that $\hat{n} = \hat{k}$.

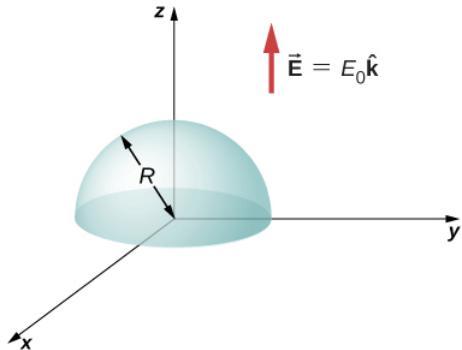


- 73.** Repeat the preceding problem, with $\vec{E} = 2x\hat{i} + 3x^2\hat{k}$.

- 74.** A circular area S is concentric with the origin, has radius a , and lies in the yz -plane. Calculate $\int_S \vec{E} \cdot \hat{n} dA$ for $\vec{E} = 3z^2\hat{i}$.

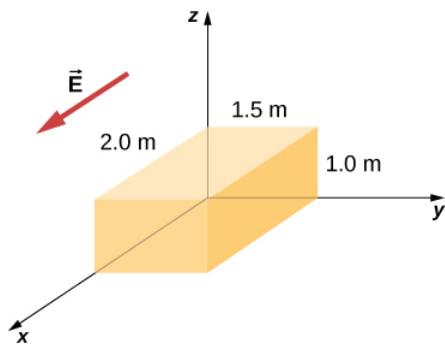
- 75.** (a) Calculate the electric flux through the open hemispherical surface due to the electric field $\vec{E} = E_0\hat{k}$ (see below).

- (b) If the hemisphere is rotated by 90° around the x -axis, what is the flux through it?

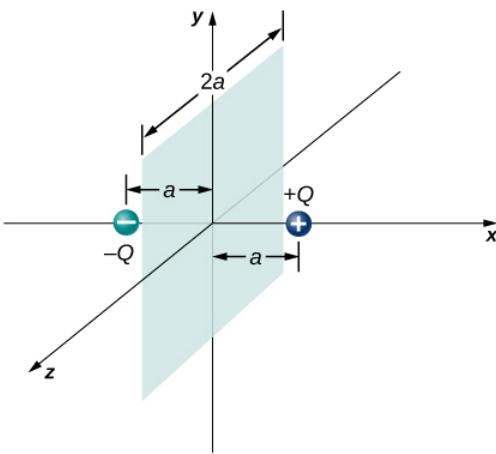


- 76.** Suppose that the electric field of an isolated point charge were proportional to $1/r^{2+\sigma}$ rather than $1/r^2$. Determine the flux that passes through the surface of a sphere of radius R centered at the charge. Would Gauss's law remain valid?

- 77.** The electric field in a region is given by $\vec{E} = a/(b+cx)\hat{i}$, where $a = 200 N \cdot m/C$, $b = 2.0 m$, and $c = 2.0$. What is the net charge enclosed by the shaded volume shown below?



- 78.** Two equal and opposite charges of magnitude Q are located on the x -axis at the points $+a$ and $-a$, as shown below. What is the net flux due to these charges through a square surface of side $2a$ that lies in the yz -plane and is centered at the origin? (**Hint:** Determine the flux due to each charge separately, then use the principle of superposition. You may be able to make a symmetry argument.)



- 79.** A fellow student calculated the flux through the square for the system in the preceding problem and got 0. What went wrong?

- 80.** A **10cm×10cm** piece of aluminum foil of 0.1 mm thickness has a charge of $20\mu C$ that spreads on both wide side surfaces evenly. You may ignore the charges on the thin sides of the edges.

(a) Find the charge density.

(b) Find the electric field 1 cm from the center, assuming approximate planar symmetry.

- 81.** Two **10cm×10cm** pieces of aluminum foil of thickness 0.1 mm face each other with a separation of 5 mm. One of the foils has a charge of $+30\mu C$ and the other has $-30\mu C$.

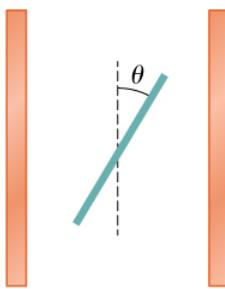
(a) Find the charge density at all surfaces, i.e., on those facing each other and those facing away.

(b) Find the electric field between the plates near the center assuming planar symmetry.

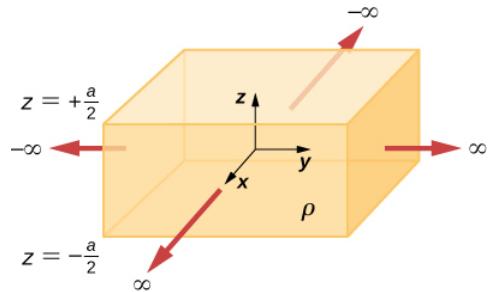
- 82.** Two large copper plates facing each other have charge densities $\pm 4.0 C/m^2$ on the surface facing the other plate, and zero in between the plates. Find the electric flux through a **3cm×4cm** rectangular area between the plates, as shown below, for the following orientations of the area.

(a) If the area is parallel to the plates, and

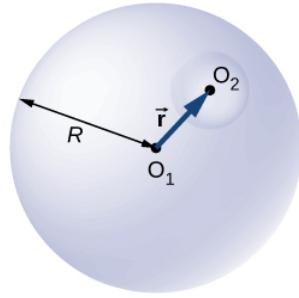
(b) if the area is tilted $\theta = 30^\circ$ from the parallel direction. Note, this angle can also be $\theta = 180^\circ + 30^\circ$.



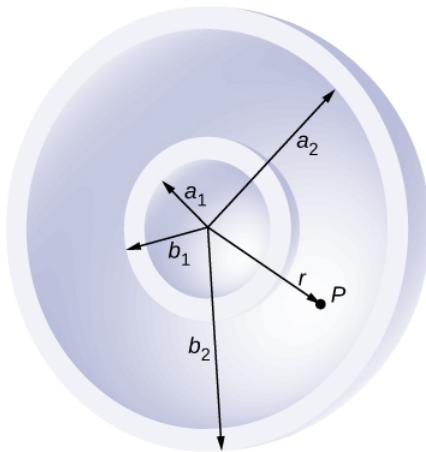
- 83.** The infinite slab between the planes defined by $z = -a/2$ and $z = a/2$ contains a uniform volume charge density ρ (see below). What is the electric field produced by this charge distribution, both inside and outside the distribution?



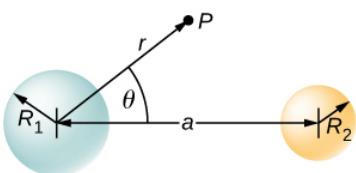
- 84.** A total charge \mathbf{Q} is distributed uniformly throughout a spherical volume that is centered at O_1 and has a radius R . Without disturbing the charge remaining, charge is removed from the spherical volume that is centered at O_2 (see below). Show that the electric field everywhere in the empty region is given by $\vec{E} = \frac{Q\vec{r}}{4\pi\epsilon_0 R^3}$, where \vec{r} is the displacement vector directed from O_1 to O_2 .



- 85.** A non-conducting spherical shell of inner radius a_1 and outer radius b_1 is uniformly charged with charge density ρ_1 inside another non-conducting spherical shell of inner radius a_2 and outer radius b_2 that is also uniformly charged with charge density ρ_2 . See below. Find the electric field at space point \mathbf{P} at a distance r from the common center such that (a) $r > b_2$, (b) $a_2 < r < b_2$, (c) $b_1 < r < a_2$, (d) $a_1 < r < b_1$, and (e) $r < a_1$.

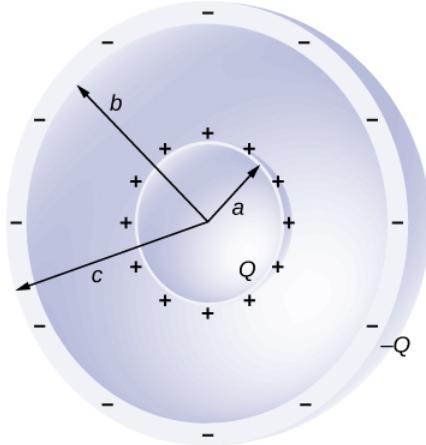


86. Two non-conducting spheres of radii R_1 and R_2 are uniformly charged with charge densities ρ_1 and ρ_2 , respectively. They are separated at center-to-center distance a (see below). Find the electric field at point P located at a distance r from the center of sphere 1 and is in the direction θ from the line joining the two spheres assuming their charge densities are not affected by the presence of the other sphere. (**Hint:** Work one sphere at a time and use the superposition principle.)

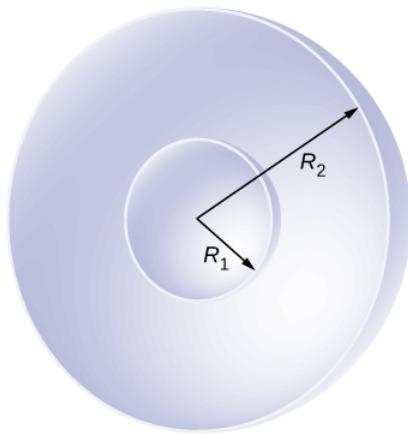


87. A disk of radius \mathbf{R} is cut in a non-conducting large plate that is uniformly charged with charge density σ (coulomb per square meter). See below. Find the electric field at a height h above the center of the disk. ($h \gg R, h \ll l$ or w). (**Hint:** Fill the hole with $\pm\sigma$.)

88. Concentric conducting spherical shells carry charges Q and $-Q$, respectively (see below). The inner shell has negligible thickness. Determine the electric field for (a) $r < a$; (b) $a < r < b$; (c) $b < r < c$; and (d) $r > c$.



89. Shown below are two concentric conducting spherical shells of radii R_1 and R_2 , each of finite thickness much less than either radius. The inner and outer shell carry net charges q_1 and q_2 , respectively, where both q_1 and q_2 are positive. What is the electric field for (a) $r < R_1$; (b) $R_1 < r < R_2$; and (c) $r > R_2$? (d) What is the net charge on the inner surface of the inner shell, the outer surface of the inner shell, the inner surface of the outer shell, and the outer surface of the outer shell?



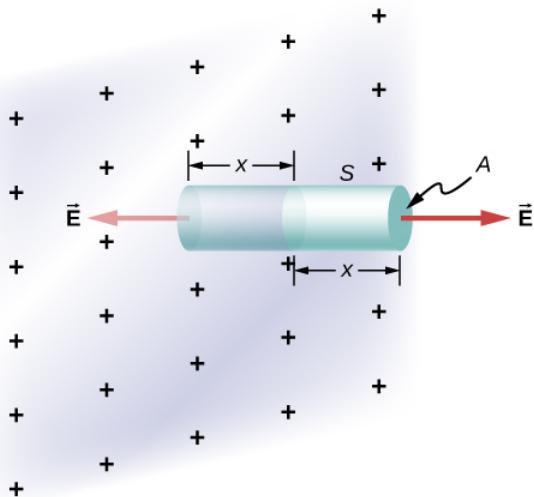
- 90.** A point charge of $q=5.0\times 10^{-8}\text{C}$ is placed at the center of an uncharged spherical conducting shell of inner radius 6.0 cm and outer radius 9.0 cm. Find the electric field at (a) $r=4.0\text{cm}$, (b) $r=8.0\text{cm}$, and (c) $r=12.0\text{cm}$. (d) What are the charges induced on the inner and outer surfaces of the shell?

Challenge Problems

91. The Hubble Space Telescope can measure the energy flux from distant objects such as supernovae and stars. Scientists then use this data to calculate the energy emitted by that object. Choose an interstellar object which scientists have observed the flux at the Hubble with (for example, *Vega*³), find the distance to that object and the size of Hubble's primary mirror, and calculate the total energy flux. (**Hint:** The Hubble intercepts only a small part of the total flux.)

92. Re-derive Gauss's law for the gravitational field, with \vec{g} directed positively outward.

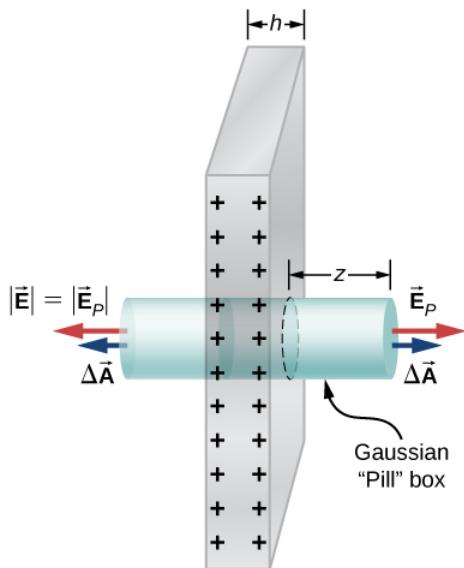
93. An infinite plate sheet of charge of surface charge density σ is shown below. What is the electric field at a distance x from the sheet? Compare the result of this calculation with that of worked out in the text.



- 94.** A spherical rubber balloon carries a total charge Q distributed uniformly over its surface. At $t = 0$, the radius of the balloon is R . The balloon is then slowly inflated until its radius reaches $2R$ at the time t_0 . Determine the electric field due to this charge as a function of time

- (a) at the surface of the balloon,
- (b) at the surface of radius R , and
- (c) at the surface of radius $2R$. Ignore any effect on the electric field due to the material of the balloon and assume that the radius increases uniformly with time.

95. Find the electric field of a large conducting plate containing a net charge \mathbf{q} . Let \mathbf{A} be area of one side of the plate and \mathbf{h} the thickness of the plate (see below). The charge on the metal plate will distribute mostly on the two planar sides and very little on the edges if the plate is thin.



This page titled [17.7: Gauss's Law \(Exercises\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.7: Gauss's Law \(Exercises\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

17.8: Gauss's Law (Answers)

Check Your Understanding

6.1. Place it so that its unit normal is perpendicular to \vec{E} .

6.2. $mab^2/2$

6.3 a. $3.4 \times 10^5 N \cdot m^2/C$;

b. $-3.4 \times 10^5 N \cdot m^2/C$;

c. $3.4 \times 10^5 N \cdot m^2/C$;

d. 0

6.4. In this case, there is only \vec{E}_{out} . So, yes.

6.5. $\vec{E} = \frac{\lambda_0}{2\pi\epsilon_0} \frac{1}{d} \hat{r}$; This agrees with the calculation of Example 5.5 where we found the electric field by integrating over the charged wire. Notice how much simpler the calculation of this electric field is with Gauss's law.

6.6. If there are other charged objects around, then the charges on the surface of the sphere will not necessarily be spherically symmetrical; there will be more in certain directions than in other directions.

Conceptual Questions

1. a. If the planar surface is perpendicular to the electric field vector, the maximum flux would be obtained. b. If the planar surface were parallel to the electric field vector, the minimum flux would be obtained.

3. true

5. Since the electric field vector has a $\frac{1}{r^2}$ dependence, the fluxes are the same since $A = 4\pi r^2$.

7. a. no;

b. zero

9. Both fields vary as $\frac{1}{r^2}$. Because the gravitational constant is so much smaller than $\frac{1}{4\pi\epsilon_0}$, the gravitational field is orders of magnitude weaker than the electric field.

11. No, it is produced by all charges both inside and outside the Gaussian surface.

13. No, since the situation does not have symmetry, making Gauss's law challenging to simplify.

15. Any shape of the Gaussian surface can be used. The only restriction is that the Gaussian integral must be calculable; therefore, a box or a cylinder are the most convenient geometrical shapes for the Gaussian surface.

17. yes

19. Since the electric field is zero inside a conductor, a charge of $-2.0\mu C$ is induced on the inside surface of the cavity. This will put a charge of $+2.0\mu C$ on the outside surface leaving a net charge of $-3.0\mu C$ on the surface.

Problems

21. $\Phi = \vec{E} \cdot \vec{A} \rightarrow EA\cos\theta = 2.2 \times 10^4 N \cdot m^2/C$ electric field in direction of unit normal; $\Phi = \vec{E} \cdot \vec{A} \rightarrow EA\cos\theta = -2.2 \times 10^4 N \cdot m^2/C$ electric field opposite to unit normal

23. $\frac{3 \times 10^{-5} N \cdot m^2/C}{(0.05m)^2} = E \Rightarrow \sigma = 2.12 \times 10^{-13} C/m^2$

25. a. $\Phi = 0.17 N \cdot m^2/C$;

b. $\Phi = 0$;

c. $\Phi = EA\cos 0^\circ = 1.0 \times 10^3 N/C (2.0 \times 10^{-4} m)^2 \cos 0^\circ = 0.20 N \cdot m^2/C$

27. $\Phi = 3.8 \times 10^4 N \cdot m^2/C$

29. $\vec{E}(z) = \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{z} \hat{k}, \int \vec{E} \cdot \hat{n} dA = \frac{\lambda}{\epsilon_0} l$

31. a. $\Phi = 3.39 \times 10^3 N \cdot m^2/C$;

b. $\Phi = 0$;

c. $\Phi = -2.25 \times 10^5 N \cdot m^2/C$;

d. $\Phi = 90.4 N \cdot m^2/C$

33. $\Phi = 1.13 \times 10^6 N \cdot m^2/C$

35. Make a cube with \mathbf{q} at the center, using the cube of side a . This would take four cubes of side a to make one side of the large cube. The shaded side of the small cube would be 1/24th of the total area of the large cube; therefore, the flux through the shaded area would be $\Phi = \frac{1}{24} \frac{q}{\epsilon_0}$.

37. $q = 3.54 \times 10^{-7} C$

39. zero, also because flux in equals flux out

41. $r > R, E = \frac{Q}{4\pi\epsilon_0 r^2}; r < R, E = \frac{qr}{4\pi\epsilon_0 R^3}$

43. $EA = \frac{\lambda l}{\epsilon_0} \Rightarrow E = 4.50 \times 10^7 N/C$

45. a. 0;

b. 0;

c. $\vec{E} = 6.74 \times 10^6 N/C(-\hat{r})$

47. a. 0;

b. $E = 2.70 \times 10^6 N/C$

49. a. Yes, the length of the rod is much greater than the distance to the point in question.

b. No, The length of the rod is of the same order of magnitude as the distance to the point in question.

c. Yes, the length of the rod is much greater than the distance to the point in question.

d. No. The length of the rod is of the same order of magnitude as the distance to the point in question.

51. a. $\vec{E} = \frac{R\sigma_0}{\epsilon_0} \frac{1}{r} \hat{r} \Rightarrow \sigma_0 = 5.31 \times 10^{-11} C/m^2, \lambda = 3.33 \times 10^{-12} C/m$;

b. $\Phi = \frac{q_{enc}}{\epsilon_0} = \frac{3.33 \times 10^{-12} C/m(0.05m)}{\epsilon + 0} = 0.019 N \cdot m^2/C$

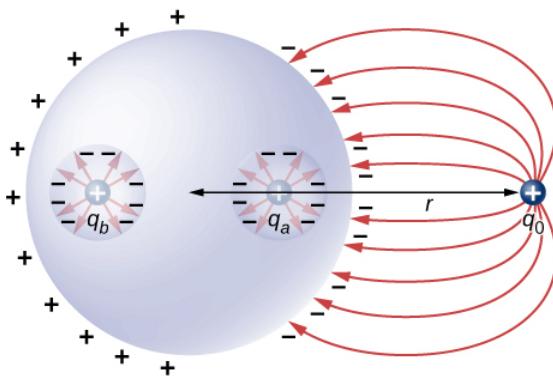
53. $E2\pi rl = \frac{\rho\pi r^2 l}{\epsilon_0} \Rightarrow E = \frac{\rho r}{2\epsilon_0} (r \leq R); E2\pi rl = \frac{\rho\pi R^2 l}{\epsilon_0} \Rightarrow E = \frac{\rho R^2}{2\epsilon_0 r} (r \geq R)$

55. $\Phi = \frac{q_{enc}}{\epsilon_0} \Rightarrow q_{enc} = -1.0 \times 10^{-9} C$

57. $q_{enc} = \frac{4}{5} \pi \alpha r^5, E4\pi r^2 = \frac{4\pi\alpha r^5}{5\epsilon_0} \Rightarrow E = \frac{\alpha r^3}{5\epsilon_0} (r \leq R), q_{enc} = \frac{4}{5} \pi \alpha R^5, E4\pi r^2 = \frac{4\pi\alpha R^5}{5\epsilon_0} \Rightarrow E = \frac{\alpha R^5}{5\epsilon_0 r^2} (r \geq R)$

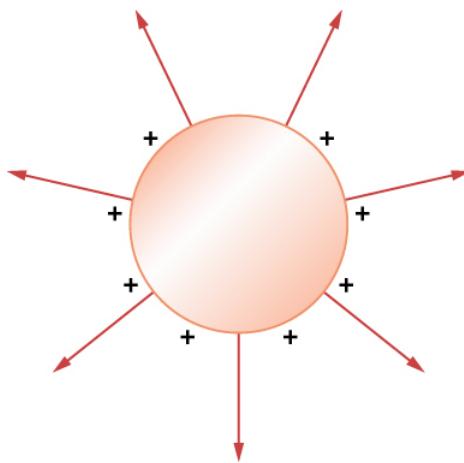
59. integrate by parts: $q_{enc} = 4\pi\rho_0 [-e^{-\alpha r} (\frac{(r)^2}{\alpha} + \frac{2r}{\alpha^2} + \frac{2}{\alpha^3}) + \frac{2}{\alpha^3}] \Rightarrow E = \frac{\rho_0}{r^2\epsilon_0} [-e^{-\alpha r} (\frac{(r)^2}{\alpha} + \frac{2r}{\alpha^2} + \frac{2}{\alpha^3}) + \frac{2}{\alpha^3}]$

61.



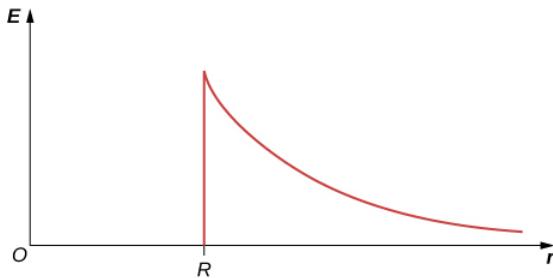
63. a. Outside: $E2\pi rl = \frac{\lambda l}{\epsilon_0} \Rightarrow E = \frac{3.0C/m}{2\pi\epsilon_0 r}$; Inside $E_{in} = 0$;

b.



65. a. $E2\pi rl = \frac{\lambda l}{\epsilon_0} \Rightarrow E = \frac{\lambda}{2\pi\epsilon_0 r}r \geq R$ E inside equals 0;

b.



67. $E = 5.65 \times 10^4 N/C$

69. $\lambda = \frac{\lambda l}{\epsilon_0} \Rightarrow E = \frac{a\sigma}{\epsilon_0 r}r \geq a$, $E = 0$ inside since $q_{\text{enclosed}} = 0$

71. a. $E = 0$;

b. $E2\pi rL = \frac{Q}{\epsilon_0} \Rightarrow E = \frac{Q}{2\pi\epsilon_0 rL}$; c. $E = 0$ since r would be either inside the second shell or if outside then q enclosed equals 0.

Additional Problems

73. $\int \vec{E} \cdot \hat{n} dA = a^4$

75. a. $\int \vec{E} \cdot \hat{n} dA = E_0 r^2 \pi$; b. zero, since the flux through the upper half cancels the flux through the lower half of the sphere

77. $\Phi = \frac{q_{enc}}{\epsilon_0}$; There are two contributions to the surface integral: one at the side of the rectangle at $x = 0$ and the other at the side at $x = 2.0m$; $-E(0)[1.5m^2] + E(2.0m)[1.5m^2] = \frac{q_{enc}}{\epsilon_0} = -100Nm2/C$

where the minus sign indicates that at $x = 0$, the electric field is along positive \mathbf{x} and the unit normal is along negative \mathbf{x} . At $x = 2$, the unit normal and the electric field vector are in the same direction: $q_{enc} = \epsilon_0 \Phi = -8.85 \times 10^{-10} C$

79. didn't keep consistent directions for the area vectors, or the electric fields

81. a. $\sigma = 3.0 \times 10^{-3} C/m^2, +3 \times 10^{-3} C/m^2$ on one and $-3 \times 10^{-3} C/m^2$ on the other;

$$b. E = 3.39 \times 10^8 N/CE = 3.39 \times 10^8 N/C$$

83. Construct a Gaussian cylinder along the \mathbf{z} -axis with cross-sectional area \mathbf{A} .

$$|z| \geq \frac{a}{2} q_{enc} = \rho A a, \Phi = \frac{\rho A a}{\epsilon_0} \Rightarrow E = \frac{\rho a}{2\epsilon_0},$$

$$|z| \leq \frac{a}{2} q_{enc} = \rho A 2z, E(2A) = \frac{\rho A 2z}{\epsilon_0} \Rightarrow E = \frac{\rho z}{\epsilon_0}$$

$$85. a. r > b_2 E 4\pi r^2 = \frac{\frac{4}{3}\pi[\rho_1(b_1^3 - a_1^3) + \rho_2(b_2^3 - a_2^3)]}{\epsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3 - a_1^3) + \rho_2(b_2^3 - a_2^3)}{3\epsilon_0 r^2};$$

$$b. a_2 < r < b_2 E 4\pi r^2 = \frac{\frac{4}{3}\pi[\rho_1(b_1^3 - a_1^3) + \rho_2(r^3 - a_2^3)]}{\epsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3 - a_1^3) + \rho_2(r^3 - a_2^3)}{3\epsilon_0 r^2};$$

$$c. b_1 < r < a_2 E 4\pi r^2 = \frac{\frac{4}{3}\pi\rho_1(b_1^3 - a_1^3)}{\epsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3 - a_1^3)}{3\epsilon_0 r^2};$$

$$d. a_1 < r < b_1 E 4\pi r^2 = \frac{\frac{4}{3}\pi\rho_1(r^3 - a_1^3)}{\epsilon_0} \Rightarrow E = \frac{\rho_1(r^3 - a_1^3)}{3\epsilon_0 r^2};$$

e. 0

87. Electric field due to plate without hole: $E = \frac{\sigma}{2\epsilon_0}$.

Electric field of just hole filled with $-\sigma$: $E = \frac{-\sigma}{2\epsilon_0} \frac{h}{\sqrt{R^2 + h^2}}$.

Thus, $E_{net} = \frac{\sigma}{2\epsilon_0} \frac{h}{\sqrt{R^2 + h^2}}$

89. a. $E = 0$; b. $E = \frac{q_1}{4\pi\epsilon_0 r^2}$; c. $E = \frac{q_1 + q_2}{4\pi\epsilon_0 r^2}$; d. 0 $q_1 - q_1, q_1 + q_2$

Challenge Problems

91. Given the referenced link, using a distance to Vega of $237 \times 10^{15} m^4$ and a diameter of 2.4 m for the primary mirror,⁵ we find that at a wavelength of 555.6 nm, Vega is emitting $2.44 \times 10^{24} J/s$ at that wavelength. Note that the flux through the mirror is essentially constant.

93. The symmetry of the system forces \vec{E} to be perpendicular to the sheet and constant over any plane parallel to the sheet. To calculate the electric field, we choose the cylindrical Gaussian surface shown. The cross-section area and the height of the cylinder are \mathbf{A} and $2x$, respectively, and the cylinder is positioned so that it is bisected by the plane sheet. Since \mathbf{E} is perpendicular to each end and parallel to the side of the cylinder, we have \mathbf{EA} as the flux through each end and there is no flux through the side. The charge enclosed by the cylinder is σA , so from Gauss's law, $2\mathbf{EA} = \frac{\sigma A}{\epsilon_0}$, and the electric field of an infinite sheet of charge is

$$E = \frac{\sigma}{2\epsilon_0}, \text{ in agreement with the calculation of in the text.}$$

95. There is $Q/2$ on each side of the plate since the net charge is $Q : \sigma = \frac{Q}{2A}$,

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{2\sigma \Delta A}{\epsilon_0} \Rightarrow E_P = \frac{\sigma}{\epsilon_0} = \frac{Q}{\epsilon_0 2A}$$

This page titled [17.8: Gauss's Law \(Answers\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.8: Gauss's Law \(Answers\)](#) by [OpenStax](#) is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

CHAPTER OVERVIEW

18: Calculation of Magnetic Quantities from Currents

- 18.1: Introduction
- 18.2: Magnetic Field due to a Thin Straight Wire
- 18.3: Magnetic Field of a Current Loop
- 18.4: Magnetic Field using Ampère's Law
- 18.5: Magnetic Field of Solenoids and Toroids
- 18.6: Magnetic Force between Two Parallel Currents
- 18.7: (edit) Magnetic Force and Torque on a Current Loop - Motors and Meters
- 18.8: Magnetic Forces in a Conductor - The Hall Effect
- 18.9: More Applications of Magnetism
- 18.10: Superconductors
- 18.11: Conclusion
- 18.12: Magnetic Forces and Fields (Summary)
- 18.13: Sources of Magnetic Fields (Summary)
- 18.14: Current and Resistance (Summary)
- 18.15: Magnetic Forces and Fields (Exercise)
- 18.16: Sources of Magnetic Fields (Exercise)
- 18.17: Magnetic Forces and Fields (Answers)
- 18.18: Sources of Magnetic Fields (Answers)

18: Calculation of Magnetic Quantities from Currents is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

18.1: Introduction

In an earlier chapter, we saw that a moving charged particle produces a magnetic field. This connection between electricity and magnetism is exploited in electromagnetic devices, such as a computer hard drive. In fact, it is the underlying principle behind most of the technology in modern society, including telephones, television, computers, and the internet.



Figure 18.1.1: An external hard drive attached to a computer works by magnetically encoding information that can be stored or retrieved quickly. A key idea in the development of digital devices is the ability to produce and use magnetic fields in this way. (credit: modification of work by "Miss Karen"/Flickr)

In this chapter, we examine how magnetic fields are created by arbitrary distributions of electric current, using the Biot-Savart law. Then we look at how current-carrying wires create magnetic fields and deduce the forces that arise between two current-carrying wires due to these magnetic fields. We also study the torques produced by the magnetic fields of current loops. We then generalize these results to an important law of electromagnetism, called Ampère's law.

We examine some devices that produce magnetic fields from currents in geometries based on loops, known as solenoids and toroids. Finally, we look at how materials behave in magnetic fields and categorize materials based on their responses to magnetic fields.

This page titled [18.1: Introduction](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.1: Prelude to Sources of Magnetic Fields](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.2: Magnetic Field due to a Thin Straight Wire

Learning Objectives

By the end of this section, you will be able to:

- Explain how the Biot-Savart law is used to determine the magnetic field due to a thin, straight wire.
- Determine the dependence of the magnetic field from a thin, straight wire based on the distance from it and the current flowing in the wire.
- Sketch the magnetic field created from a thin, straight wire by using the second right-hand rule.

How much current is needed to produce a significant magnetic field, perhaps as strong as Earth's field? Surveyors will tell you that overhead electric power lines create magnetic fields that interfere with their compass readings. Indeed, when Oersted discovered in 1820 that a current in a wire affected a compass needle, he was not dealing with extremely large currents. How does the shape of wires carrying current affect the shape of the magnetic field created? We noted in Chapter 28 that a current loop created a magnetic field similar to that of a bar magnet, but what about a straight wire? We can use the [Biot-Savart law](#) to answer all of these questions, including determining the magnetic field of a long straight wire.

Figure 18.2.1 shows a section of an infinitely long, straight wire that carries a current \mathbf{I} . What is the magnetic field at a point \mathbf{P} , located a distance \mathbf{R} from the wire?

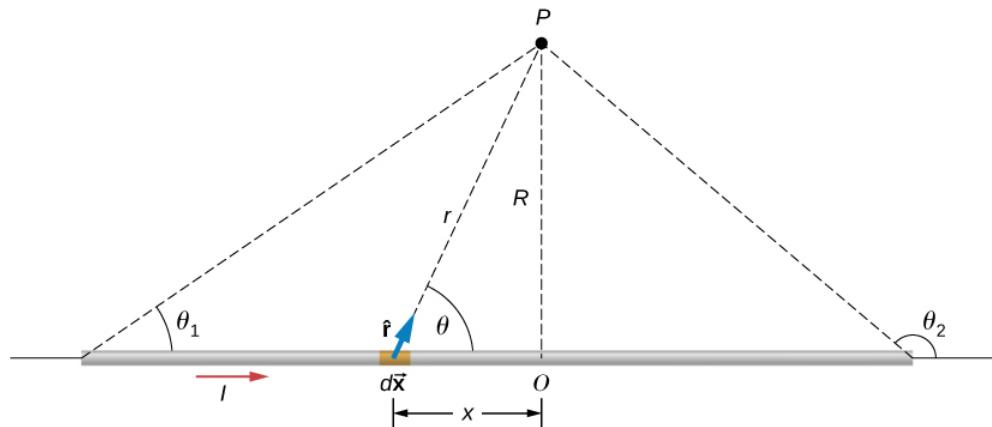


Figure 18.2.1: A section of a thin, straight current-carrying wire. The independent variable θ has the limits θ_1 and θ_2 .

Let's begin by considering the magnetic field due to the current element $I d\vec{x}$ located at the position \mathbf{x} . Using the right-hand rule 1 from the previous chapter, $d\vec{x} \times \hat{r}$ points out of the page for any element along the wire. At point \mathbf{P} , therefore, the magnetic fields due to all current elements have the same direction. This means that we can calculate the net field there by evaluating the scalar sum of the contributions of the elements. With

$$|d\vec{x} \times \hat{r}| = (dx)(1) \sin \theta$$

we have from the **Biot-Savart law**

$$B = \frac{\mu_0}{4\pi} \int_{\text{wire}} \frac{I \sin \theta dx}{r^2}. \quad (18.2.1)$$

The wire is symmetrical about point O , so we can set the limits of the integration from zero to infinity and double the answer, rather than integrate from negative infinity to positive infinity. Based on the picture and trigonometry, we can write expressions for r and $\sin \theta$ in terms of x and \mathbf{R} , namely:

$$r = \sqrt{x^2 + R^2}$$

$$\sin \theta = \frac{R}{\sqrt{x^2 + R^2}}.$$

Substituting these expressions into Equation 18.2.1, the magnetic field integration becomes

$$B = \frac{\mu_0 I}{2\pi} \int_0^\infty \frac{R dx}{(x^2 + R^2)^{3/2}}.$$

Evaluating the integral yields

$$B = \frac{\mu_0 I}{2\pi R} \left[\frac{x}{(x^2 + R^2)^{1/2}} \right]_0^\infty.$$

Substituting the limits gives us the solution

$$B = \frac{\mu_0 I}{2\pi R}.$$

The magnetic field lines of the infinite wire are circular and centered at the wire (Figure 18.2.2), and they are identical in every plane perpendicular to the wire. Since the field decreases with distance from the wire, the spacing of the field lines must increase correspondingly with distance. The direction of this magnetic field may be found with a second form of the **right-hand rule** (Figure 18.2.2). If you hold the wire with your right hand so that your thumb points along the current, then your fingers wrap around the wire in the same sense as \vec{B} .

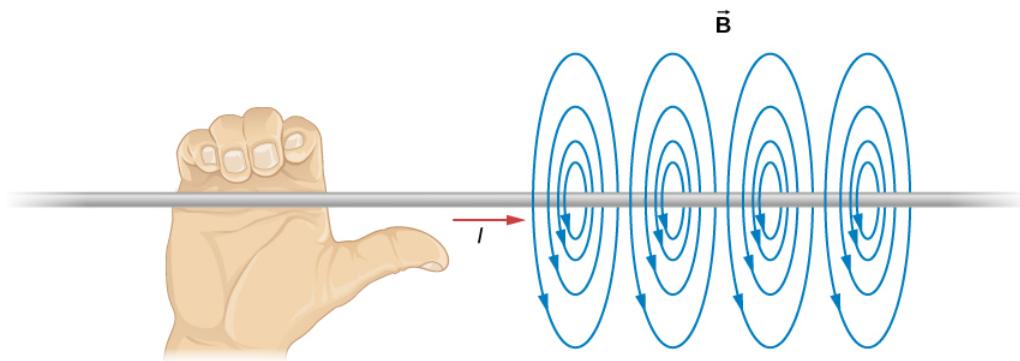


Figure 18.2.2: Some magnetic field lines of an infinite wire. The direction of B can be found with a form of the right-hand rule.

The direction of the field lines can be observed experimentally by placing several small compass needles on a circle near the wire, as illustrated in Figure 18.2.3a. When there is no current in the wire, the needles align with Earth's magnetic field. However, when a large current is sent through the wire, the compass needles all point tangent to the circle. Iron filings sprinkled on a horizontal surface also delineate the field lines, as shown in Figure 18.2.3b.

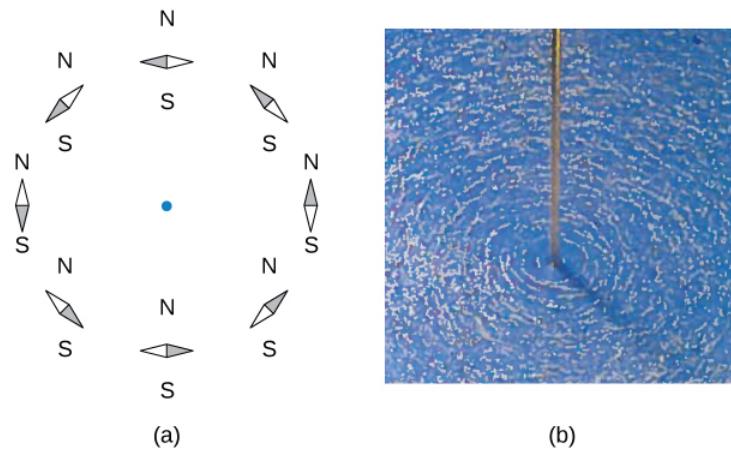


Figure 18.2.3: The shape of the magnetic field lines of a long wire can be seen using (a) small compass needles and (b) iron filings.

✓ Example 18.2.1: Calculating Magnetic Field Due to Three Wires

Three wires sit at the corners of a square, all carrying currents of 2 amps into the page as shown in Figure 18.2.4. Calculate the magnitude of the magnetic field at the other corner of the square, point **P**, if the length of each side of the square is 1 cm.

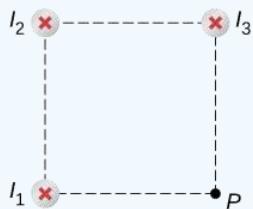


Figure 18.2.4: Three wires have current flowing into the page. The magnetic field is determined at the fourth corner of the square.

Strategy

The magnetic field due to each wire at the desired point is calculated. The diagonal distance is calculated using the Pythagorean theorem. Next, the direction of each magnetic field's contribution is determined by drawing a circle centered at the point of the wire and out toward the desired point. The direction of the magnetic field contribution from that wire is tangential to the curve. Lastly, working with these vectors, the resultant is calculated.

Solution

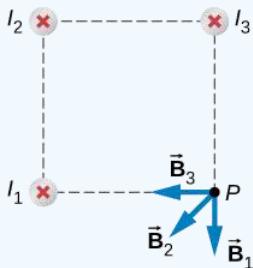
Wires 1 and 3 both have the same magnitude of magnetic field contribution at point **P**:

$$B_1 = B_3 = \frac{\mu_0 I}{2\pi R} = \frac{(4\pi \times 10^{-7} T \cdot m/A)(2 A)}{2\pi(0.01 m)} = 4 \times 10^{-5} T.$$

Wire 2 has a longer distance and a magnetic field contribution at point **P** of:

$$B_2 = \frac{\mu_0 I}{2\pi R} = \frac{(4\pi \times 10^{-7} T \cdot m/A)(2 A)}{2\pi(0.01414 m)} = 3 \times 10^{-5} T.$$

The vectors for each of these magnetic field contributions are shown.



The magnetic field in the x-direction has contributions from wire 3 and the x-component of wire 2:

$$B_{net x} = -4 \times 10^{-5} T - 2.83 \times 10^{-5} T \cos(45^\circ) = -6 \times 10^{-5} T.$$

The y-component is similarly the contributions from wire 1 and the y-component of wire 2:

$$B_{net y} = -4 \times 10^{-5} T - 2.83 \times 10^{-5} T \sin(45^\circ) = -6 \times 10^{-5} T.$$

Therefore, the net magnetic field is the resultant of these two components:

$$\begin{aligned} B_{net} &= \sqrt{B_{net x}^2 + B_{net y}^2} \\ &= \sqrt{(-6 \times 10^{-5} T)^2 + (-6 \times 10^{-5} T)^2} \\ &= 8.48 \times 10^{-5} T. \end{aligned}$$

Significance

The geometry in this problem results in the magnetic field contributions in the **x**- and **y**-directions having the same magnitude. This is not necessarily the case if the currents were different values or if the wires were located in different positions. Regardless of the numerical results, working on the components of the vectors will yield the resulting magnetic field at the point in need.

? Exercise 18.2.1

Using Example 18.2.1, keeping the currents the same in wires 1 and 3, what should the current be in wire 2 to counteract the magnetic fields from wires 1 and 3 so that there is no net magnetic field at point **P**?

Solution

4 amps flowing out of the page

This page titled [18.2: Magnetic Field due to a Thin Straight Wire](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.3: Magnetic Field due to a Thin Straight Wire](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

18.3: Magnetic Field of a Current Loop

Learning Objectives

By the end of this section, you will be able to:

- Explain how the Biot-Savart law is used to determine the magnetic field due to a current in a loop of wire at a point along a line perpendicular to the plane of the loop.
- Determine the magnetic field of an arc of current.

The circular loop of Figure 18.3.1 has a radius \mathbf{R} , carries a current \mathbf{I} , and lies in the xz -plane. What is the magnetic field due to the current at an arbitrary point \mathbf{P} along the axis of the loop?

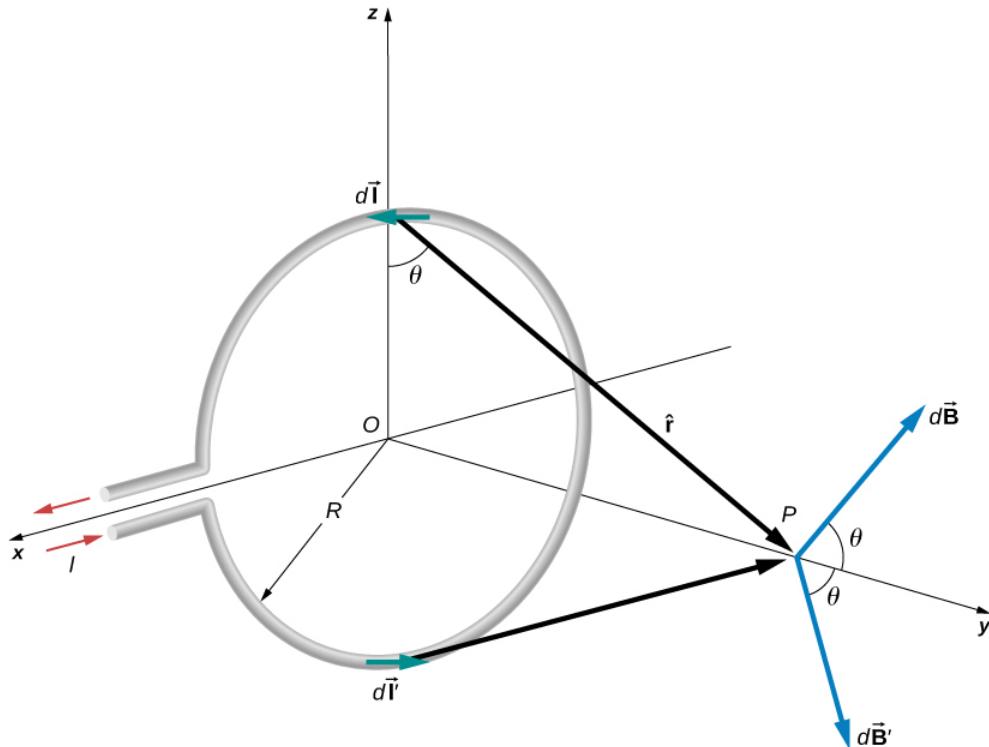


Figure 18.3.1: Determining the magnetic field at point \mathbf{P} along the axis of a current-carrying loop of wire.

We can use the Biot-Savart law to find the magnetic field due to a current. We first consider arbitrary segments on opposite sides of the loop to qualitatively show by the vector results that the net magnetic field direction is along the central axis from the loop. From there, we can use the Biot-Savart law to derive the expression for magnetic field.

Let \mathbf{P} be a distance y from the center of the loop. From the right-hand rule, the magnetic field $d\vec{\mathbf{B}}$ at \mathbf{P} , produced by the current element $I d\vec{l}$ is directed at an angle θ above the y -axis as shown. Since $d\vec{l}$ is parallel along the x -axis and $\hat{\mathbf{r}}$ is in the yz -plane, the two vectors are perpendicular, so we have

$$dB = \frac{\mu_0}{4\pi} \frac{I dl \sin \pi/2}{r^2} = \frac{\mu_0}{4\pi} \frac{I dl}{y^2 + R^2} \quad (18.3.1)$$

where we have used $r^2 = y^2 + R^2$.

Now consider the magnetic field $d\vec{\mathbf{B}}'$ due to the current element $I d\vec{l}'$, which is directly opposite $I d\vec{l}$ on the loop. The magnitude of $d\vec{\mathbf{B}}'$ is also given by Equation 18.3.1, but it is directed at an angle θ below the y -axis. The components of $d\vec{\mathbf{B}}$ and $d\vec{\mathbf{B}}'$ perpendicular to the y -axis therefore cancel, and in calculating the net magnetic field, only the components along the y -axis need to be considered. The components perpendicular to the axis of the loop sum to zero in pairs. Hence at point \mathbf{P} :

$$\vec{B} = \hat{j} \int_{loop} dB \cos \theta = \hat{j} \frac{\mu_0 I}{4\pi} \int_{loop} \frac{\cos \theta dl}{y^2 + R^2}. \quad (18.3.2)$$

For all elements $d\vec{l}$ on the wire, y , \mathbf{R} , and θ are constant and are related by

$$\cos \theta = \frac{R}{\sqrt{y^2 + R^2}}.$$

Now from Equation 18.3.2, the magnetic field at \mathbf{P} is

$$\vec{B} = \hat{j} \frac{\mu_0 IR}{4\pi(y^2 + R^2)^{3/2}} \int_{loop} dl = \frac{\mu_0 IR^2}{2(y^2 + R^2)^{3/2}} \hat{j} \quad (18.3.3)$$

where we have used $\int_{loop} dl = 2\pi R$. As discussed in the previous chapter, the closed current loop is a magnetic dipole of moment $\vec{\mu} = IA\hat{n}$. For this example, $A = \pi R^2$ and $\hat{n} = \hat{j}$, so the magnetic field at \mathbf{P} can also be written as

$$\vec{B} = \frac{\mu_0 \mu \hat{j}}{2\pi(y^2 + R^2)^{3/2}}. \quad (18.3.4)$$

By setting $y = 0$ in Equation 18.3.4, we obtain the magnetic field at the center of the loop:

✓ Note

$$\vec{B} = \frac{\mu_0 I}{2R} \hat{j}. \quad (18.3.5)$$

This equation becomes $B = \mu_0 n I / (2R)$ for a flat coil of n loops per length. It can also be expressed as

$$\vec{B} = \frac{\mu_0 \vec{\mu}}{2\pi R^3}. \quad (18.3.6)$$

If we consider $y \gg R$ in Equation 18.3.4, the expression reduces to an expression known as the magnetic field from a dipole:

$$\vec{B} = \frac{\mu_0 \vec{\mu}}{2\pi y^3}. \quad (18.3.7)$$

The calculation of the magnetic field due to the circular current loop at points off-axis requires rather complex mathematics, so we'll just look at the results. The magnetic field lines are shaped as shown in Figure 18.3.2. Notice that one field line follows the axis of the loop. This is the field line we just found. Also, very close to the wire, the field lines are almost circular, like the lines of a long straight wire.

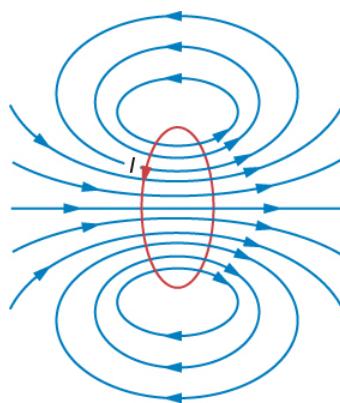


Figure 18.3.2: Sketch of the magnetic field lines of a circular current loop.

✓ Magnetic Field between Two Loops

Two loops of wire carry the same current of 10 mA, but flow in opposite directions as seen in Figure 18.3.3. One loop is measured to have a radius of $R = 50 \text{ cm}$ while the other loop has a radius of $2R = 100 \text{ cm}$. The distance from the first loop to the

point where the magnetic field is measured is 0.25 m, and the distance from that point to the second loop is 0.75 m. What is the magnitude of the net magnetic field at point **P**?

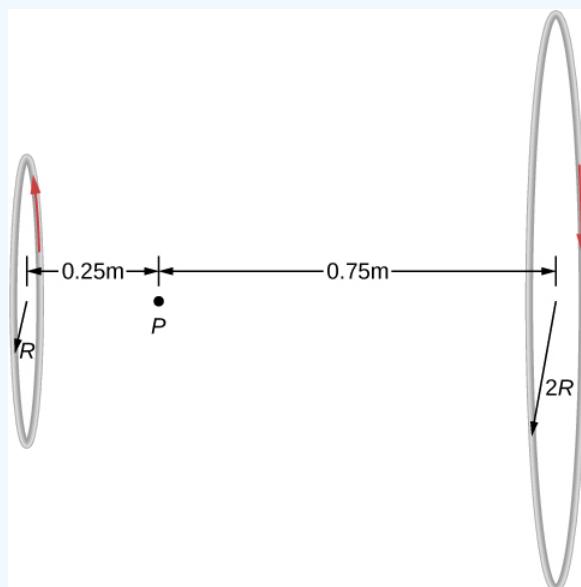


Figure 18.3.3: Two loops of different radii have the same current but flowing in opposite directions. The magnetic field at point **P** is measured to be zero.

Strategy

The magnetic field at point **P** has been determined in Equation 18.3.3. Since the currents are flowing in opposite directions, the net magnetic field is the difference between the two fields generated by the coils. Using the given quantities in the problem, the net magnetic field is then calculated.

Solution

Solving for the net magnetic field using Equation 18.3.3 and the given quantities in the problem yields

$$B = \frac{\mu_0 I R_1^2}{2(y_1^2 + R_1^2)^{3/2}} - \frac{\mu_0 I R_2^2}{2(y_2^2 + R_2^2)^{3/2}}$$

$$B = \frac{(4\pi \times 10^{-7} T \cdot m/A)(0.010 A)(0.5 m)^2}{2((0.25 m)^2 + (0.5 m)^2)^{3/2}} - \frac{(4\pi \times 10^{-7} T \cdot m/A)(0.010 A)(1.0 m)^2}{2((0.75 m)^2 + (1.0 m)^2)^{3/2}}$$

$B = 5.77 \times 10^{-9} T$ to the right.

Significance

Helmholtz coils typically have loops with equal radii with current flowing in the same direction to have a strong uniform field at the midpoint between the loops. A similar application of the magnetic field distribution created by Helmholtz coils is found in a magnetic bottle that can temporarily trap charged particles. See [Magnetic Forces and Fields](#) for a discussion on this.

Exercise 18.3.1

Using Example 18.3.1, at what distance would you have to move the first coil to have zero measurable magnetic field at point **P**?

Solution

0.608 meters

- **12.5: Magnetic Field of a Current Loop** by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.4: Magnetic Field using Ampère's Law

Learning Objectives

By the end of this section, you will be able to:

- Explain how Ampère's law relates the magnetic field produced by a current to the value of the current
- Calculate the magnetic field from a long straight wire, either thin or thick, by Ampère's law

A fundamental property of a static magnetic field is that, unlike an electrostatic field, it is not conservative. A conservative field is one that does the same amount of work on a particle moving between two different points regardless of the path chosen. Magnetic fields do not have such a property. Instead, there is a relationship between the magnetic field and its source, electric current. It is expressed in terms of the line integral of \vec{B} and is known as **Ampère's law**. This law can also be derived directly from the Biot-Savart law. We now consider that derivation for the special case of an infinite, straight wire.

Figure 18.4.1 shows an arbitrary plane perpendicular to an infinite, straight wire whose current I is directed out of the page. The magnetic field lines are circles centered on the wire. To begin, let's consider $\oint \vec{B} \cdot d\vec{l}$ over the closed paths **M** and **N**. Notice that one path (**M**) encloses the wire, whereas the other (**N**) does not. Since the field lines are circular, $\vec{B} \cdot d\vec{l}$ is the product of \mathbf{B} and the projection of $d\mathbf{l}$ onto the circle passing through $d\vec{l}$. If the radius of this particular circle is r , the projection is $r d\theta$, and

$$\vec{B} \cdot d\vec{l} = Br d\theta.$$

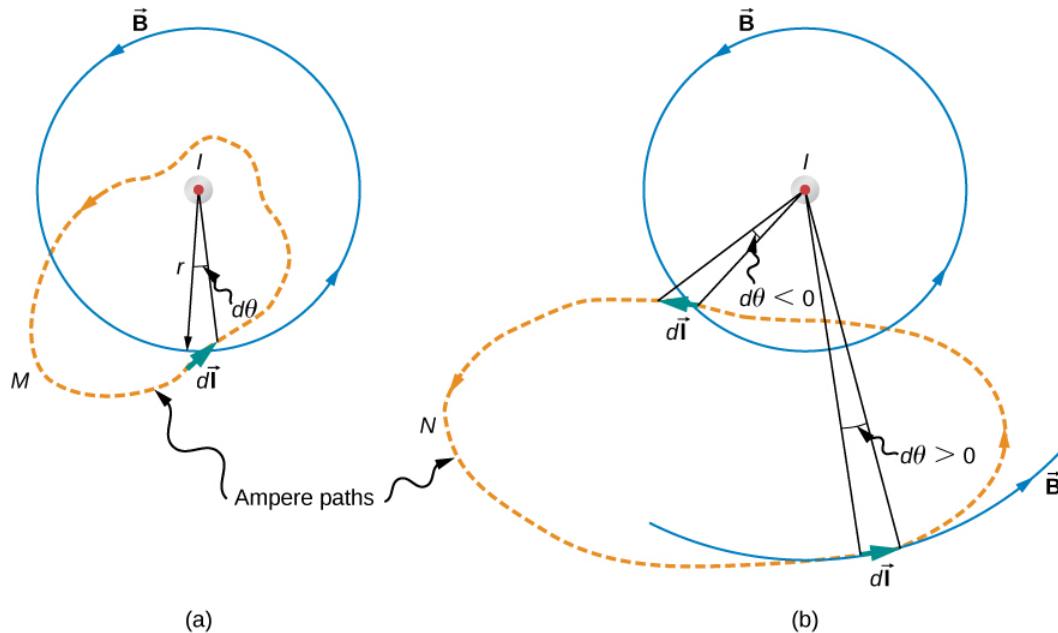


Figure 18.4.1: The current I of a long, straight wire is directed out of the page. The integral $\oint d\theta$ equals 2π and 0, respectively, for paths **M** and **N**.

With \vec{B} given by Equation 12.4.1,

$$\oint \vec{B} \cdot d\vec{l} = \oint \left(\frac{\mu_0 I}{2\pi r} \right) r d\theta = \frac{\mu_0 I}{2\pi} \oint d\theta.$$

For path **M**, which circulates around the wire, $\oint_M d\theta = 2\pi$ and

$$\oint_M \vec{B} \cdot d\vec{l} = \mu_0 I.$$

Path **N**, on the other hand, circulates through both positive (counterclockwise) and negative (clockwise) $d\theta$ (see Figure 18.4.1), and since it is closed, $\oint_N d\theta = 0$. Thus for path **N**,

$$\oint_N \vec{B} \cdot d\vec{l} = 0.$$

The extension of this result to the general case is Ampère's law.

Ampere's Law

Over an arbitrary closed path,

$$\oint \vec{B} \cdot d\vec{l} = \mu_0 I$$

where I is the total current passing through any open surface S whose perimeter is the path of integration. Only currents inside the path of integration need be considered.

To determine whether a specific current I is positive or negative, curl the fingers of your right hand in the direction of the path of integration, as shown in Figure 18.4.1. If I passes through S in the same direction as your extended thumb, I is positive; if I passes through S in the direction opposite to your extended thumb, it is negative.

Problem-Solving Strategy: Ampère's Law

To calculate the magnetic field created from current in wire(s), use the following steps:

1. Identify the symmetry of the current in the wire(s). If there is no symmetry, use the Biot-Savart law to determine the magnetic field.
2. Determine the direction of the magnetic field created by the wire(s) by right-hand rule 2.
3. Choose a path loop where the magnetic field is either constant or zero.
4. Calculate the current inside the loop.
5. Calculate the line integral $\oint \vec{B} \cdot d\vec{l}$ around the closed loop.
6. Equate $\oint \vec{B} \cdot d\vec{l}$ with $\mu_0 I_{enc}$ with $\mu_0 I_{enc}$ and solve for \vec{B} .

Using Ampère's Law to Calculate the Magnetic Field Due to a Wire

Use Ampère's law to calculate the magnetic field due to a steady current I in an infinitely long, thin, straight wire as shown in Figure 18.4.2.

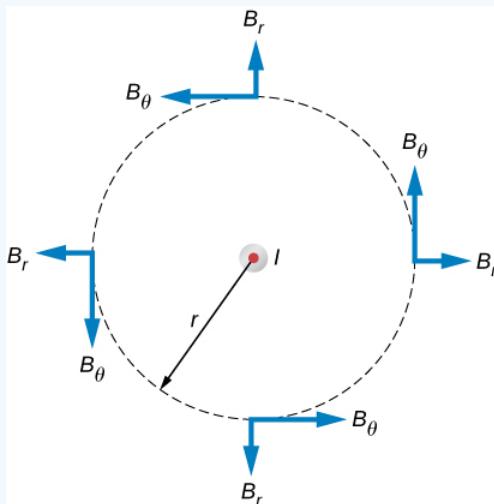


Figure 18.4.2: The possible components of the magnetic field \vec{B} due to a current I , which is directed out of the page. The radial component is zero because the angle between the magnetic field and the path is at a right angle.

Strategy

Consider an arbitrary plane perpendicular to the wire, with the current directed out of the page. The possible magnetic field components in this plane, B_r and B_θ are shown at arbitrary points on a circle of radius r centered on the wire. Since the field is cylindrically symmetric, neither B_r nor B_θ varies with the position on this circle. Also from symmetry, the radial lines, if they exist, must be directed either all inward or all outward from the wire. This means, however, that there must be a net magnetic flux across an arbitrary cylinder concentric with the wire. The radial component of the magnetic field must be zero because $\vec{B} \cdot d\vec{l} = 0$. Therefore, we can apply Ampère's law to the circular path as shown.

Solution

Over this path \vec{B} is constant and parallel to $d\vec{l}$, so

$$\oint \vec{B} \cdot d\vec{l} = B_\theta \oint dl = B_\theta(2\pi r).$$

Thus Ampère's law reduces to

$$B_\theta(2\pi r) = \mu_0 I.$$

Finally, since B_θ is the only component of \vec{B} , we can drop the subscript and write

$$B = \frac{\mu_0 I}{2\pi r}.$$

This agrees with the Biot-Savart calculation above.

Significance

Ampère's law works well if you have a path to integrate over which $\vec{B} \cdot d\vec{l}$ has results that are easy to simplify. For the infinite wire, this works easily with a path that is circular around the wire so that the magnetic field factors out of the integration. If the path dependence looks complicated, you can always go back to the Biot-Savart law and use that to find the magnetic field.

✓ Example 18.4.2: Calculating the Magnetic Field of a Thick Wire with Ampère's Law

The radius of the long, straight wire of Figure 18.4.3 is a , and the wire carries a current I_0 that is distributed uniformly over its cross-section. Find the magnetic field both inside and outside the wire.

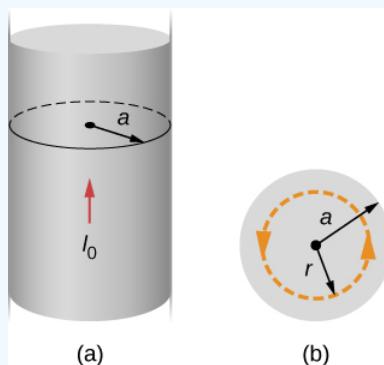


Figure 18.4.3: (a) A model of a current-carrying wire of radius a and current I_0 . (b) A cross-section of the same wire showing the radius a and the Ampère's loop of radius r .

Strategy

This problem has the same geometry as Example 18.4.1, but the enclosed current changes as we move the integration path from outside the wire to inside the wire, where it doesn't capture the entire current enclosed (see Figure 18.4.3).

Solution

For any circular path of radius r that is centered on the wire,

$$\oint \vec{B} \cdot d\vec{l} = \oint B dl = B \oint dl = B(2\pi r).$$

From Ampère's law, this equals the total current passing through any surface bounded by the path of integration.

Consider first a circular path that is inside the wire ($r \leq a$) such as that shown in part (a) of Figure 18.4.3. We need the current I passing through the area enclosed by the path. It's equal to the current density J times the area enclosed. Since the current is uniform, the current density inside the path equals the current density in the whole wire, which is $I_0/\pi a^2$. Therefore the current I passing through the area enclosed by the path is

$$I = \frac{\pi r^2}{\pi a^2} I_0 = \frac{r^2}{a^2} I_0.$$

We can consider this ratio because the current density J is constant over the area of the wire. Therefore, the current density of a part of the wire is equal to the current density in the whole area. Using Ampère's law, we obtain

$$B(2\pi r) = \mu_0 \left(\frac{r^2}{a^2} \right) I_0,$$

and the magnetic field inside the wire is

$$B = \frac{\mu_0 I_0}{2\pi} \frac{r}{a^2} (r \leq a).$$

Outside the wire, the situation is identical to that of the infinite thin wire of the previous example; that is,

$$B = \frac{\mu_0 I_0}{2\pi r} (r \geq a).$$

The variation of \mathbf{B} with r is shown in Figure 18.4.4.

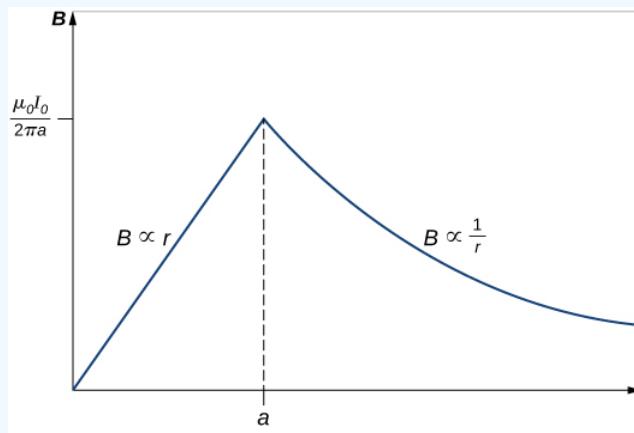


Figure 18.4.4: Variation of the magnetic field produced by a current I_0 in a long, straight wire of radius a .

Significance

The results show that as the radial distance increases inside the thick wire, the magnetic field increases from zero to a familiar value of the magnetic field of a thin wire. Outside the wire, the field drops off regardless of whether it was a thick or thin wire.

This result is similar to how Gauss's law for electrical charges behaves inside a uniform charge distribution, except that Gauss's law for electrical charges has a uniform volume distribution of charge, whereas Ampère's law here has a uniform area of current distribution. Also, the drop-off outside the thick wire is similar to how an electric field drops off outside of a linear charge distribution, since the two cases have the same geometry and neither case depends on the configuration of charges or currents once the loop is outside the distribution.

✓ Using Ampère's Law with Arbitrary Paths

Use Ampère's law to evaluate $\oint \vec{B} \cdot d\vec{l}$ for the current configurations and paths in Figure 18.4.5.

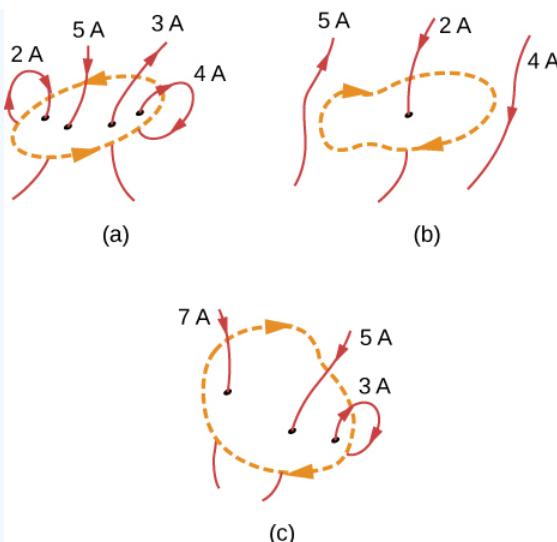


Figure 18.4.5: Current configurations and paths for Example 18.4.3.

Strategy

Ampère's law states that $\oint \vec{B} \cdot d\vec{l} = \mu_0 I$ where I is the total current passing through the enclosed loop. The quickest way to evaluate the integral is to calculate $\mu_0 I$ by finding the net current through the loop. Positive currents flow with your right-hand thumb if your fingers wrap around in the direction of the loop. This will tell us the sign of the answer.

Solution

- (a) The current going downward through the loop equals the current going out of the loop, so the net current is zero. Thus, $\oint \vec{B} \cdot d\vec{l} = 0$.
- (b) The only current to consider in this problem is 2A because it is the only current inside the loop. The right-hand rule shows us the current going downward through the loop is in the positive direction. Therefore, the answer is $\oint \vec{B} \cdot d\vec{l} = \mu_0(2 \text{ A}) = 2.51 \times 10^{-6} \text{T} \cdot \text{m}$.
- (c) The right-hand rule shows us the current going downward through the loop is in the positive direction. There are $7\text{A} + 5\text{A} = 12\text{A}$ of current going downward and -3 A going upward. Therefore, the total current is 9 A and $\oint \vec{B} \cdot d\vec{l} = \mu_0(9 \text{ A}) = 5.65 \times 10^{-6} \text{T} \cdot \text{m}$.

Significance

If the currents all wrapped around so that the same current went into the loop and out of the loop, the net current would be zero and no magnetic field would be present. This is why wires are very close to each other in an electrical cord. The currents flowing toward a device and away from a device in a wire equal zero total current flow through an Ampère loop around these wires. Therefore, no stray magnetic fields can be present from cords carrying current.

Exercise 18.4.1

Consider using Ampère's law to calculate the magnetic fields of a finite straight wire and of a circular loop of wire. Why is it not useful for these calculations?

Answer

In these cases the integrals around the Ampèrean loop are very difficult because there is no symmetry, so this method would not be useful.

This page titled [18.4: Magnetic Field using Ampère's Law](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.6: Ampère's Law](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.5: Magnetic Field of Solenoids and Toroids

Learning Objectives

By the end of this section, you will be able to:

- Establish a relationship for how the magnetic field of a solenoid varies with distance and current by using both the Biot-Savart law and Ampère's law
- Establish a relationship for how the magnetic field of a toroid varies with distance and current by using Ampère's law

Two of the most common and useful electromagnetic devices are called solenoids and toroids. In one form or another, they are part of numerous instruments, both large and small. In this section, we examine the magnetic field typical of these devices.

Solenoids

A long wire wound in the form of a helical coil is known as a **solenoid**. Solenoids are commonly used in experimental research requiring magnetic fields. A solenoid is generally easy to wind, and near its center, its magnetic field is quite uniform and directly proportional to the current in the wire.

Figure 18.5.1 shows a solenoid consisting of **N** turns of wire tightly wound over a length **L**. A current **I** is flowing along the wire of the solenoid. The number of turns per unit length is **N/L**; therefore, the number of turns in an infinitesimal length **dy** are $(N/L)dy$ turns. This produces a current

$$dI = \frac{NI}{L} dy. \quad (18.5.1)$$

We first calculate the magnetic field at the point **P** of Figure 18.5.1. This point is on the central axis of the solenoid. We are basically cutting the solenoid into thin slices that are **dy** thick and treating each as a current loop. Thus, **dI** is the current through each slice. The magnetic field $d\vec{B}$ due to the current **dI** in **dy** can be found with the help of Equation 12.5.3 and Equation 18.5.1:

$$d\vec{B} = \frac{\mu_0 R^2 dI}{2(y^2 + R^2)^{3/2}} \hat{j} = \left(\frac{\mu_0 I R^2 N}{2L} \hat{j} \right) \frac{dy}{(y^2 + R^2)^{3/2}} \quad (18.5.2)$$

where we used Equation 18.5.1 to replace **dI**. The resultant field at **P** is found by integrating $d\vec{B}$ along the entire length of the solenoid. It's easiest to evaluate this integral by changing the independent variable from **y** to **θ**. From inspection of Figure 18.5.1, we have:

$$\sin \theta = \frac{y}{\sqrt{y^2 + R^2}}. \quad (18.5.3)$$

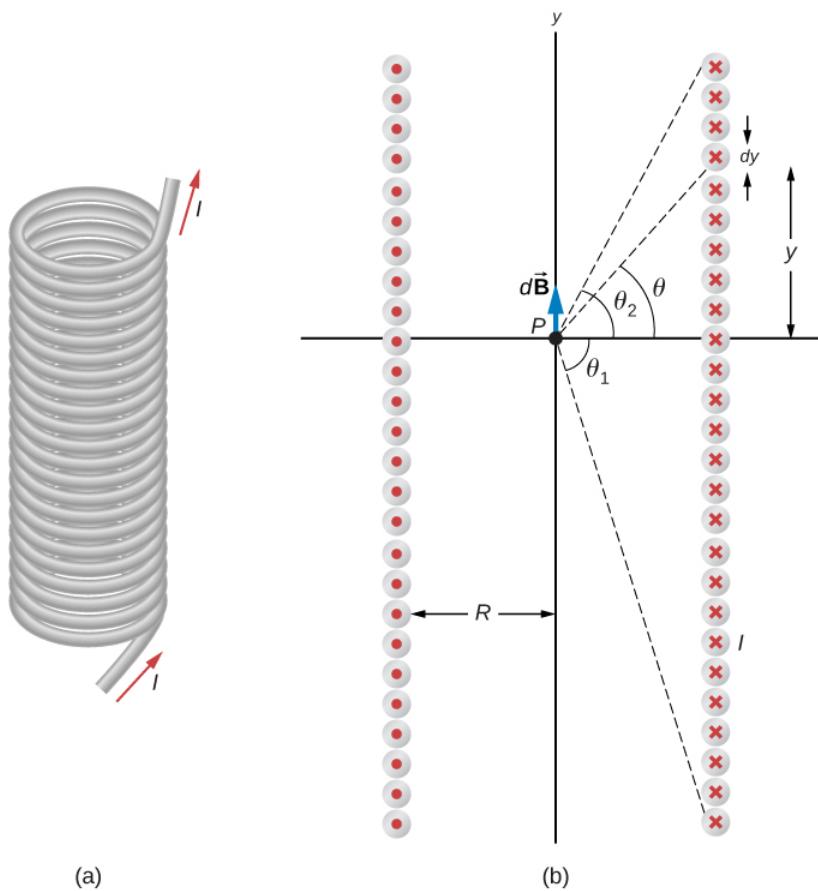


Figure 18.5.1: (a) A solenoid is a long wire wound in the shape of a helix. (b) The magnetic field at the point **P** on the axis of the solenoid is the net field due to all of the current loops.

Taking the differential of both sides of this equation, we obtain

$$\begin{aligned} \cos \theta d\theta &= \left[-\frac{y^2}{(y^2 + R^2)^{3/2}} + \frac{1}{\sqrt{y^2 + R^2}} \right] dy \\ &= \frac{R^2 dy}{(y^2 + R^2)^{3/2}}. \end{aligned}$$

When this is substituted into the equation for $d\vec{B}$, we have

$$\vec{B} = \frac{\mu_0 I N}{2L} \hat{j} \int_{\theta_1}^{\theta_2} \cos \theta d\theta = \frac{\mu_0 I N}{2L} (\sin \theta_2 - \sin \theta_1) \hat{j}, \quad (18.5.4)$$

which is the magnetic field along the central axis of a finite solenoid.

Of special interest is the infinitely long solenoid, for which $L \rightarrow \infty$. From a practical point of view, the infinite solenoid is one whose length is much larger than its radius ($L \gg R$). In this case, $\theta_1 = -\frac{\pi}{2}$ and $\theta_2 = \frac{\pi}{2}$. Then from Equation 18.5.4, the magnetic field along the central axis of an infinite solenoid is

$$\vec{B} = \frac{\mu_0 I N}{2L} \hat{j} [\sin(\pi/2) - \sin(-\pi/2)] = \frac{\mu_0 I N}{L} \hat{j}$$

or

$$\vec{B} = \mu_0 n I \hat{j}, \quad (18.5.5)$$

where n is the number of turns per unit length. You can find the direction of \vec{B} with a right-hand rule: Curl your fingers in the direction of the current, and your thumb points along the magnetic field in the interior of the solenoid.

We now use these properties, along with Ampère's law, to calculate the magnitude of the magnetic field at any location inside the infinite solenoid. Consider the closed path of Figure 18.5.2. Along segment 1, \vec{B} is uniform and parallel to the path. Along segments 2 and 4, \vec{B} is perpendicular to part of the path and vanishes over the rest of it. Therefore, segments 2 and 4 do not contribute to the line integral in Ampère's law. Along segment 3, $\vec{B} = 0$ because the magnetic field is zero outside the solenoid. If you consider an Ampère's law loop outside of the solenoid, the current flows in opposite directions on different segments of wire. Therefore, there is no enclosed current and no magnetic field according to Ampère's law. Thus, there is no contribution to the line integral from segment 3. As a result, we find

$$\oint \vec{B} \cdot d\vec{l} = \int_1 \vec{B} \cdot d\vec{l} = Bl. \quad (18.5.6)$$

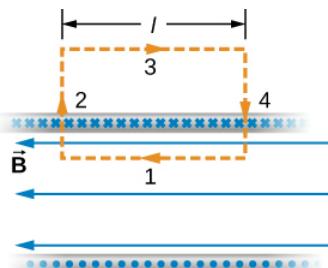


Figure 18.5.2: The path of integration used in Ampère's law to evaluate the magnetic field of an infinite solenoid.

The solenoid has n turns per unit length, so the current that passes through the surface enclosed by the path is nIl . Therefore, from Ampère's law,

$$Bl = \mu_0 nIl$$

and

 Note

$$B = \mu_0 nI \quad (18.5.7)$$

within the solenoid. This agrees with what we found earlier for \mathbf{B} on the central axis of the solenoid. Here, however, the location of segment 1 is arbitrary, so we have found that this equation gives the magnetic field everywhere inside the infinite solenoid.

Outside the solenoid, one can draw an Ampère's law loop around the entire solenoid. This would enclose current flowing in both directions. Therefore, the net current inside the loop is zero. According to Ampère's law, if the net current is zero, the magnetic field must be zero. Therefore, for locations outside of the solenoid's radius, the magnetic field is zero.

When a patient undergoes a **magnetic resonance imaging** (MRI) scan, the person lies down on a table that is moved into the center of a large solenoid that can generate very large magnetic fields. The solenoid is capable of these high fields from high currents flowing through superconducting wires. The large magnetic field is used to change the spin of protons in the patient's body. The time it takes for the spins to align or relax (return to original orientation) is a signature of different tissues that can be analyzed to see if the structures of the tissues is normal (Figure 18.5.3).



Figure 18.5.3: In an MRI machine, a large magnetic field is generated by the cylindrical solenoid surrounding the patient. (credit: Liz West)

✓ Example 18.5.1: Magnetic Field Inside a Solenoid

A solenoid has 300 turns wound around a cylinder of diameter 1.20 cm and length 14.0 cm. If the current through the coils is 0.410 A, what is the magnitude of the magnetic field inside and near the middle of the solenoid?

Strategy

We are given the number of turns and the length of the solenoid so we can find the number of turns per unit length. Therefore, the magnetic field inside and near the middle of the solenoid is given by Equation 18.5.7. Outside the solenoid, the magnetic field is zero.

Solution

The number of turns per unit length is

$$n = \frac{300 \text{ turns}}{0.140 \text{ m}} = 2.14 \times 10^3 \text{ turns/m.}$$

The magnetic field produced inside the solenoid is

$$\begin{aligned} B &= \mu_0 n I = (4\pi \times 10^{-7} \text{ T} \cdot \text{m/A})(2.14 \times 10^3 \text{ turns/m})(0.410 \text{ A}) \\ B &= 1.10 \times 10^{-3} \text{ T}. \end{aligned}$$

Significance

This solution is valid only if the length of the solenoid is reasonably large compared with its diameter. This example is a case where this is valid.

? Exercise 18.5.1

What is the ratio of the magnetic field produced from using a finite formula over the infinite approximation for an angle θ of (a) 85° ? (b) 89° ? The solenoid has 1000 turns in 50 cm with a current of 1.0 A flowing through the coils

Solution

- a. 1.00382; b. 1.00015

Toroids

A toroid is a donut-shaped coil closely wound with one continuous wire, as illustrated in part (a) of Figure 18.5.4. If the toroid has N windings and the current in the wire is I , what is the magnetic field both inside and outside the toroid?

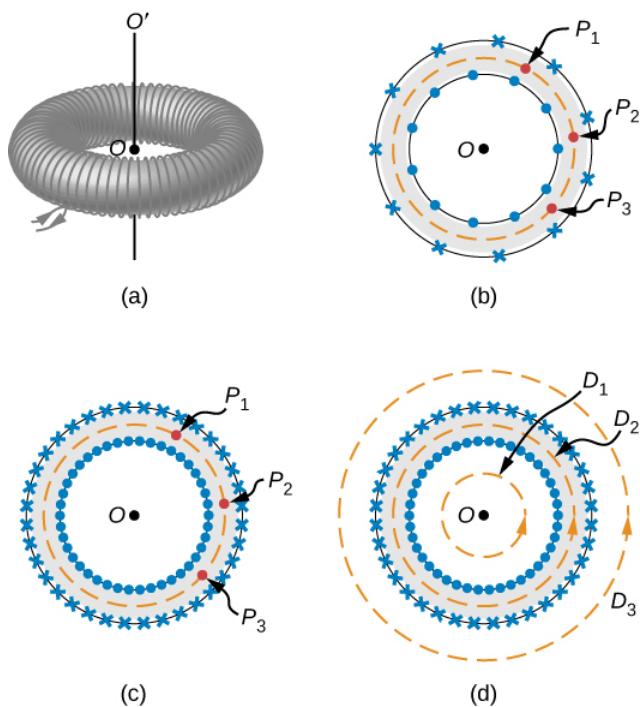


Figure 18.5.4: (a) A toroid is a coil wound into a donut-shaped object. (b) A loosely wound toroid does not have cylindrical symmetry. (c) In a tightly wound toroid, cylindrical symmetry is a very good approximation. (d) Several paths of integration for Ampère's law.

We begin by assuming cylindrical symmetry around the axis OO' . Actually, this assumption is not precisely correct, for as part (b) of Figure 18.5.4 shows, the view of the toroidal coil varies from point to point (for example, P_1 , P_2 and P_3) on a circular path centered around OO' . However, if the toroid is tightly wound, all points on the circle become essentially equivalent [part (c) of Figure 18.5.4], and cylindrical symmetry is an accurate approximation.

With this symmetry, the magnetic field must be tangent to and constant in magnitude along any circular path centered on OO' . This allows us to write for each of the paths D_1 , D_2 and D_3 shown in part (d) of Figure 18.5.4,

$$\oint \vec{B} \cdot d\vec{l} = B(2\pi r). \quad (18.5.8)$$

Ampère's law relates this integral to the net current passing through any surface bounded by the path of integration. For a path that is external to the toroid, either no current passes through the enclosing surface (path D_1), or the current passing through the surface in one direction is exactly balanced by the current passing through it in the opposite direction (path D_3). In either case, there is no net current passing through the surface, so

$$\oint B(2\pi r) = 0$$

and

$$B = 0 \text{ (outside the toroid)}. \quad (18.5.9)$$

The turns of a toroid form a helix, rather than circular loops. As a result, there is a small field external to the coil; however, the derivation above holds if the coils were circular.

For a circular path within the toroid (path D_2), the current in the wire cuts the surface N times, resulting in a net current NI through the surface. We now find with Ampère's law,

$$B(2\pi r) = \mu_0 NI$$

and

✓ Note

$$B = \frac{\mu_0 NI}{2\pi r} \text{ (within the toroid).} \quad (18.5.10)$$

The magnetic field is directed in the counterclockwise direction for the windings shown. When the current in the coils is reversed, the direction of the magnetic field also reverses.

The magnetic field inside a toroid is not uniform, as it varies inversely with the distance r from the axis $\mathbf{OO'}$. However, if the central radius \mathbf{R} (the radius midway between the inner and outer radii of the toroid) is much larger than the cross-sectional diameter of the coils \mathbf{r} , the variation is fairly small, and the magnitude of the magnetic field may be calculated by Equation 18.5.10 where $r = R$.

This page titled [18.5: Magnetic Field of Solenoids and Toroids](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.7: Solenoids and Toroids](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.6: Magnetic Force between Two Parallel Currents

Learning Objectives

By the end of this section, you will be able to:

- Explain how parallel wires carrying currents can attract or repel each other
- Define the ampere and describe how it is related to current-carrying wires
- Calculate the force of attraction or repulsion between two current-carrying wires

You might expect that two current-carrying wires generate significant forces between them, since ordinary currents produce magnetic fields and these fields exert significant forces on ordinary currents. But you might not expect that the force between wires is used to define the ampere. It might also surprise you to learn that this force has something to do with why large circuit breakers burn up when they attempt to interrupt large currents.

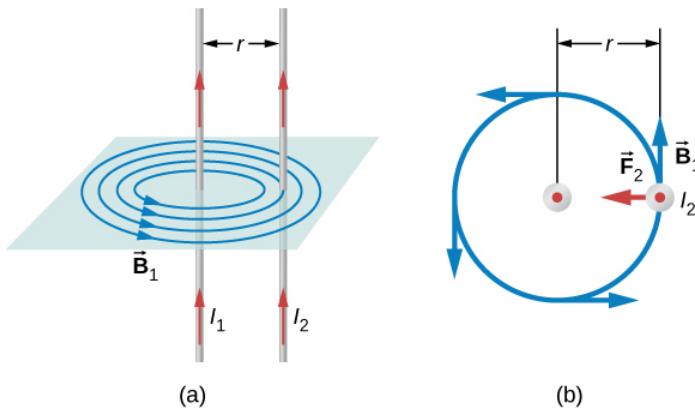


Figure 18.6.1: (a) The magnetic field produced by a long straight conductor is perpendicular to a parallel conductor, as indicated by right-hand rule (RHR)-2. (b) A view from above of the two wires shown in (a), with one magnetic field line shown for wire 1. RHR-1 shows that the force between the parallel conductors is attractive when the currents are in the same direction. A similar analysis shows that the force is repulsive between currents in opposite directions.

The force between two long, straight, and parallel conductors separated by a distance \mathbf{r} can be found by applying what we have developed in the preceding sections. Figure 18.6.1 shows the wires, their currents, the field created by one wire, and the consequent force the other wire experiences from the created field. Let us consider the field produced by wire 1 and the force it exerts on wire 2 (call the force \mathbf{F}_2). The field due to I_1 at a distance \mathbf{r} is

$$B_1 = \frac{\mu_0 I_1}{2\pi r}$$

This field is uniform from the wire 1 and perpendicular to it, so the force \mathbf{F}_2 it exerts on a length \mathbf{l} of wire 2 is given by $\mathbf{F} = \mathbf{IB} \sin \theta$ with $\sin \theta = 1$:

$$\mathbf{F}_2 = I_2 l B_1. \quad (18.6.1)$$

The forces on the wires are equal in magnitude, so we just write \mathbf{F} for the magnitude of \mathbf{F}_2 (Note that $\vec{F}_1 = -\vec{F}_2$.) Since the wires are very long, it is convenient to think in terms of \mathbf{F}/\mathbf{l} , the force per unit length. Substituting the expression for B_1 into Equation 18.6.1 and rearranging terms gives

Note

$$\frac{\mathbf{F}}{\mathbf{l}} = \frac{\mu_0 I_1 I_2}{2\pi r}. \quad (18.6.2)$$

The ratio \mathbf{F}/\mathbf{l} is the force per unit length between two parallel currents I_1 and I_2 separated by a distance r . The force is attractive if the currents are in the same direction and repulsive if they are in opposite directions.

This force is responsible for the **pinch effect** in electric arcs and other plasmas. The force exists whether the currents are in wires or not. It is only apparent if the overall charge density is zero; otherwise, the Coulomb repulsion overwhelms the magnetic attraction. In an electric arc, where charges are moving parallel to one another, an attractive force squeezes currents into a smaller tube. In large circuit breakers, such as those used in neighborhood power distribution systems, the pinch effect can concentrate an arc between plates of a switch trying to break a large current, burn holes, and even ignite the equipment. Another example of the pinch effect is found in the solar plasma, where jets of ionized material, such as solar flares, are shaped by magnetic forces.

The definition of the **ampere** is based on the force between current-carrying wires. Note that for long, parallel wires separated by 1 meter with each carrying 1 ampere, the force per meter is

$$\frac{F}{l} = \frac{(4\pi \times 10^{-7} T \cdot m/A)(1 A)^2}{(2\pi)(1 m)} = 2 \times 10^{-7} N/m.$$

Since μ_0 is exactly $4\pi \times 10^{-7} T \cdot m/A$ by definition, and because $1 T = 1 N/(A \cdot m)$, the force per meter is exactly $2 \times 10^{-7} N/m$. This is the basis of the definition of the ampere.

Infinite-length wires are impractical, so in practice, a current balance is constructed with coils of wire separated by a few centimeters. Force is measured to determine current. This also provides us with a method for measuring the coulomb. We measure the charge that flows for a current of one ampere in one second. That is, $1 C = 1 A \cdot s$. For both the ampere and the coulomb, the method of measuring force between conductors is the most accurate in practice.

✓ Example 18.6.1: Calculating Forces on Wires

Two wires, both carrying current out of the page, have a current of magnitude 5.0 mA. The first wire is located at (0.0 cm, 3.0 cm) while the other wire is located at (4.0 cm, 0.0 cm) as shown in Figure 18.6.2. What is the magnetic force per unit length of the first wire on the second and the second wire on the first?

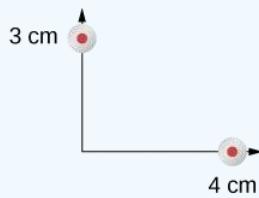


Figure 18.6.2: Two current-carrying wires at given locations with currents out of the page.

Strategy

Each wire produces a magnetic field felt by the other wire. The distance along the hypotenuse of the triangle between the wires is the radial distance used in the calculation to determine the force per unit length. Since both wires have currents flowing in the same direction, the direction of the force is toward each other.

Solution

The distance between the wires results from finding the hypotenuse of a triangle:

$$r = \sqrt{(3.0 \text{ cm})^2 + (4.0 \text{ cm})^2} = 5.0 \text{ cm}.$$

The force per unit length can then be calculated using the known currents in the wires:

$$\frac{F}{l} = \frac{(4\pi \times 10^{-7} T \cdot m/A)(5 \times 10^{-3} A)^2}{(2\pi)(5 \times 10^{-2} m)} = 1 \times 10^{-10} N/m.$$

The force from the first wire pulls the second wire. The angle between the radius and the x-axis is

$$\theta = \tan^{-1} \left(\frac{3 \text{ cm}}{4 \text{ cm}} \right) = 36.9^\circ.$$

The unit vector for this is calculated by

$$\cos(36.9^\circ)\hat{i} - \sin(36.9^\circ)\hat{j} = 0.8\hat{i} - 0.6\hat{j}.$$

Therefore, the force per unit length from wire one on wire 2 is

$$\frac{\vec{F}}{l} = (1 \times 10^{-10} \text{ N/m}) \times (0.8\hat{i} - 0.6\hat{j}) = (8 \times 10^{-11}\hat{i} - 6 \times 10^{-11}\hat{j}) \text{ N/m.}$$

The force per unit length from wire 2 on wire 1 is the negative of the previous answer:

$$\frac{\vec{F}}{l} = (-8 \times 10^{-11}\hat{i} + 6 \times 10^{-11}\hat{j}) \text{ N/m.}$$

Significance

These wires produced magnetic fields of equal magnitude but opposite directions at each other's locations. Whether the fields are identical or not, the forces that the wires exert on each other are always equal in magnitude and opposite in direction (Newton's third law).

Exercise 18.6.1

Two wires, both carrying current out of the page, have a current of magnitude 2.0 mA and 3.0 mA, respectively. The first wire is located at (0.0 cm, 5.0 cm) while the other wire is located at (12.0 cm, 0.0 cm). What is the magnitude of the magnetic force per unit length of the first wire on the second and the second wire on the first?

Answer

Both have a force per unit length of $9.23 \times 10^{-12} \text{ N/m}$

This page titled [18.6: Magnetic Force between Two Parallel Currents](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.4: Magnetic Force between Two Parallel Currents](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

18.7: (edit) Magnetic Force and Torque on a Current Loop - Motors and Meters

Learning Objectives

By the end of this section, you will be able to:

- Evaluate the net force on a current loop in an external magnetic field
- Evaluate the net torque on a current loop in an external magnetic field
- Define the magnetic dipole moment of a current loop

Motors are the most common application of magnetic force on current-carrying wires. Motors contain loops of wire in a magnetic field. When current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft. Electrical energy is converted into mechanical work in the process. Once the loop's surface area is aligned with the magnetic field, the direction of current is reversed, so there is a continual torque on the loop (Figure 18.7.1). This reversal of the current is done with commutators and brushes. The commutator is set to reverse the current flow at set points to keep continual motion in the motor. A basic commutator has three contact areas to avoid dead spots where the loop would have zero instantaneous torque at that point. The brushes press against the commutator, creating electrical contact between parts of the commutator during the spinning motion.

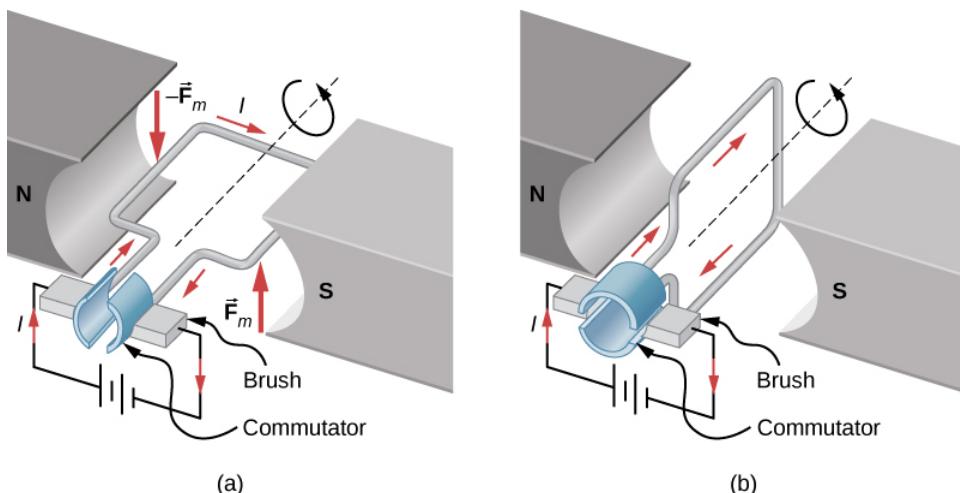


Figure 18.7.1: A simplified version of a dc electric motor. (a) The rectangular wire loop is placed in a magnetic field. The forces on the wires closest to the magnetic poles (N and S) are opposite in direction as determined by the right-hand rule-1. Therefore, the loop has a net torque and rotates to the position shown in (b). (b) The brushes now touch the commutator segments so that no current flows through the loop. No torque acts on the loop, but the loop continues to spin from the initial velocity given to it in part (a). By the time the loop flips over, current flows through the wires again but now in the opposite direction, and the process repeats as in part (a). This causes continual rotation of the loop.

In a uniform magnetic field, a current-carrying loop of wire, such as a loop in a motor, experiences both forces and torques on the loop. Figure 18.7.1 shows a rectangular loop of wire that carries a current I and has sides of lengths a and b . The loop is in a uniform magnetic field: $\vec{B} = B\hat{j}$. The magnetic force on a straight current-carrying wire of length l is given by $I\vec{l} \times \vec{B}$. To find the net force on the loop, we have to apply this equation to each of the four sides. The force on side 1 is

$$\vec{F}_1 = IaB \sin(90^\circ - \theta)\hat{i} = IaB \cos \theta \hat{i}$$

where the direction has been determined with the RHR-1. The current in side 3 flows in the opposite direction to that of side 1, so

$$\vec{F}_3 = -IaB \sin(90^\circ + \theta)\hat{i} = -IaB \cos \theta \hat{i}$$

The currents in sides 2 and 4 are perpendicular to \vec{B} and the forces on these sides are

$$\vec{F}_2 = IbB\hat{k}$$

$$\vec{F}_4 = -IbB\hat{k}.$$

We can now find the net force on the loop:

$$\sum \vec{F}_{net} = \vec{F}_1 + \vec{F}_2 + \vec{F}_3 + \vec{F}_4 = 0.$$

Although this result ($\sum F = 0$) has been obtained for a rectangular loop, it is far more general and holds for current-carrying loops of arbitrary shapes; that is, there is no net force on a current loop in a uniform magnetic field.

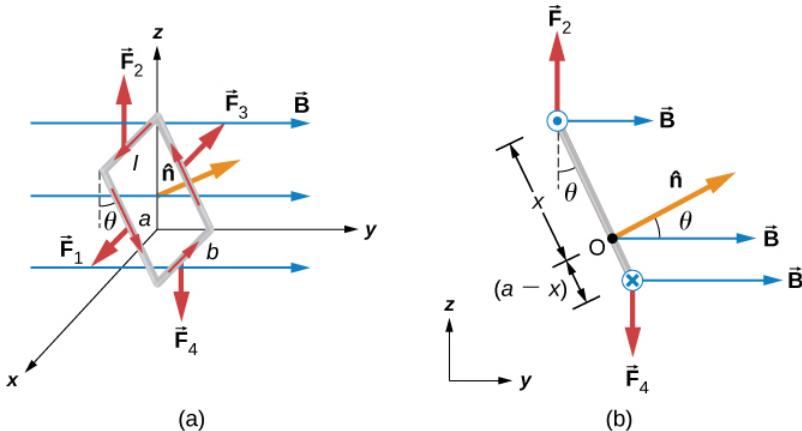


Figure 18.7.2: (a) A rectangular current loop in a uniform magnetic field is subjected to a net torque but not a net force. (b) A side view of the coil.

To find the net torque on the current loop shown in Figure 18.7.2a, we first consider \vec{F}_1 and \vec{F}_3 . Since they have the same line of action and are equal and opposite, the sum of their torques about any axis is zero (see [Fixed-Axis Rotation](#)). Thus, if there is any torque on the loop, it must be furnished by \vec{F}_2 and \vec{F}_4 . Let's calculate the torques around the axis that passes through point **O** of Figure 18.7.2b (a side view of the coil) and is perpendicular to the plane of the page. The point **O** is a distance x from side 2 and a distance $(a - x)$ from side 4 of the loop. The moment arms of \vec{F}_2 and \vec{F}_4 are $x \sin \theta \hat{i}$ and $(a - x) \sin \theta \hat{i}$, respectively, so the net torque on the loop is

$$\begin{aligned} \sum \vec{\tau} &= \vec{\tau}_1 + \vec{\tau}_2 + \vec{\tau}_3 + \vec{\tau}_4 = F_2 x \sin \theta \hat{i} - F_4 (a - x) \sin \theta \hat{i} \\ &\quad - IbBx \sin \theta \hat{i} - IbB(a - x) \sin \theta \hat{i}. \end{aligned}$$

This simplifies to

$$\vec{\tau} = -IAB \sin \theta \hat{i}$$

where $A = ab$ is the area of the loop.

Notice that this torque is independent of x ; it is therefore independent of where point **O** is located in the plane of the current loop. Consequently, the loop experiences the same torque from the magnetic field about any axis in the plane of the loop and parallel to the x -axis.

A closed-current loop is commonly referred to as a **magnetic dipole** and the term IA is known as its **magnetic dipole moment** μ . Actually, the magnetic dipole moment is a vector that is defined as

$$\vec{\mu} = IA \hat{n}$$

where \hat{n} is a unit vector directed perpendicular to the plane of the loop (see Figure 18.7.2). The direction of \hat{n} is obtained with the RHR-2—if you curl the fingers of your right hand in the direction of current flow in the loop, then your thumb points along \hat{n} . If the loop contains N turns of wire, then its magnetic dipole moment is given by

$$\vec{\mu} = NIA \hat{n}.$$

In terms of the magnetic dipole moment, the torque on a current loop due to a uniform magnetic field can be written simply as

$$\vec{\tau} = \vec{\mu} \times \vec{B}.$$

This equation holds for a current loop in a two-dimensional plane of arbitrary shape.

Using a calculation analogous to that found in [Capacitance](#) for an electric dipole, the potential energy of a magnetic dipole is

$$U = -\vec{\mu} \cdot \vec{B}.$$

✓ Example 18.7.1: Forces and Torques on Current-Carrying Loops

A circular current loop of radius 2.0 cm carries a current of 2.0 mA. (a) What is the magnitude of its magnetic dipole moment? (b) If the dipole is oriented at 30 degrees to a uniform magnetic field of magnitude 0.50 T, what is the magnitude of the torque it experiences and what is its potential energy?

Strategy

The dipole moment is defined by the current times the area of the loop. The area of the loop can be calculated from the area of the circle. The torque on the loop and potential energy are calculated from identifying the magnetic moment, magnetic field, and angle oriented in the field.

Solution

1. The magnetic moment μ is calculated by the current times the area of the loop or πr^2 .

$$\mu = IA = (2.0 \times 10^{-3} A)(\pi(0.02 m)^2) = 2.5 \times 10^{-6} A \cdot m^2$$

2. The torque and potential energy are calculated by identifying the magnetic moment, magnetic field, and the angle between these two vectors. The calculations of these quantities are:

$$\tau = \vec{\mu} \times \vec{B} = \mu B \sin \theta = (2.5 \times 10^{-6} A \cdot m^2)(0.50 T) \sin(30^\circ) = 6.3 \times 10^{-7} N \cdot m$$

$$U = -\vec{\mu} \cdot \vec{B} = -\mu B \cos \theta = -(2.5 \times 10^{-6} A \cdot m^2)(0.50 T) \cos(30^\circ) = -1.1 \times 10^{-6} J.$$

Significance

The concept of magnetic moment at the atomic level is discussed in the next chapter. The concept of aligning the magnetic moment with the magnetic field is the functionality of devices like magnetic motors, whereby switching the external magnetic field results in a constant spinning of the loop as it tries to align with the field to minimize its potential energy.

? Exercise 18.7.1

In what orientation would a magnetic dipole have to be to produce (a) a maximum torque in a magnetic field? (b) A maximum energy of the dipole?

Solution

- a. aligned or anti-aligned; b. perpendicular

This page titled [18.7: \(edit\) Magnetic Force and Torque on a Current Loop - Motors and Meters](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [11.6: Force and Torque on a Current Loop](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

18.8: Magnetic Forces in a Conductor - The Hall Effect

Learning Objectives

By the end of this section, you will be able to:

- Explain a scenario where the magnetic and electric fields are crossed and their forces balance each other as a charged particle moves through a velocity selector
- Compare how charge carriers move in a conductive material and explain how this relates to the Hall effect

In 1879, E.H. Hall devised an experiment that can be used to identify the sign of the predominant charge carriers in a conducting material. From a historical perspective, this experiment was the first to demonstrate that the charge carriers in most metals are negative.

Visit this [website](#) to find more information about the Hall effect.

We investigate the **Hall effect** by studying the motion of the free electrons along a metallic strip of width \mathbf{I} in a constant magnetic field (Figure 18.8.1). The electrons are moving from left to right, so the magnetic force they experience pushes them to the bottom edge of the strip. This leaves an excess of positive charge at the top edge of the strip, resulting in an electric field \mathbf{E} directed from top to bottom. The charge concentration at both edges builds up until the electric force on the electrons in one direction is balanced by the magnetic force on them in the opposite direction. Equilibrium is reached when:

$$e\mathbf{E} = ev_d\mathbf{B} \quad (18.8.1)$$

where e is the magnitude of the electron charge, v_d is the drift speed of the electrons, and \mathbf{E} is the magnitude of the electric field created by the separated charge. Solving this for the drift speed results in

$$v_d = \frac{\mathbf{E}}{B}. \quad (18.8.2)$$

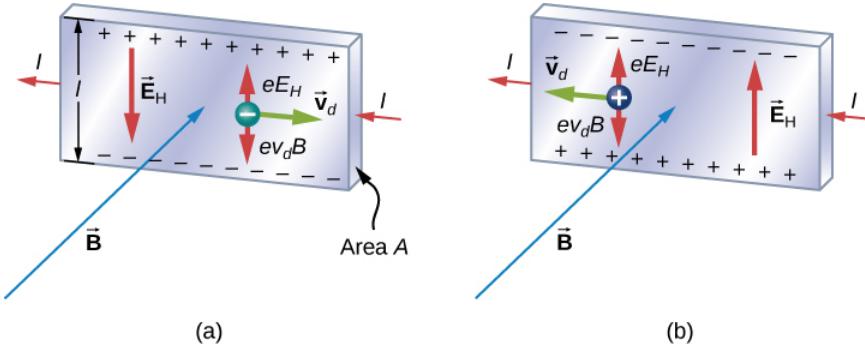


Figure 18.8.1: In the Hall effect, a potential difference between the top and bottom edges of the metal strip is produced when moving charge carriers are deflected by the magnetic field. (a) Hall effect for negative charge carriers; (b) Hall effect for positive charge carriers.

A scenario where the electric and magnetic fields are perpendicular to one another is called a crossed-field situation. If these fields produce equal and opposite forces on a charged particle with the velocity that equates the forces, these particles are able to pass through an apparatus, called a **velocity selector**, undeflected. This velocity is represented in Equation 18.8.3. Any other velocity of a charged particle sent into the same fields would be deflected by the magnetic force or electric force.

Going back to the Hall effect, if the current in the strip is \mathbf{I} , then from [Current and Resistance](#), we know that

$$\mathbf{I} = nev_d\mathbf{A} \quad (18.8.3)$$

where n is the number of charge carriers per volume and \mathbf{A} is the cross-sectional area of the strip. Combining the equations for v_d and \mathbf{I} results in

$$\mathbf{I} = ne \left(\frac{\mathbf{E}}{B} \right) \mathbf{A}. \quad (18.8.4)$$

The field **E** is related to the potential difference **V** between the edges of the strip by

$$E = \frac{V}{l}. \quad (18.8.5)$$

The quantity **V** is called the **Hall potential** and can be measured with a voltmeter. Finally, combining the equations for **I** and **E** gives us

$$V = \frac{IBl}{neA} \quad (18.8.6)$$

where the upper edge of the strip in Figure 18.8.1 is positive with respect to the lower edge.

We can also combine Equation 18.8.1 and Equation 18.8.5 to get an expression for the Hall voltage in terms of the magnetic field:

$$V = Blv_d.$$

What if the charge carriers are positive, as in Figure 18.8.1? For the same current **I**, the magnitude of **V** is still given by Equation 18.8.6. However, the upper edge is now negative with respect to the lower edge. Therefore, by simply measuring the sign of **V**, we can determine the sign of the majority charge carriers in a metal.

Hall potential measurements show that electrons are the dominant charge carriers in most metals. However, Hall potentials indicate that for a few metals, such as tungsten, beryllium, and many semiconductors, the majority of charge carriers are positive. It turns out that conduction by positive charge is caused by the migration of missing electron sites (called holes) on ions. Conduction by holes is studied later in [Condensed Matter Physics](#).

The Hall effect can be used to measure magnetic fields. If a material with a known density of charge carriers **n** is placed in a magnetic field and **V** is measured, then the field can be determined from Equation ???. In research laboratories where the fields of electromagnets used for precise measurements have to be extremely steady, a “Hall probe” is commonly used as part of an electronic circuit that regulates the field.

✓ Example 18.8.1: Velocity Selector

An electron beam enters a crossed-field velocity selector with magnetic and electric fields of 2.0 mT and $6.0 \times 10^3 \text{ N/C}$, respectively. (a) What must the velocity of the electron beam be to traverse the crossed fields undeflected? If the electric field is turned off, (b) what is the acceleration of the electron beam and (c) what is the radius of the circular motion that results?

Strategy

The electron beam is not deflected by either of the magnetic or electric fields if these forces are balanced. Based on these balanced forces, we calculate the velocity of the beam. Without the electric field, only the magnetic force is used in Newton’s second law to find the acceleration. Lastly, the radius of the path is based on the resulting circular motion from the magnetic force.

Solution

1. The velocity of the unperturbed beam of electrons with crossed fields is calculated by Equation 18.8.2:

$$v_d = \frac{E}{B} = \frac{6 \times 10^3 \text{ N/C}}{2 \times 10^{-3} \text{ T}} = 3 \times 10^6 \text{ m/s.}$$

2. The acceleration is calculated from the net force from the magnetic field, equal to mass times acceleration. The magnitude of the acceleration is:

$$ma = qvB$$

$$a = \frac{qvB}{m} = \frac{(1.6 \times 10^{-19} \text{ C})(3 \times 10^6 \text{ m/s})(2 \times 10^{-3} \text{ T})}{0.1 \times 10^{-31} \text{ kg}} = 1.1 \times 10^{15} \text{ m/s}^2.$$

3. The radius of the path comes from a balance of the circular and magnetic forces, or Equation 18.8.2:

$$r = \frac{mv}{qB} = \frac{(9.1 \times 10^{-31} \text{ kg})(3 \times 10^6 \text{ m/s})}{(1.6 \times 10^{-19} \text{ C})(2 \times 10^{-3} \text{ T})} = 8.5 \times 10^{-3} \text{ m.}$$

Significance

If electrons in the beam had velocities above or below the answer in part (a), those electrons would have a stronger net force exerted by either the magnetic or electric field. Therefore, only those electrons at this specific velocity would make it through.

✓ The Hall Potential in a Silver Ribbon

Figure 18.8.2 shows a silver ribbon whose cross section is 1.0 cm by 0.20 cm. The ribbon carries a current of 100 A from left to right, and it lies in a uniform magnetic field of magnitude 1.5 T. Using a density value of $n = 5.9 \times 10^{28}$ electrons per cubic meter for silver, find the Hall potential between the edges of the ribbon.

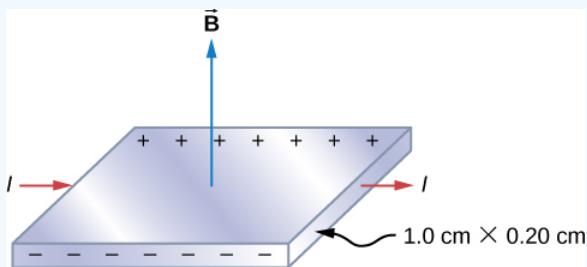


Figure 18.8.2: Finding the Hall potential in a silver ribbon in a magnetic field is shown.

Strategy

Since the majority of charge carriers are electrons, the polarity of the Hall voltage is that indicated in the figure. The value of the Hall voltage is calculated using Equation 18.8.6.

Solution

When calculating the Hall voltage, we need to know the current through the material, the magnetic field, the length, the number of charge carriers, and the area. Since all of these are given, the Hall voltage is calculated as:

$$\begin{aligned} v &= \frac{IBl}{neA} \\ &= \frac{(100 \text{ A})(1.5 \text{ T})(1.0 \times 10^{-2} \text{ m})}{(5.9 \times 10^{28}/\text{m}^3)(1.6 \times 10^{-19} \text{ C})(2.0 \times 10^{-5} \text{ m}^2)} \\ &= 7.9 \times 10^{-6} \text{ V}. \end{aligned}$$

Significance

As in this example, the Hall potential is generally very small, and careful experimentation with sensitive equipment is required for its measurement.

? Exercise 18.8.1

A Hall probe consists of a copper strip, $n = 8.5 \times 10^{28}$ electrons per cubic meter, which is 2.0 cm wide and 0.10 cm thick. What is the magnetic field when $I = 50 \text{ A}$ and the Hall potential is

- a. $4.0 \mu\text{V}$ and
- b. $6.0 \mu\text{V}$?

Answer a

1.1 T

Answer b

1.6 T

- **11.7: The Hall Effect** by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.9: More Applications of Magnetism

Learning Objectives

By the end of this section, you will be able to:

- Describe some applications of magnetism.

Mass Spectrometry

The curved paths followed by charged particles in magnetic fields can be put to use. A charged particle moving perpendicular to a magnetic field travels in a circular path having a radius r .

$$r = \frac{mv}{qB} \quad (18.9.1)$$

It was noted that this relationship could be used to measure the mass of charged particles such as ions. A mass spectrometer is a device that measures such masses. Most mass spectrometers use magnetic fields for this purpose, although some of them have extremely sophisticated designs. Since there are five variables in the relationship, there are many possibilities. However, if v , q , and B can be fixed, then the radius of the path r is simply proportional to the mass m of the charged particle. Let us examine one such mass spectrometer that has a relatively simple design (Figure 18.9.1). The process begins with an ion source, a device like an electron gun. The ion source gives ions their charge, accelerates them to some velocity v , and directs a beam of them into the next stage of the spectrometer. This next region is a *velocity selector* that only allows particles with a particular value of v to get through.

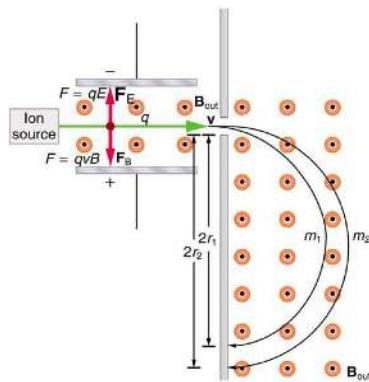


Figure 18.9.1: This mass spectrometer uses a velocity selector to fix v so that the radius of the path is proportional to mass.

The velocity selector has both an electric field and a magnetic field, perpendicular to one another, producing forces in opposite directions on the ions. Only those ions for which the forces balance travel in a straight line into the next region. If the forces balance, then the electric force $\mathbf{F} = q\mathbf{E}$ equals the magnetic force $\mathbf{F} = q\mathbf{v}\mathbf{B}$, so that $q\mathbf{E} = q\mathbf{v}\mathbf{B}$. Noting that q cancels, we see that

$$v = \frac{E}{B} \quad (18.9.2)$$

is the velocity particles must have to make it through the velocity selector, and further, that v can be selected by varying E and B . In the final region, there is only a uniform magnetic field, and so the charged particles move in circular arcs with radii proportional to particle mass. The paths also depend on charge q , but since q is in multiples of electron charges, it is easy to determine and to discriminate between ions in different charge states.

Mass spectrometry today is used extensively in chemistry and biology laboratories to identify chemical and biological substances according to their mass-to-charge ratios. In medicine, mass spectrometers are used to measure the concentration of isotopes used as tracers. Usually, biological molecules such as proteins are very large, so they are broken down into smaller fragments before analyzing. Recently, large virus particles have been analyzed as a whole on mass spectrometers. Sometimes a gas chromatograph or high-performance liquid chromatograph provides an initial separation of the large molecules, which are then input into the mass spectrometer.

Cathode Ray Tubes—CRTs—and the Like

What do non-flat-screen TVs, old computer monitors, x-ray machines, and the 2-mile-long Stanford Linear Accelerator have in common? All of them accelerate electrons, making them different versions of the electron gun. Many of these devices use magnetic fields to steer the accelerated electrons. Figure 18.9.2 shows the construction of the type of cathode ray tube (CRT) found in some TVs, oscilloscopes, and old computer monitors. Two pairs of coils are used to steer the electrons, one vertically and the other horizontally, to their desired destination.

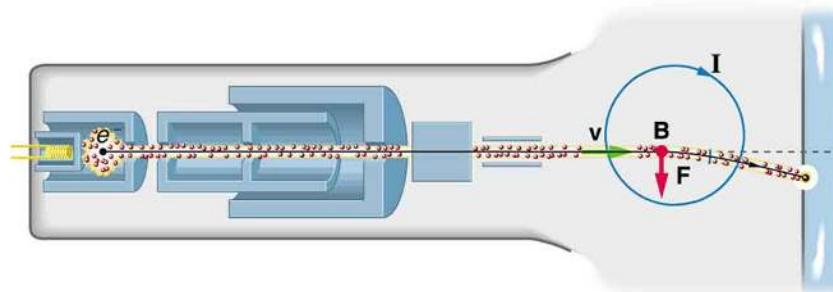


Figure 18.9.2: The cathode ray tube (CRT) is so named because rays of electrons originate at the cathode in the electron gun. Magnetic coils are used to steer the beam in many CRTs. In this case, the beam is moved down. Another pair of horizontal coils would steer the beam horizontally.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is one of the most useful and rapidly growing medical imaging tools. It non-invasively produces two-dimensional and three-dimensional images of the body that provide important medical information with none of the hazards of x-rays. MRI is based on an effect called **nuclear magnetic resonance (NMR)** in which an externally applied magnetic field interacts with the nuclei of certain atoms, particularly those of hydrogen (protons). These nuclei possess their own small magnetic fields, similar to those of electrons and the current loops discussed earlier in this chapter.

When placed in an external magnetic field, such nuclei experience a torque that pushes or aligns the nuclei into one of two new energy states—depending on the orientation of its spin (analogous to the N pole and S pole in a bar magnet). Transitions from the lower to higher energy state can be achieved by using an external radio frequency signal to “flip” the orientation of the small magnets. (This is actually a quantum mechanical process. The direction of the nuclear magnetic field is quantized as is energy in the radio waves. We will return to these topics in later chapters.) The specific frequency of the radio waves that are absorbed and reemitted depends sensitively on the type of nucleus, the chemical environment, and the external magnetic field strength. Therefore, this is a *resonance* phenomenon in which *nuclei* in a *magnetic* field act like resonators (analogous to those discussed in the treatment of sound in "Oscillatory Motion and Waves") that absorb and reemit only certain frequencies. Hence, the phenomenon is named *nuclear magnetic resonance (NMR)*.

NMR has been used for more than 50 years as an analytical tool. It was formulated in 1946 by F. Bloch and E. Purcell, with the 1952 Nobel Prize in Physics going to them for their work. Over the past two decades, NMR has been developed to produce detailed images in a process now called magnetic resonance imaging (MRI), a name coined to avoid the use of the word “nuclear” and the concomitant implication that nuclear radiation is involved. (It is not.) The 2003 Nobel Prize in Medicine went to P. Lauterbur and P. Mansfield for their work with MRI applications.

The largest part of the MRI unit is a superconducting magnet that creates a magnetic field, typically between 1 and 2 T in strength, over a relatively large volume. MRI images can be both highly detailed and informative about structures and organ functions. It is helpful that normal and non-normal tissues respond differently for slight changes in the magnetic field. In most medical images, the protons that are hydrogen nuclei are imaged. (About 2/3 of the atoms in the body are hydrogen.) Their location and density give a variety of medically useful information, such as organ function, the condition of tissue (as in the brain), and the shape of structures, such as vertebral disks and knee-joint surfaces. MRI can also be used to follow the movement of certain ions across membranes, yielding information on active transport, osmosis, dialysis, and other phenomena. With excellent spatial resolution, MRI can provide information about tumors, strokes, shoulder injuries, infections, etc.

An image requires position information as well as the density of a nuclear type (usually protons). By varying the magnetic field slightly over the volume to be imaged, the resonant frequency of the protons is made to vary with position. Broadcast radio frequencies are swept over an appropriate range and nuclei absorb and reemit them only if the nuclei are in a magnetic field with the correct strength. The imaging receiver gathers information through the body almost point by point, building up a tissue map. The reception of reemitted radio waves as a function of frequency thus gives position information. These “slices” or cross sections through the body are only several mm thick. The intensity of the reemitted radio waves is proportional to the concentration of the nuclear type being flipped, as well as information on the chemical environment in that area of the body. Various techniques are available for enhancing contrast in images and for obtaining more information. Scans called T1, T2, or proton density scans rely on different relaxation mechanisms of nuclei. Relaxation refers to the time it takes for the protons to return to equilibrium after the external field is turned off. This time depends upon tissue type and status (such as inflammation).

While MRI images are superior to x rays for certain types of tissue and have none of the hazards of x rays, they do not completely supplant x-ray images. MRI is less effective than x rays for detecting breaks in bone, for example, and in imaging breast tissue, so the two diagnostic tools complement each other. MRI images are also expensive compared to simple x-ray images and tend to be used most often where they supply information not readily obtained from x rays. Another disadvantage of MRI is that the patient is totally enclosed with detectors close to the body for about 30 minutes or more, leading to claustrophobia. It is also difficult for the obese patient to be in the magnet tunnel. New “open-MRI” machines are now available in which the magnet does not completely surround the patient.

Over the last decade, the development of much faster scans, called “functional MRI” (fMRI), has allowed us to map the functioning of various regions in the brain responsible for thought and motor control. This technique measures the change in blood flow for activities (thought, experiences, action) in the brain. The nerve cells increase their consumption of oxygen when active. Blood hemoglobin releases oxygen to active nerve cells and has somewhat different magnetic properties when oxygenated than when deoxygenated. With MRI, we can measure this and detect a blood oxygen-dependent signal. Most of the brain scans today use fMRI.

Other Medical Uses of Magnetic Fields

Currents in nerve cells and the heart create magnetic fields like any other currents. These can be measured but with some difficulty since their strengths are about 10^{-6} to 10^{-8} less than the Earth’s magnetic field. Recording of the heart’s magnetic field as it beats is called a **magnetocardiogram (MCG)**, while measurements of the brain’s magnetic field is called a **magnetoencephalogram (MEG)**. Both give information that differs from that obtained by measuring the electric fields of these organs (ECGs and EEGs), but they are not yet of sufficient importance to make these difficult measurements common.

In both of these techniques, the sensors do not touch the body. MCG can be used in fetal studies, and is probably more sensitive than echocardiography. MCG also looks at the heart’s electrical activity whose voltage output is too small to be recorded by surface electrodes as in EKG. It has the potential of being a rapid scan for early diagnosis of cardiac ischemia (obstruction of blood flow to the heart) or problems with the fetus.

MEG can be used to identify abnormal electrical discharges in the brain that produce weak magnetic signals. Therefore, it looks at brain activity, not just brain structure. It has been used for studies of Alzheimer’s disease and epilepsy. Advances in instrumentation to measure very small magnetic fields have allowed these two techniques to be used more in recent years. What is used is a sensor called a SQUID, for superconducting quantum interference device. This operates at liquid helium temperatures and can measure magnetic fields thousands of times smaller than the Earth’s.

Finally, there is a burgeoning market for magnetic cures in which magnets are applied in a variety of ways to the body, from magnetic bracelets to magnetic mattresses. The best that can be said for such practices is that they are apparently harmless, unless the magnets get close to the patient’s computer or magnetic storage disks. Claims are made for a broad spectrum of benefits from cleansing the blood to giving the patient more energy, but clinical studies have not verified these claims, nor is there an identifiable mechanism by which such benefits might occur.

PHET EXPLORATIONS: MAGNET AND COMPASS

Ever wonder how a compass worked to point you to the Arctic? Explore the interactions between a compass and bar magnet, and then add the Earth and find the surprising answer! Vary the magnet's strength, and see how things change both inside and outside. Use the field meter to measure how the magnetic field changes.

Summary

- Crossed (perpendicular) electric and magnetic fields act as a velocity filter, giving equal and opposite forces on any charge with velocity perpendicular to the fields and of magnitude

$$v = \frac{E}{B}.$$

Glossary

magnetic resonance imaging (MRI)

a medical imaging technique that uses magnetic fields to create detailed images of internal tissues and organs

nuclear magnetic resonance (NMR)

a phenomenon in which an externally applied magnetic field interacts with the nuclei of certain atoms

magnetocardiogram (MCG)

a recording of the heart's magnetic field as it beats

magnetoencephalogram (MEG)

a measurement of the brain's magnetic field

This page titled [18.9: More Applications of Magnetism](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [22.11: More Applications of Magnetism](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/college-physics>.

18.10: Superconductors

LEARNING OBJECTIVES

By the end of this section, you will be able to:

- Describe the phenomenon of superconductivity
- List applications of superconductivity

Touch the power supply of your laptop computer or some other device. It probably feels slightly warm. That heat is an unwanted byproduct of the process of converting household electric power into a current that can be used by your device. Although electric power is reasonably efficient, other losses are associated with it. As discussed in the section on power and energy, transmission of electric power produces I^2R line losses. These line losses exist whether the power is generated from conventional power plants (using coal, oil, or gas), nuclear plants, solar plants, hydroelectric plants, or wind farms. These losses can be reduced, but not eliminated, by transmitting using a higher voltage. It would be wonderful if these line losses could be eliminated, but that would require transmission lines that have zero resistance. In a world that has a global interest in not wasting energy, the reduction or elimination of this unwanted thermal energy would be a significant achievement. Is this possible?

The Resistance of Mercury

In 1911, Heike Kamerlingh Onnes of Leiden University, a Dutch physicist, was looking at the temperature dependence of the resistance of the element mercury. He cooled the sample of mercury and noticed the familiar behavior of a linear dependence of resistance on temperature; as the temperature decreased, the resistance decreased. Kamerlingh Onnes continued to cool the sample of mercury, using liquid helium. As the temperature approached 4.2 K (-269.2°C), the resistance abruptly went to zero (Figure 18.10.1). This temperature is known as the **critical temperature** T_c for mercury. The sample of mercury entered into a phase where the resistance was absolutely zero. This phenomenon is known as **superconductivity**. (Note: If you connect the leads of a three-digit ohmmeter across a conductor, the reading commonly shows up as $0.00\ \Omega$. The resistance of the conductor is not actually zero, it is less than $0.01\ \Omega$.) There are various methods to measure very small resistances, such as the four-point method, but an ohmmeter is not an acceptable method to use for testing resistance in superconductivity.

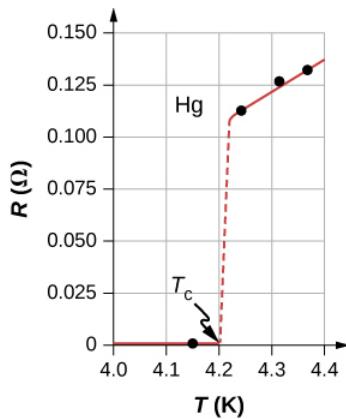


Figure 18.10.1: The resistance of a sample of mercury is zero at very low temperatures—it is a superconductor up to the temperature of about 4.2 K. Above that critical temperature, its resistance makes a sudden jump and then increases nearly linearly with temperature.

Other Superconducting Materials

As research continued, several other materials were found to enter a superconducting phase, when the temperature reached near absolute zero. In 1941, an alloy of niobium-nitride was found that could become superconducting at $T_c = 16\text{ K}$ (-257°C) and in 1953, vanadium-silicon was found to become superconductive at $T_c = 17.5\text{ K}$ (-255.7°C). The temperatures for the transition into superconductivity were slowly creeping higher. Strangely, many materials that make good conductors, such as copper, silver, and gold, do not exhibit superconductivity. Imagine the energy savings if transmission lines for electric power-generating stations could be made to be superconducting at temperatures near room temperature! A resistance of zero ohms means no I^2R losses and a great

boost to reducing energy consumption. The problem is that $T_c = 17.5\text{ K}$ is still very cold and in the range of liquid helium temperatures. At this temperature, it is not cost effective to transmit electrical energy because of the cooling requirements.

A large jump was seen in 1986, when a team of researchers, headed by Dr. Ching Wu Chu of Houston University, fabricated a brittle, ceramic compound with a transition temperature of $T_c = 92\text{ K} (-181^\circ\text{C})$. The ceramic material, composed of yttrium barium copper oxide (YBCO), was an insulator at room temperature. Although this temperature still seems quite cold, it is near the boiling point of liquid nitrogen, a liquid commonly used in refrigeration. You may have noticed refrigerated trucks traveling down the highway labeled as “Liquid Nitrogen Cooled.”

YBCO ceramic is a material that could be useful for transmitting electrical energy because the cost saving of reducing the I^2R losses are larger than the cost of cooling the superconducting cable, making it financially feasible. There were and are many engineering problems to overcome. For example, unlike traditional electrical cables, which are flexible and have a decent tensile strength, ceramics are brittle and would break rather than stretch under pressure. Processes that are rather simple with traditional cables, such as making connections, become difficult when working with ceramics. The problems are difficult and complex, and material scientists and engineers are coming up with innovative solutions.

An interesting consequence of the resistance going to zero is that once a current is established in a superconductor, it persists without an applied voltage source. Current loops in a superconductor have been set up and the current loops have been observed to persist for years without decaying.

Zero resistance is not the only interesting phenomenon that occurs as the materials reach their transition temperatures. A second effect is the exclusion of magnetic fields. This is known as the **Meissner effect** (Figure 18.10.2). A light, permanent magnet placed over a superconducting sample will levitate in a stable position above the superconductor. High-speed trains have been developed that levitate on strong superconducting magnets, eliminating the friction normally experienced between the train and the tracks. In Japan, the Yamanashi Maglev test line opened on April 3, 1997. In April 2015, the MLX01 test vehicle attained a speed of 374 mph (603 km/h).

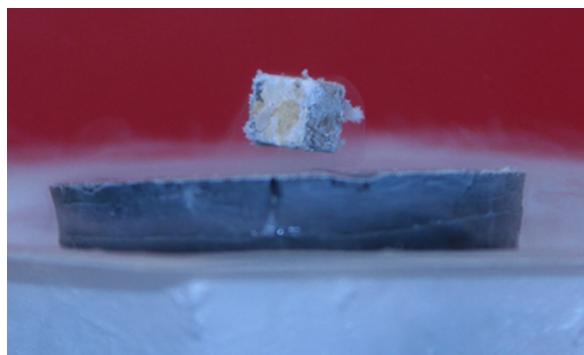


Figure 18.10.1: A small, strong magnet levitates over a superconductor cooled to liquid nitrogen temperature. The magnet levitates because the superconductor excludes magnetic fields.

Table 18.10.1 shows a select list of elements, compounds, and high-temperature superconductors, along with the critical temperatures for which they become superconducting. Each section is sorted from the highest critical temperature to the lowest. Also listed is the critical magnetic field for some of the materials. This is the strength of the magnetic field that destroys superconductivity. Finally, the type of the superconductor is listed.

There are two types of superconductors. There are 30 pure metals that exhibit zero resistivity below their critical temperature and exhibit the Meissner effect, the property of excluding magnetic fields from the interior of the superconductor while the superconductor is at a temperature below the critical temperature. These metals are called Type I superconductors. The superconductivity exists only below their critical temperatures and below a critical magnetic field strength. Type I superconductors are well described by the BCS theory (described next). Type I superconductors have limited practical applications because the strength of the critical magnetic field needed to destroy the superconductivity is quite low.

Type II superconductors are found to have much higher critical magnetic fields and therefore can carry much higher current densities while remaining in the superconducting state. A collection of various ceramics containing barium-copper-oxide have

much higher critical temperatures for the transition into a superconducting state. Superconducting materials that belong to this subcategory of the Type II superconductors are often categorized as high-temperature superconductors.

Introduction to BCS Theory

Type I superconductors, along with some Type II superconductors can be modeled using the **BCS theory**, proposed by John **Bardeen**, Leon **Cooper**, and Robert **Schrieffer**. Although the theory is beyond the scope of this chapter, a short summary of the theory is provided here. (More detail is provided in [Condensed Matter Physics](#).) The theory considers pairs of electrons and how they are coupled together through lattice-vibration interactions. Through the interactions with the crystalline lattice, electrons near the Fermi energy level feel a small attractive force and form pairs (**Cooper pairs**), and the coupling is known as a phonon interaction. Single electrons are fermions, which are particles that obey the [Pauli exclusion principle](#). The Pauli exclusion principle in quantum mechanics states that two identical fermions (particles with half-integer spin) cannot occupy the same quantum state simultaneously. Each electron has four quantum numbers (n , ℓ , m_ℓ , m_s). The principal quantum number (n) describes the energy of the electron, the orbital angular momentum quantum number (ℓ) indicates the most probable distance from the nucleus, the magnetic quantum number m_ℓ describes the energy levels in the subshell, and the electron spin quantum number m_s describes the orientation of the spin of the electron, either up or down. As the material enters a superconducting state, pairs of electrons act more like bosons, which can condense into the same energy level and need not obey the Pauli exclusion principle. The electron pairs have a slightly lower energy and leave an energy gap above them on the order of 0.001 eV. This energy gap inhibits collision interactions that lead to ordinary resistivity. When the material is below the critical temperature, the thermal energy is less than the band gap and the material exhibits zero resistivity.

Table 18.10.1: Superconductor Critical Temperatures

Material	Symbol or Formula	Critical Temperature	Critical Magnetic Field	Type
		K	Tesla	
Elements				
Lead	Pb	7.19	0.08	I
Lanthanum	La	(α) 4.90 - (β) 6.30		I
Tantalum	Ta	4.48	0.09	I
Mercury	Hg	(α) 4.15 - (β) 3.95	0.04	I
Tin	Sn	3.72	0.03	I
Indium	In	3.40	0.03	I
Thallium	Tl	2.39	0.03	I
Rhenium	Re	2.40	0.03	I
Thorium	Th	1.37	0.013	I
Protactinium	Pa	1.40		I
Aluminum	Al	1.20	0.01	I
Gallium	Ga	1.10	0.005	I
Zinc	Zn	0.86	0.014	I
Titanium	Ti	0.39	0.01	I
Uranium	U	(α) 0.68 - (β) 1.80		I
Cadmium	Cd	11.4	4.00	I

Material	Symbol or Formula	Critical Temperature $T_c(K)$	Critical Magnetic Field $H_c(T)$	Type
Compounds				
Niobium-germanium	Nb_3Ge	23.20	37.00	II
Niobium-tin	Nb_3Sn	18.30	30.00	II
Niobium-nitrite	NbN	16.00		II
Niobium-titanium	NbTi	10.00	15.00	II
High-Temperature Oxides				
	$HgBa_2CaCu_2O_8$	134.00		II
	$Ti_2Ba_2Ca_2Cu_3O_{10}$	125.00		II
	$YBa_2Cu_3O_7$	92.00	120.00	II

Applications of Superconductors

Superconductors can be used to make superconducting magnets. These magnets are 10 times stronger than the strongest electromagnets. These magnets are currently in use in magnetic resonance imaging (MRI), which produces high-quality images of the body interior without dangerous radiation.

Another interesting application of superconductivity is the **SQUID** (superconducting quantum interference device). A SQUID is a very sensitive magnetometer used to measure extremely subtle magnetic fields. The operation of the SQUID is based on superconducting loops containing Josephson junctions. A **Josephson junction** is the result of a theoretical prediction made by B. D. Josephson in an article published in 1962. In the article, Josephson described how a supercurrent can flow between two pieces of superconductor separated by a thin layer of insulator. This phenomenon is now called the **Josephson effect**. The SQUID consists of a superconducting current loop containing two Josephson junctions, as shown in Figure 18.10.3. When the loop is placed in even a very weak magnetic field, there is an interference effect that depends on the strength of the magnetic field.

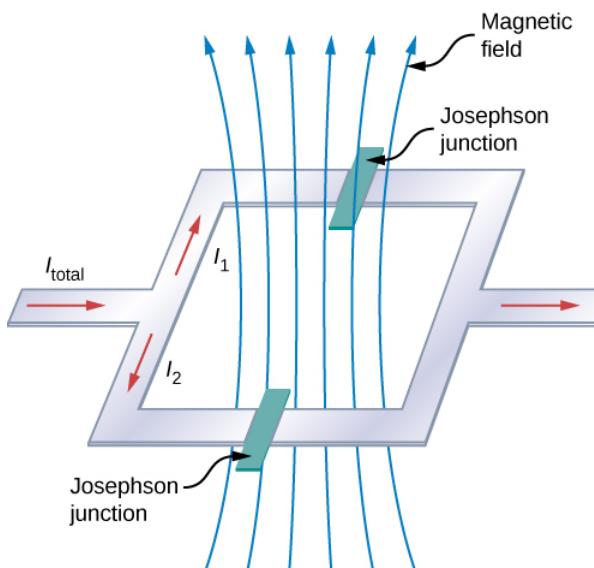


Figure 18.10.3: The SQUID (superconducting quantum interference device) uses a superconducting current loop and two Josephson junctions to detect magnetic fields as low as 10^{-14} (Earth's magnet field is on the order of $0.3 \times 10^{-5} T$).

Superconductivity is a fascinating and useful phenomenon. At critical temperatures near the boiling point of liquid nitrogen, superconductivity has special applications in MRIs, particle accelerators, and high-speed trains. Will we reach a state where we can

have materials enter the superconducting phase at near room temperatures? It seems a long way off, but if scientists in 1911 were asked if we would reach liquid-nitrogen temperatures with a ceramic, they might have thought it implausible.

This page titled [18.10: Superconductors](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [9.7: Superconductors](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.11: Conclusion

18.11: Conclusion is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

18.12: Magnetic Forces and Fields (Summary)

Key Terms

cosmic rays	comprised of particles that originate mainly from outside the solar system and reach Earth
cyclotron	device used to accelerate charged particles to large kinetic energies
dees	large metal containers used in cyclotrons that serve contain a stream of charged particles as their speed is increased
gauss	G, unit of the magnetic field strength; $1G = 10^{-4}T$
Hall effect	creation of voltage across a current-carrying conductor by a magnetic field
helical motion	superposition of circular motion with a straight-line motion that is followed by a charged particle moving in a region of magnetic field at an angle to the field
magnetic dipole	closed-current loop
magnetic dipole moment	term \mathbf{IA} of the magnetic dipole, also called μ
magnetic field lines	continuous curves that show the direction of a magnetic field; these lines point in the same direction as a compass points, toward the magnetic south pole of a bar magnet
magnetic force	force applied to a charged particle moving through a magnetic field
mass spectrometer	device that separates ions according to their charge-to-mass ratios
motor (dc)	loop of wire in a magnetic field; when current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft; electrical energy is converted into mechanical work in the process
north magnetic pole	currently where a compass points to north, near the geographic North Pole; this is the effective south pole of a bar magnet but has flipped between the effective north and south poles of a bar magnet multiple times over the age of Earth
right-hand rule-1	using your right hand to determine the direction of either the magnetic force, velocity of a charged particle, or magnetic field
south magnetic pole	currently where a compass points to the south, near the geographic South Pole; this is the effective north pole of a bar magnet but has flipped just like the north magnetic pole
tesla	SI unit for magnetic field: $1T = 1N/A - m$
velocity selector	apparatus where the crossed electric and magnetic fields produce equal and opposite forces on a charged particle moving with a specific velocity; this particle moves through the velocity selector not affected by either field while particles moving with different velocities are deflected by the apparatus

Key Equations

Force on a charge in a magnetic field	$\vec{F} = q\vec{v} \times \vec{B}$
Magnitude of magnetic force	$F = qvB\sin\theta$
Radius of a particle's path in a magnetic field	$r = \frac{mv}{qB}$
Period of a particle's motion in a magnetic field	$T = \frac{2\pi m}{qB}$
Force on a current-carrying wire in a uniform magnetic field	$\vec{F} = I\vec{l} \times \vec{B}$
Magnetic dipole moment	$\vec{\mu} = NIA\hat{n}$
Torque on a current loop	$\vec{\tau} = \vec{\mu} \times \vec{B}$
Energy of a magnetic dipole	$U = -\vec{\mu} \cdot \vec{B}$
Drift velocity in crossed electric and magnetic fields	$v_d = \frac{E}{B}$
Hall potential	$V = \frac{IBl}{neA}$
Hall potential in terms of drift velocity	$V = Blv_d$
Charge-to-mass ratio in a mass spectrometer	$\frac{q}{m} = \frac{E}{BB_0R}$
Maximum speed of a particle in a cyclotron	$v_{max} = \frac{qBR}{m}$

Summary

11.2 Magnetism and Its Historical Discoveries

- Magnets have two types of magnetic poles, called the north magnetic pole and the south magnetic pole. North magnetic poles are those that are attracted toward Earth's geographic North Pole.
- Like poles repel and unlike poles attract.
- Discoveries of how magnets respond to currents by Oersted and others created a framework that led to the invention of modern electronic devices, electric motors, and magnetic imaging technology.

11.3 Magnetic Fields and Lines

- Charges moving across a magnetic field experience a force determined by $\vec{F} = q\vec{v} \times \vec{B}$. The force is perpendicular to the plane formed by \vec{v} and \vec{B} .
- The direction of the force on a moving charge is given by the right hand rule 1 (RHR-1): Sweep your fingers in a velocity, magnetic field plane. Start by pointing them in the direction of velocity and sweep towards the magnetic field. Your thumb points in the direction of the magnetic force for positive charges.
- Magnetic fields can be pictorially represented by magnetic field lines, which have the following properties:
 1. The field is tangent to the magnetic field line.
 2. Field strength is proportional to the line density.
 3. Field lines cannot cross.
 4. Field lines form continuous, closed loops.
- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.

11.4 Motion of a Charged Particle in a Magnetic Field

- A magnetic force can supply centripetal force and cause a charged particle to move in a circular path of radius $r = \frac{mv}{qB}$.
- The period of circular motion for a charged particle moving in a magnetic field perpendicular to the plane of motion is $T = \frac{2\pi m}{qB}$.

- Helical motion results if the velocity of the charged particle has a component parallel to the magnetic field as well as a component perpendicular to the magnetic field.

11.5 Magnetic Force on a Current-Carrying Conductor

- An electrical current produces a magnetic field around the wire.
- The directionality of the magnetic field produced is determined by the right hand rule-2, where your thumb points in the direction of the current and your fingers wrap around the wire in the direction of the magnetic field.
- The magnetic force on current-carrying conductors is given by $\vec{F} = I\vec{l} \times \vec{B}$ where I is the current and l is the length of a wire in a uniform magnetic field B .

11.6 Force and Torque on a Current Loop

- The net force on a current-carrying loop of any plane shape in a uniform magnetic field is zero.
- The net torque τ on a current-carrying loop of any shape in a uniform magnetic field is calculated using $\tau = \vec{\mu} \times \vec{B}$ where $\vec{\mu}$ is the magnetic dipole moment and \vec{B} is the magnetic field strength.
- The magnetic dipole moment μ is the product of the number of turns of wire N , the current in the loop I , and the area of the loop A or $\vec{\mu} = NIA\hat{n}$.

11.7 The Hall Effect

- Perpendicular electric and magnetic fields exert equal and opposite forces for a specific velocity of entering particles, thereby acting as a velocity selector. The velocity that passes through undeflected is calculated by $v = \frac{E}{B}$.
- The Hall effect can be used to measure the sign of the majority of charge carriers for metals. It can also be used to measure a magnetic field.

11.8 Applications of Magnetic Forces and Fields

- A mass spectrometer is a device that separates ions according to their charge-to-mass ratios by first sending them through a velocity selector, then a uniform magnetic field.
- Cyclotrons are used to accelerate charged particles to large kinetic energies through applied electric and magnetic fields.

This page titled [18.12: Magnetic Forces and Fields \(Summary\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [11.9: Magnetic Forces and Fields \(Summary\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.13: Sources of Magnetic Fields (Summary)

Key Terms

Ampère's law	physical law that states that the line integral of the magnetic field around an electric current is proportional to the current
Biot-Savart law	an equation giving the magnetic field at a point produced by a current-carrying wire
diamagnetic materials	their magnetic dipoles align oppositely to an applied magnetic field; when the field is removed, the material is unmagnetized
ferromagnetic materials	contain groups of dipoles, called domains, that align with the applied magnetic field; when this field is removed, the material is still magnetized
hysteresis	property of ferromagnets that is seen when a material's magnetic field is examined versus the applied magnetic field; a loop is created resulting from sweeping the applied field forward and reverse
magnetic domains	groups of magnetic dipoles that are all aligned in the same direction and are coupled together quantum mechanically
magnetic susceptibility	ratio of the magnetic field in the material over the applied field at that time; positive susceptibilities are either paramagnetic or ferromagnetic (aligned with the field) and negative susceptibilities are diamagnetic (aligned oppositely with the field)
paramagnetic materials	their magnetic dipoles align partially in the same direction as the applied magnetic field; when this field is removed, the material is unmagnetized
permeability of free space	μ_0 , measure of the ability of a material, in this case free space, to support a magnetic field
solenoid	thin wire wound into a coil that produces a magnetic field when an electric current is passed through it
toroid	donut-shaped coil closely wound around that is one continuous wire

Key Equations

Permeability of free space	$\mu_0 = 4\pi \times 10^{-7} T \cdot m/A$
Contribution to magnetic field from a current element	$dB = \frac{\mu_0}{4\pi} \frac{Idl \sin\theta}{r^2}$
Biot-Savart law	$\vec{B} = \frac{\mu_0}{4\pi} \int_{\text{wire}} \frac{Id\vec{l} \times \hat{r}}{r^2}$
Magnetic field due to a long straight wire	$B = \frac{\mu_0 I}{2\pi R}$
Force between two parallel currents	$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}$
Magnetic field of a current loop	$B = \frac{\mu_0 I}{2R}$ (at center of loop)
Ampère's law	$\oint \vec{B} \cdot d\vec{l} = \mu_0 I$
Magnetic field strength inside a solenoid	$B = \mu_0 nI$
Magnetic field strength inside a toroid	$B = \frac{\mu_0 NI}{2\pi r}$
Magnetic permeability	$\mu = (1 + \chi)\mu_0$
Magnetic field of a solenoid filled with paramagnetic material	$B = \mu nI$

Summary

12.2 The Biot-Savart Law

- The magnetic field created by a current-carrying wire is found by the Biot-Savart law.
- The current element $Id\vec{l}$ produces a magnetic field a distance r away.

12.3 Magnetic Field Due to a Thin Straight Wire

- The strength of the magnetic field created by current in a long straight wire is given by $B = \frac{\mu_0 I}{2\pi R}$ (long straight wire) where I is the current, R is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \times 10^{-7} T \cdot m/A$ is the permeability of free space.
- The direction of the magnetic field created by a long straight wire is given by right-hand rule 2 (RHR-2): Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops created by it.

12.4 Magnetic Force between Two Parallel Currents

- The force between two parallel currents I_1 and I_2 , separated by a distance r , has a magnitude per unit length given by $\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}$.
- The force is attractive if the currents are in the same direction, repulsive if they are in opposite directions.

12.5 Magnetic Field of a Current Loop

- The magnetic field strength at the center of a circular loop is given by $B = \frac{\mu_0 I}{2R}$ (at center of loop), where R is the radius of the loop. RHR-2 gives the direction of the field about the loop.

12.6 Ampère's Law

- The magnetic field created by current following any path is the sum (or integral) of the fields due to segments along the path (magnitude and direction as for a straight wire), resulting in a general relationship between current and field known as Ampère's law.
- Ampère's law can be used to determine the magnetic field from a thin wire or thick wire by a geometrically convenient path of integration. The results are consistent with the Biot-Savart law.

12.7 Solenoids and Toroids

- The magnetic field strength inside a solenoid is

$$\mathbf{B} = \mu_0 n \mathbf{I}$$
 (inside a solenoid)

where n is the number of loops per unit length of the solenoid. The field inside is very uniform in magnitude and direction.

- The magnetic field strength inside a toroid is

$$B = \frac{\mu_0 N I}{2\pi r}$$
 (within the toroid)

where N is the number of windings. The field inside a toroid is not uniform and varies with the distance as $1/r$.

12.8 Magnetism in Matter

- Materials are classified as paramagnetic, diamagnetic, or ferromagnetic, depending on how they behave in an applied magnetic field.
- Paramagnetic materials have partial alignment of their magnetic dipoles with an applied magnetic field. This is a positive magnetic susceptibility. Only a surface current remains, creating a solenoid-like magnetic field.
- Diamagnetic materials exhibit induced dipoles opposite to an applied magnetic field. This is a negative magnetic susceptibility.
- Ferromagnetic materials have groups of dipoles, called domains, which align with the applied magnetic field. However, when the field is removed, the ferromagnetic material remains magnetized, unlike paramagnetic materials. This magnetization of the material versus the applied field effect is called hysteresis.

This page titled [18.13: Sources of Magnetic Fields \(Summary\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.9: Sources of Magnetic Fields \(Summary\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

18.14: Current and Resistance (Summary)

Key Terms

ampere (amp)	SI unit for current; $1 \text{ A} = 1 \text{ C/s}$
circuit	complete path that an electrical current travels along
conventional current	current that flows through a circuit from the positive terminal of a battery through the circuit to the negative terminal of the battery
critical temperature	temperature at which a material reaches superconductivity
current density	flow of charge through a cross-sectional area divided by the area
diode	nonohmic circuit device that allows current flow in only one direction
drift velocity	velocity of a charge as it moves nearly randomly through a conductor, experiencing multiple collisions, averaged over a length of a conductor, whose magnitude is the length of conductor traveled divided by the time it takes for the charges to travel the length
electrical conductivity	measure of a material's ability to conduct or transmit electricity
electrical current	rate at which charge flows, $I = \frac{dQ}{dt}$
electrical power	time rate of change of energy in an electric circuit
Josephson junction	junction of two pieces of superconducting material separated by a thin layer of insulating material, which can carry a supercurrent
Meissner effect	phenomenon that occurs in a superconducting material where all magnetic fields are expelled
nonohmic	type of a material for which Ohm's law is not valid
ohm	(Ω) unit of electrical resistance, $1 \Omega = 1 \text{ V/A}$
ohmic	type of a material for which Ohm's law is valid, that is, the voltage drop across the device is equal to the current times the resistance
Ohm's law	empirical relation stating that the current I is proportional to the potential difference V; it is often written as $V = IR$, where R is the resistance
resistance	electric property that impedes current; for ohmic materials, it is the ratio of voltage to current, $R = V/I$
resistivity	intrinsic property of a material, independent of its shape or size, directly proportional to the resistance, denoted by ρ
schematic	graphical representation of a circuit using standardized symbols for components and solid lines for the wire connecting the components
SQUID	(Superconducting Quantum Interference Device) device that is a very sensitive magnetometer, used to measure extremely subtle magnetic fields
superconductivity	phenomenon that occurs in some materials where the resistance goes to exactly zero and all magnetic fields are expelled, which occurs dramatically at some low critical temperature (T_c)

Key Equations

Average electrical current	$I_{ave} = \frac{\Delta Q}{\Delta t}$
Definition of an ampere	$1A = 1C/s$
Electrical current	$I = \frac{dQ}{dt}$
Drift velocity	$v_d = \frac{I}{nqA}$
Current density	$I = \iint_{area} \vec{J} \cdot d\vec{A}$
Resistivity	$\rho = \frac{E}{J}$
Common expression of Ohm's law	$V = IR$
Resistivity as a function of temperature	$\rho = \rho_0[1 + \alpha(T - T_0)]$
Definition of resistance	$R \equiv \frac{V}{I}$
Resistance of a cylinder of material	$R = \rho \frac{L}{A}$
Temperature dependence of resistance	$R = R_0(1 + \alpha\Delta T)$
Electric power	$P = IV$
Power dissipated by a resistor	$P = I^2R = \frac{V^2}{R}$

Summary

9.2 Electrical Current

- The average electrical current I_{ave} is the rate at which charge flows, given by $I_{ave} = \frac{\Delta Q}{\Delta t}$, where ΔQ is the amount of charge passing through an area in time Δt .
- The instantaneous electrical current, or simply the current I , is the rate at which charge flows. Taking the limit as the change in time approaches zero, we have $I = \frac{dQ}{dt}$, where $\frac{dQ}{dt}$ is the time derivative of the charge.
- The direction of conventional current is taken as the direction in which positive charge moves. In a simple direct-current (DC) circuit, this will be from the positive terminal of the battery to the negative terminal.
- The SI unit for current is the ampere, or simply the amp (A), where $1 A = 1C/mathrms.$
- Current consists of the flow of free charges, such as electrons, protons, and ions.

9.3 Model of Conduction in Metals

- The current through a conductor depends mainly on the motion of free electrons.
- When an electrical field is applied to a conductor, the free electrons in a conductor do not move through a conductor at a constant speed and direction; instead, the motion is almost random due to collisions with atoms and other free electrons.
- Even though the electrons move in a nearly random fashion, when an electrical field is applied to the conductor, the overall velocity of the electrons can be defined in terms of a drift velocity.
- The current density is a vector quantity defined as the current through an infinitesimal area divided by the area.
- The current can be found from the current density, $I = \iint_{area} \vec{J} \cdot d\vec{A}$.
- An incandescent light bulb is a filament of wire enclosed in a glass bulb that is partially evacuated. Current runs through the filament, where the electrical energy is converted to light and heat.

9.4 Resistivity and Resistance

- Resistance has units of ohms (Ω), related to volts and amperes by $1\Omega = 1V/A$.
- The resistance R of a cylinder of length L and cross-sectional area A is $R = \frac{\rho L}{A}$, where ρ is the resistivity of the material.

- Values of ρ in Table 9.1 show that materials fall into three groups—conductors, semiconductors, and insulators.
- Temperature affects resistivity; for relatively small temperature changes ΔT , resistivity is $\rho = \rho_0(1 + \alpha\Delta T)$, where ρ_0 is the original resistivity and α is the temperature coefficient of resistivity.
- The resistance R of an object also varies with temperature: $R = R_0(1 + \alpha\Delta T)$, where R_0 is the original resistance, and R is the resistance after the temperature change.

9.5 Ohm's Law

- Ohm's law is an empirical relationship for current, voltage, and resistance for some common types of circuit elements, including resistors. It does not apply to other devices, such as diodes.
- One statement of Ohm's law gives the relationship among current I , voltage V , and resistance R in a simple circuit as $V = IR$.
- Another statement of Ohm's law, on a microscopic level, is $J = \sigma E$.

9.6 Electrical Energy and Power

- Electric power is the rate at which electric energy is supplied to a circuit or consumed by a load.
- Power dissipated by a resistor depends on the square of the current through the resistor and is equal to $P = I^2R = \frac{V^2}{R}$.
- The SI unit for electric power is the watt and the SI unit for electric energy is the joule. Another common unit for electric energy, used by power companies, is the kilowatt-hour ($kW \cdot h$).
- The total energy used over a time interval can be found by $E = \int Pdt$.

9.7 Superconductors

- Superconductivity is a phenomenon that occurs in some materials when cooled to very low critical temperatures, resulting in a resistance of exactly zero and the expulsion of all magnetic fields.
- Materials that are normally good conductors (such as copper, gold, and silver) do not experience superconductivity.
- Superconductivity was first observed in mercury by Heike Kamerlingh Onnes in 1911. In 1986, Dr. Ching Wu Chu of Houston University fabricated a brittle, ceramic compound with a critical temperature close to the temperature of liquid nitrogen.
- Superconductivity can be used in the manufacture of superconducting magnets for use in MRIs and high-speed, levitated trains.

This page titled [18.14: Current and Resistance \(Summary\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [9.8: Current and Resistance \(Summary\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.15: Magnetic Forces and Fields (Exercise)

Conceptual Questions

11.3 Magnetic Fields and Lines

1. Discuss the similarities and differences between the electrical force on a charge and the magnetic force on a charge.
2. (a) Is it possible for the magnetic force on a charge moving in a magnetic field to be zero?
 - (b) Is it possible for the electric force on a charge moving in an electric field to be zero?
 - (c) Is it possible for the resultant of the electric and magnetic forces on a charge moving simultaneously through both fields to be zero?

11.4 Motion of a Charged Particle in a Magnetic Field

3. At a given instant, an electron and a proton are moving with the same velocity in a constant magnetic field. Compare the magnetic forces on these particles. Compare their accelerations.
4. Does increasing the magnitude of a uniform magnetic field through which a charge is traveling necessarily mean increasing the magnetic force on the charge? Does changing the direction of the field necessarily mean a change in the force on the charge?
5. An electron passes through a magnetic field without being deflected. What do you conclude about the magnetic field?
6. If a charged particle moves in a straight line, can you conclude that there is no magnetic field present?
7. How could you determine which pole of an electromagnet is north and which pole is south?

11.5 Magnetic Force on a Current-Carrying Conductor

8. Describe the error that results from accidentally using your left rather than your right hand when determining the direction of a magnetic force.
9. Considering the magnetic force law, are the velocity and magnetic field always perpendicular? Are the force and velocity always perpendicular? What about the force and magnetic field?
10. Why can a nearby magnet distort a cathode ray tube television picture?
11. A magnetic field exerts a force on the moving electrons in a current carrying wire. What exerts the force on a wire?
12. There are regions where the magnetic field of earth is almost perpendicular to the surface of Earth. What difficulty does this cause in the use of a compass?

11.7 The Hall Effect

13. Hall potentials are much larger for poor conductors than for good conductors. Why?

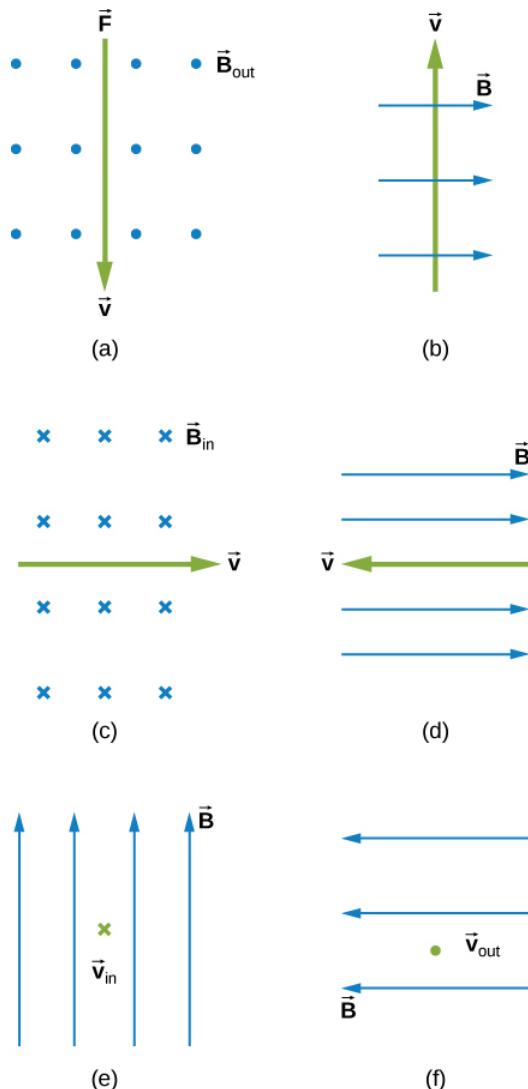
11.8 Applications of Magnetic Forces and Fields

14. Describe the primary function of the electric field and the magnetic field in a cyclotron.

Problems

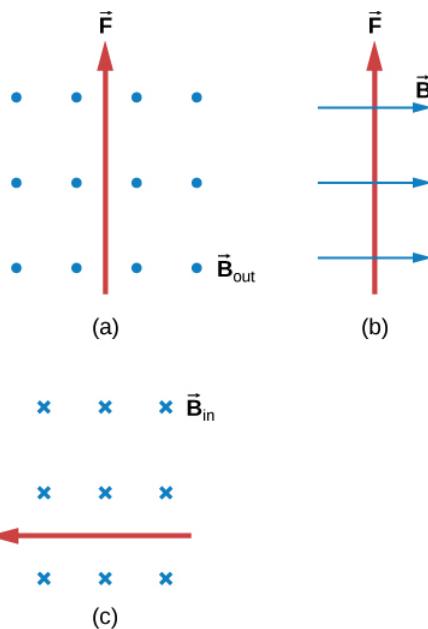
11.3 Magnetic Fields and Lines

15. What is the direction of the magnetic force on a positive charge that moves as shown in each of the six cases?



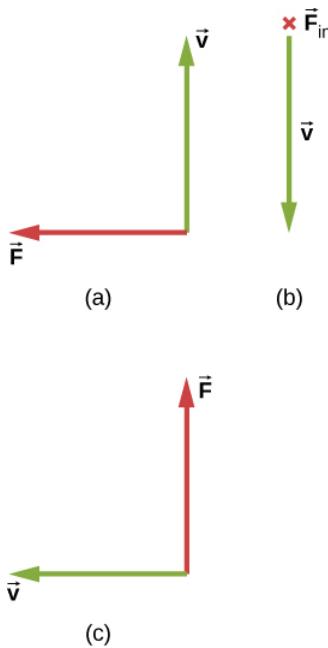
16. Repeat previous exercise for a negative charge.

17. What is the direction of the velocity of a negative charge that experiences the magnetic force shown in each of the three cases, assuming it moves perpendicular to B ?



18. Repeat previous exercise for a positive charge.

19. What is the direction of the magnetic field that produces the magnetic force on a positive charge as shown in each of the three cases, assuming \vec{B} is perpendicular to \vec{v} ?



20. Repeat previous exercise for a negative charge.

21. (a) Aircraft sometimes acquire small static charges. Suppose a supersonic jet has a $0.500\text{-}\mu\text{C}$ charge and flies due west at a speed of $660.\text{ m/s}$ over Earth's south magnetic pole, where the $8.00 \times 10^{-5} - \text{T}$ magnetic field points straight up. What are the direction and the magnitude of the magnetic force on the plane?

(b) Discuss whether the value obtained in part (a) implies this is a significant or negligible effect.

22. (a) A cosmic ray proton moving toward Earth at $5.00 \times 10^7 \text{ m/s}$ experiences a magnetic force of $1.70 \times 10^{-16} \text{ N}$. What is the strength of the magnetic field if there is a 45° angle between it and the proton's velocity?

(b) Is the value obtained in part a. consistent with the known strength of Earth's magnetic field on its surface? Discuss.

23. An electron moving at $4.00 \times 10^3 \text{ m/s}$ in a 1.25-T magnetic field experiences a magnetic force of $1.40 \times 10^{-16} \text{ N}$. What angle does the velocity of the electron make with the magnetic field? There are two answers.

24. (a) A physicist performing a sensitive measurement wants to limit the magnetic force on a moving charge in her equipment to less than $1.00 \times 10^{-12} \text{ N}$. What is the greatest the charge can be if it moves at a maximum speed of 30.0 m/s in Earth's field?

(b) Discuss whether it would be difficult to limit the charge to less than the value found in (a) by comparing it with typical static electricity and noting that static is often absent.

11.4 Motion of a Charged Particle in a Magnetic Field

25. A cosmic-ray electron moves at $7.5 \times 10^6 \text{ m/s}$ perpendicular to Earth's magnetic field at an altitude where the field strength is $1.0 \times 10^{-5} \text{ T}$. What is the radius of the circular path the electron follows?

26. (a) Viewers of Star Trek have heard of an antimatter drive on the Starship Enterprise. One possibility for such a futuristic energy source is to store antimatter charged particles in a vacuum chamber, circulating in a magnetic field, and then extract them as needed. Antimatter annihilates normal matter, producing pure energy. What strength magnetic field is needed to hold antiprotons, moving at $5.0 \times 10^7 \text{ m/s}$ in a circular path 2.00 m in radius? Antiprotons have the same mass as protons but the opposite (negative) charge.

(b) Is this field strength obtainable with today's technology or is it a futuristic possibility?

27. (a) An oxygen-16 ion with a mass of $2.66 \times 10^{-26} \text{ kg}$ travels at $5.0 \times 10^6 \text{ m/s}$ perpendicular to a 1.20-T magnetic field, which makes it move in a circular arc with a 0.231-m radius. What positive charge is on the ion?

(b) What is the ratio of this charge to the charge of an electron? (c) Discuss why the ratio found in (b) should be an integer.

28. An electron in a TV CRT moves with a speed of $6.0 \times 10^7 \text{ m/s}$, in a direction perpendicular to Earth's field, which has a strength of $5.0 \times 10^{-5} \text{ T}$. (a) What strength electric field must be applied perpendicular to the Earth's field to make the electron move in a straight line? (b) If this is done between plates separated by 1.00 cm , what is the voltage applied? (Note that TVs are usually surrounded by a ferromagnetic material to shield against external magnetic fields and avoid the need for such a correction.)

29. (a) At what speed will a proton move in a circular path of the same radius as the electron in the previous exercise?

(b) What would the radius of the path be if the proton had the same speed as the electron?

(c) What would the radius be if the proton had the same kinetic energy as the electron?

(d) The same momentum?

30. (a) What voltage will accelerate electrons to a speed of $6.00 \times 10^7 \text{ m/s}$? (b) Find the radius of curvature of the path of a proton accelerated through this potential in a 0.500-T field and compare this with the radius of curvature of an electron accelerated through the same potential.

31. An alpha-particle ($m = 6.64 \times 10^{-27} \text{ kg}$, $q = 3.2 \times 10^{-19} \text{ C}$) travels in a circular path of radius 25 cm in a uniform magnetic field of magnitude 1.5 T .

(a) What is the speed of the particle?

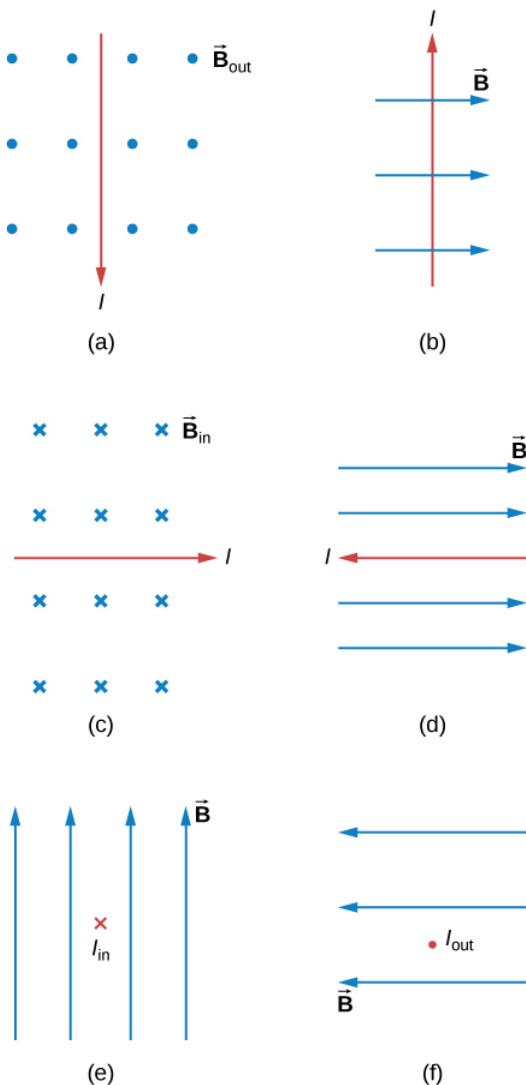
(b) What is the kinetic energy in electron-volts?

(c) Through what potential difference must the particle be accelerated in order to give it this kinetic energy?

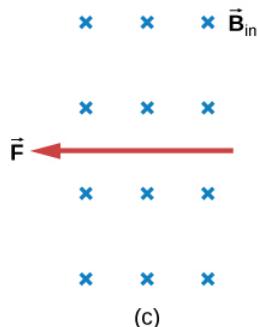
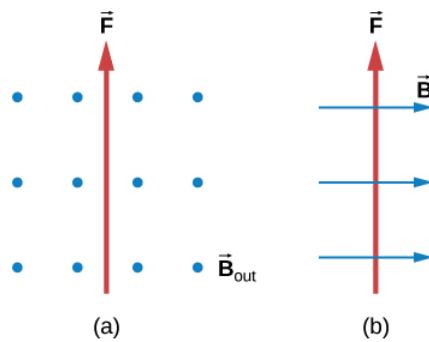
32. A particle of charge q and mass m is accelerated from rest through a potential difference V , after which it encounters a uniform magnetic field B . If the particle moves in a plane perpendicular to B , what is the radius of its circular orbit?

11.5 Magnetic Force on a Current-Carrying Conductor

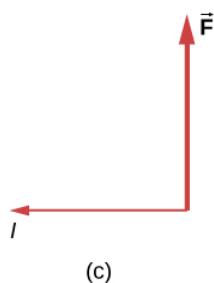
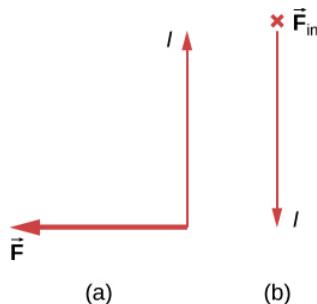
33. What is the direction of the magnetic force on the current in each of the six cases?



34. What is the direction of a current that experiences the magnetic force shown in each of the three cases, assuming the current runs perpendicular to \vec{B} ?



35. What is the direction of the magnetic field that produces the magnetic force shown on the currents in each of the three cases, assuming \vec{B} is perpendicular to \vec{I} ?



36. (a) What is the force per meter on a lightning bolt at the equator that carries 20,000 A perpendicular to Earth's $3.0 \times 10^{-5} T$ field? (b) What is the direction of the force if the current is straight up and Earth's field direction is due north, parallel to the ground?

37. (a) A dc power line for a light-rail system carries 1000 A at an angle of 30.0° to Earth's $5.0 \times 10^{-5} T$ field. What is the force on a 100-m section of this line?

(b) Discuss practical concerns this presents, if any.

38. A wire carrying a 30.0-A current passes between the poles of a strong magnet that is perpendicular to its field and experiences a 2.16-N force on the 4.00 cm of wire in the field. What is the average field strength?

11.6 Force and Torque on a Current Loop

39. (a) By how many percent is the torque of a motor decreased if its permanent magnets lose 5.0% of their strength?

(b) How many percent would the current need to be increased to return the torque to original values?

40. (a) What is the maximum torque on a 150-turn square loop of wire 18.0 cm on a side that carries a 50.0-A current in a 1.60-T field?

(b) What is the torque when θ is 10.9° ?

41. Find the current through a loop needed to create a maximum torque of **9.0 N·m**. The loop has 50 square turns that are 15.0 cm on a side and is in a uniform 0.800-T magnetic field.

42. Calculate the magnetic field strength needed on a 200-turn square loop 20.0 cm on a side to create a maximum torque of 300 N · m if the loop is carrying 25.0 A.

43. Since the equation for torque on a current-carrying loop is $\tau = NIAB \sin \theta$, the units of N · m must equal units of $A \cdot m^2 T$. Verify this.

44. (a) At what angle θ is the torque on a current loop 90.0% of maximum?

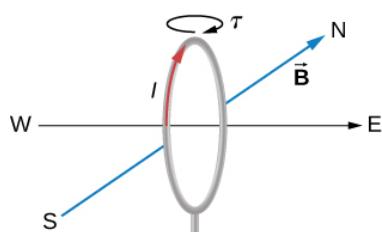
(b) 50.0% of maximum?

(c) 10.0% of maximum?

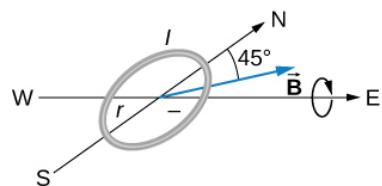
45. A proton has a magnetic field due to its spin. The field is similar to that created by a circular current loop $0.65 \times 10^{-15} m$ in radius with a current of $1.05 \times 10^4 A$. Find the maximum torque on a proton in a 2.50-T field. (This is a significant torque on a small particle.)

46. (a) A 200-turn circular loop of radius 50.0 cm is vertical, with its axis on an east-west line. A current of 100 A circulates clockwise in the loop when viewed from the east. Earth's field here is due north, parallel to the ground, with a strength of $3.0 \times 10^{-5} T$. What are the direction and magnitude of the torque on the loop?

(b) Does this device have any practical applications as a motor?



47. Repeat the previous problem, but with the loop lying flat on the ground with its current circulating counterclockwise (when viewed from above) in a location where Earth's field is north, but at an angle 45.0° below the horizontal and with a strength of $6.0 \times 10^{-5} T$.



11.7 The Hall Effect

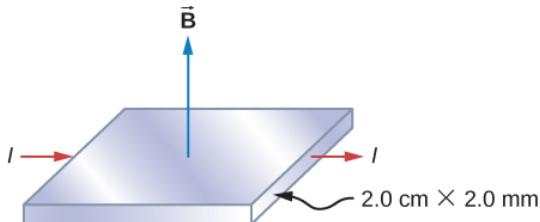
48. A strip of copper is placed in a uniform magnetic field of magnitude 2.5 T. The Hall electric field is measured to be $1.5 \times 10^{-3} V/m$.

(a) What is the drift speed of the conduction electrons?

(b) Assuming that $n = 8.0 \times 10^{28}$ electrons per cubic meter and that the cross-sectional area of the strip is $5.0 \times 10^{-6} m^2$, calculate the current in the strip.

(c) What is the Hall coefficient $1/nq$?

49. The cross-sectional dimensions of the copper strip shown are 2.0 cm by 2.0 mm. The strip carries a current of 100 A, and it is placed in a magnetic field of magnitude $B = 1.5$ T. What are the value and polarity of the Hall potential in the copper strip?



50. The magnitudes of the electric and magnetic fields in a velocity selector are $1.8 \times 10^5 V/m$ and 0.080 T, respectively.

(a) What speed must a proton have to pass through the selector?

(b) Also calculate the speeds required for an alpha-particle and a singly ionized O^{16} atom to pass through the selector.

51. A charged particle moves through a velocity selector at constant velocity. In the selector, $E = 1.0 \times 10^4 N/C$ and $B = 0.250$ T. When the electric field is turned off, the charged particle travels in a circular path of radius 3.33 mm. Determine the charge-to-mass ratio of the particle.

52. A Hall probe gives a reading of $1.5\mu V$ for a current of 2 A when it is placed in a magnetic field of 1 T. What is the magnetic field in a region where the reading is $2\mu V$ for 1.7 A of current?

11.8 Applications of Magnetic Forces and Fields

53. A physicist is designing a cyclotron to accelerate protons to one-tenth the speed of light. The magnetic field will have a strength of 1.5 T. Determine

(a) the rotational period of the circulating protons and

(b) the maximum radius of the protons' orbit.

54. The strengths of the fields in the velocity selector of a Bainbridge mass spectrometer are $B = 0.500$ T and $E = 1.2 \times 10^5 V/m$, and the strength of the magnetic field that separates the ions is $B_o = 0.750$ T. A stream of singly charged Li ions is found to bend in a circular arc of radius 2.32 cm. What is the mass of the Li ions?

55. The magnetic field in a cyclotron is 1.25 T, and the maximum orbital radius of the circulating protons is 0.40 m.

(a) What is the kinetic energy of the protons when they are ejected from the cyclotron?

(b) What is this energy in MeV?

(c) Through what potential difference would a proton have to be accelerated to acquire this kinetic energy?

(d) What is the period of the voltage source used to accelerate the protons?

(e) Repeat the calculations for alpha-particles.

56. A mass spectrometer is being used to separate common oxygen-16 from the much rarer oxygen-18, taken from a sample of old glacial ice. (The relative abundance of these oxygen isotopes is related to climatic temperature at the time the ice was deposited.) The ratio of the masses of these two ions is 16 to 18, the mass of oxygen-16 is $2.66 \times 10^{-26} kg$, and they are singly charged and travel at $5.00 \times 10^6 m/s$ in a 1.20-T magnetic field. What is the separation between their paths when they hit a target after traversing a semicircle?

57. (a) Triply charged uranium-235 and uranium-238 ions are being separated in a mass spectrometer. (The much rarer uranium-235 is used as reactor fuel.) The masses of the ions are $3.90 \times 10^{-25} kg$ and $3.95 \times 10^{-25} kg$, respectively, and they travel at $3.0 \times 10^6 m/s$ in a 0.250-T field. What is the separation between their paths when they hit a target after traversing a

semicircle? (b) Discuss whether this distance between their paths seems to be big enough to be practical in the separation of uranium-235 from uranium-238.

Additional Problems

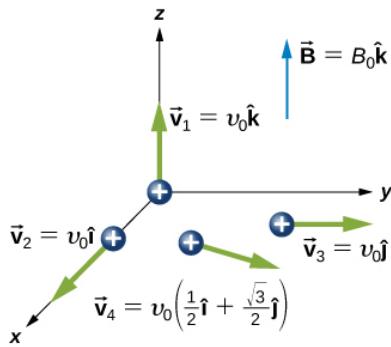
58. Calculate the magnetic force on a hypothetical particle of charge $1.0 \times 10^{-19} C$ moving with a velocity of $6.0 \times 10^4 \hat{i} m/s$ in a magnetic field of $1.2 \hat{k} T$.

59. Repeat the previous problem with a new magnetic field of $(0.4\hat{i} + 1.2\hat{k}) T$.

60. An electron is projected into a uniform magnetic field $(0.5\hat{i} + 0.8\hat{k}) T$ with a velocity of $(3.0\hat{i} + 4.0\hat{j}) \times 10^6 m/s$. What is the magnetic force on the electron?

61. The mass and charge of a water droplet are $1.0 \times 10^{-4} g$ and $2.0 \times 10^{-8} C$, respectively. If the droplet is given an initial horizontal velocity of $5.0 \times 10^5 \hat{i} m/s$, what magnetic field will keep it moving in this direction? Why must gravity be considered here?

62. Four different proton velocities are given. For each case, determine the magnetic force on the proton in terms of e , v_0 , and B_0 .



63. An electron of kinetic energy 2000 eV passes between parallel plates that are 1.0 cm apart and kept at a potential difference of 300 V. What is the strength of the uniform magnetic field B that will allow the electron to travel undeflected through the plates? Assume E and B are perpendicular.

64. An alpha-particle ($m = 6.64 \times 10^{-27} kg$, $q = 3.2 \times 10^{-19} C$) moving with a velocity $\vec{v} = (2.0\hat{i} - 4.0\hat{k}) \times 10^6 m/s$ enters a region where $\vec{E} = (5.0\hat{i} - 2.0\hat{j}) \times 10^4 V/m$ and $\vec{B} = (1.0\hat{i} + 4.0\hat{k}) \times 10^{-2} T$. What is the initial force on it?

65. An electron moving with a velocity $\vec{v} = (4.0\hat{i} + 3.0\hat{j} + 2.0\hat{k}) \times 10^6 m/s$ enters a region where there is a uniform electric field and a uniform magnetic field. The magnetic field is given by $\vec{B} = (1.0\hat{i} - 2.0\hat{j} + 4.0\hat{k}) \times 10^{-2} T$. If the electron travels through a region without being deflected, what is the electric field?

66. At a particular instant, an electron is traveling west to east with a kinetic energy of 10 keV. Earth's magnetic field has a horizontal component of $1.8 \times 10^{-5} T$ north and a vertical component of $5.0 \times 10^{-5} T$ down. (a) What is the path of the electron? (b) What is the radius of curvature of the path?

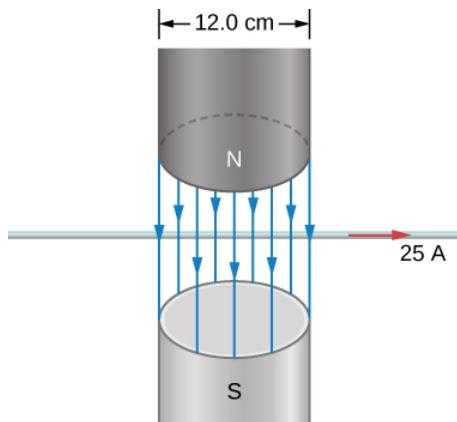
67. What is the (a) path of a proton and (b) the magnetic force on the proton that is traveling west to east with a kinetic energy of 10 keV in Earth's magnetic field that has a horizontal component of $1.8 \times 10^{-5} T$ north and a vertical component of $5.0 \times 10^{-5} T$ down?

68. What magnetic field is required in order to confine a proton moving with a speed of $4.0 \times 10^6 m/s$ to a circular orbit of radius 10 cm?

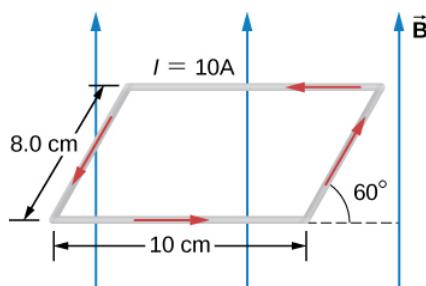
69. An electron and a proton move with the same speed in a plane perpendicular to a uniform magnetic field. Compare the radii and periods of their orbits.

70. A proton and an alpha-particle have the same kinetic energy and both move in a plane perpendicular to a uniform magnetic field. Compare the periods of their orbits.

- 71.** A singly charged ion takes $2.0 \times 10^{-3}\text{s}$ to complete eight revolutions in a uniform magnetic field of magnitude $2.0 \times 10^{-2}\text{T}$. What is the mass of the ion?
- 72.** A particle moving downward at a speed of $6.0 \times 10^6\text{m/s}$ enters a uniform magnetic field that is horizontal and directed from east to west.
- If the particle is deflected initially to the north in a circular arc, is its charge positive or negative?
 - If $B = 0.25\text{ T}$ and the charge-to-mass ratio (q/m) of the particle is $4.0 \times 10^7\text{C/kg}$, what is the radius of the path?
 - What is the speed of the particle after it has moved in the field for $1.0 \times 10^{-5}\text{s}$? for 2.0 s ?
- 73.** A proton, deuteron, and an alpha-particle are all accelerated through the same potential difference. They then enter the same magnetic field, moving perpendicular to it. Compute the ratios of the radii of their circular paths. Assume that $m_d = 2m_p$ and $m_\alpha = 4m_p$.
- 74.** A singly charged ion is moving in a uniform magnetic field of $7.5 \times 10^{-2}\text{T}$ completes 10 revolutions in $3.47 \times 10^{-4}\text{s}$. Identify the ion.
- 75.** Two particles have the same linear momentum, but particle A has four times the charge of particle B. If both particles move in a plane perpendicular to a uniform magnetic field, what is the ratio R_A/R_B of the radii of their circular orbits?
- 76.** A uniform magnetic field of magnitude B is directed parallel to the z-axis. A proton enters the field with a velocity $\vec{v} = (4\hat{j} + 3\hat{k}) \times 10^6\text{m/s}$ and travels in a helical path with a radius of 5.0 cm.
- What is the value of B?
 - What is the time required for one trip around the helix?
 - Where is the proton $5.0 \times 10^{-7}\text{s}$ after entering the field?
- 77.** An electron moving at $5.0 \times 10^6\text{m/s}$ enters a magnetic field that makes a 75° angle with the x-axis of magnitude 0.20 T. Calculate the
- pitch and
 - radius of the trajectory.
- 78.** (a) A 0.750-m-long section of cable carrying current to a car starter motor makes an angle of 60° with Earth's $5.5 \times 10^{-5}\text{T}$ field. What is the current when the wire experiences a force of $7.0 \times 10^{-3}\text{N}$?
- If you run the wire between the poles of a strong horseshoe magnet, subjecting 5.00 cm of it to a 1.75-T field, what force is exerted on this segment of wire?
- 79.** (a) What is the angle between a wire carrying an 8.00-A current and the 1.20-T field it is in if 50.0 cm of the wire experiences a magnetic force of 2.40 N?
- What is the force on the wire if it is rotated to make an angle of 90° with the field?
- 80.** A 1.0-m-long segment of wire lies along the x-axis and carries a current of 2.0 A in the positive x-direction. Around the wire is the magnetic field of $(3.0\hat{i} + 4.0\hat{k}) \times 10^{-3}\text{T}$. Find the magnetic force on this segment.
- 81.** A 5.0-m section of a long, straight wire carries a current of 10 A while in a uniform magnetic field of magnitude $8.0 \times 10^{-3}\text{T}$. Calculate the magnitude of the force on the section if the angle between the field and the direction of the current is
- 45° ;
 - 90° ;
 - 0° ; or
 - 180° .
- 82.** An electromagnet produces a magnetic field of magnitude 1.5 T throughout a cylindrical region of radius 6.0 cm. A straight wire carrying a current of 25 A passes through the field as shown in the accompanying figure. What is the magnetic force on the wire?



- 83.** The current loop shown in the accompanying figure lies in the plane of the page, as does the magnetic field. Determine the net force and the net torque on the loop if $I = 10 \text{ A}$ and $\mathbf{B} = 1.5 \text{ T}$.



- 84.** A circular coil of radius 5.0 cm is wound with five turns and carries a current of 5.0 A. If the coil is placed in a uniform magnetic field of strength 5.0 T, what is the maximum torque on it?

- 85.** A circular coil of wire of radius 5.0 cm has 20 turns and carries a current of 2.0 A. The coil lies in a magnetic field of magnitude 0.50 T that is directed parallel to the plane of the coil.

(a) What is the magnetic dipole moment of the coil?

(b) What is the torque on the coil?

- 86.** A current-carrying coil in a magnetic field experiences a torque that is 75% of the maximum possible torque. What is the angle between the magnetic field and the normal to the plane of the coil?

- 87.** A 4.0-cm by 6.0-cm rectangular current loop carries a current of 10 A. What is the magnetic dipole moment of the loop?

- 88.** A circular coil with 200 turns has a radius of 2.0 cm.

(a) What current through the coil results in a magnetic dipole moment of 3.0 Am^2 ?

(b) What is the maximum torque that the coil will experience in a uniform field of strength $5.0 \times 10^{-2} \text{ T}$?

(c) If the angle between μ and B is 45° , what is the magnitude of the torque on the coil?

(d) What is the magnetic potential energy of coil for this orientation?

- 89.** The current through a circular wire loop of radius 10 cm is 5.0 A.

(a) Calculate the magnetic dipole moment of the loop.

(b) What is the torque on the loop if it is in a uniform 0.20-T magnetic field such that μ and B are directed at 30° to each other?

(c) For this position, what is the potential energy of the dipole?

- 90.** A wire of length 1.0 m is wound into a single-turn planar loop. The loop carries a current of 5.0 A, and it is placed in a uniform magnetic field of strength 0.25 T.

- (a) What is the maximum torque that the loop will experience if it is square?
- (b) If it is circular?
- (c) At what angle relative to B would the normal to the circular coil have to be oriented so that the torque on it would be the same as the maximum torque on the square coil?

91. Consider an electron rotating in a circular orbit of radius r . Show that the magnitudes of the magnetic dipole moment μ and the angular momentum L of the electron are related by:

$$\frac{\mu}{L} = \frac{e}{2m}.$$

92. The Hall effect is to be used to find the sign of charge carriers in a semiconductor sample. The probe is placed between the poles of a magnet so that magnetic field is pointed up. A current is passed through a rectangular sample placed horizontally. As current is passed through the sample in the east direction, the north side of the sample is found to be at a higher potential than the south side. Decide if the number density of charge carriers is positively or negatively charged.

93. The density of charge carriers for copper is 8.47×10^{28} electrons per cubic meter. What will be the Hall voltage reading from a probe made up of **3cm×2 cm×1 cm(L×W×T)** copper plate when a current of **1.5 A** is passed through it in a magnetic field of **2.5 T** perpendicular to the **3cm×2 cm**.

94. The Hall effect is to be used to find the density of charge carriers in an unknown material. A Hall voltage **40 μ V** for **3-A** current is observed in a **3-T** magnetic field for a rectangular sample with length **2 cm**, width **1.5 cm**, and height **0.4 cm**. Determine the density of the charge carriers.

95. Show that the Hall voltage across wires made of the same material, carrying identical currents, and subjected to the same magnetic field is inversely proportional to their diameters. (Hint: Consider how drift velocity depends on wire diameter.)

96. A velocity selector in a mass spectrometer uses a **0.100-T** magnetic field.

- (a) What electric field strength is needed to select a speed of **$4.0 \times 10^6 m/s$** ?
- (b) What is the voltage between the plates if they are separated by **1.00 cm**?

97. Find the radius of curvature of the path of a **25.0-MeV** proton moving perpendicularly to the **1.20-T** field of a cyclotron.

98. Unreasonable results To construct a non-mechanical water meter, a **0.500-T** magnetic field is placed across the supply water pipe to a home and the Hall voltage is recorded.

- (a) Find the flow rate through a **3.00-cm-diameter** pipe if the Hall voltage is **60.0 mV**.
- (b) What would the Hall voltage be for the same flow rate through a **10.0-cm-diameter** pipe with the same field applied?

99. Unreasonable results A charged particle having mass **$6.64 \times 10^{-27} kg$** (that of a helium atom) moving at **$8.70 \times 10^5 m/s$** perpendicular to a **1.50-T** magnetic field travels in a circular path of radius **16.0 mm**.

- (a) What is the charge of the particle?
- (b) What is unreasonable about this result?
- (c) Which assumptions are responsible?

100. Unreasonable results An inventor wants to generate **120-V** power by moving a **1.00-m-long** wire perpendicular to Earth's **$5.00 \times 10^{-5} T$** field.

- (a) Find the speed with which the wire must move.
- (b) What is unreasonable about this result? (c) Which assumption is responsible?

101. Unreasonable results Frustrated by the small Hall voltage obtained in blood flow measurements, a medical physicist decides to increase the applied magnetic field strength to get a **0.500-V** output for blood moving at **30.0 cm/s** in a **1.50-cm-diameter** vessel.

- (a) What magnetic field strength is needed?

(b) What is unreasonable about this result?

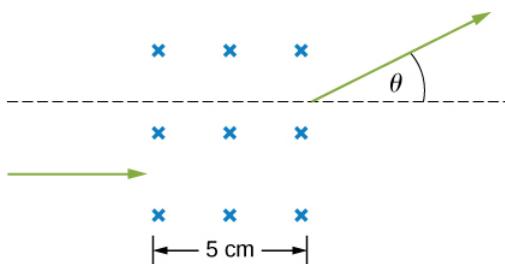
(c) Which premise is responsible?

Challenge Problems

102. A particle of charge $+q$ and mass m moves with velocity \hat{v}_0 pointed in the $+y$ -direction as it crosses the x -axis at $x = R$ at a particular time. There is a negative charge $-Q$ fixed at the origin, and there exists a uniform magnetic field \hat{B}_0 pointed in the $+z$ -direction. It is found that the particle describes a circle of radius R about $-Q$. Find \hat{B}_0 in terms of the given quantities.

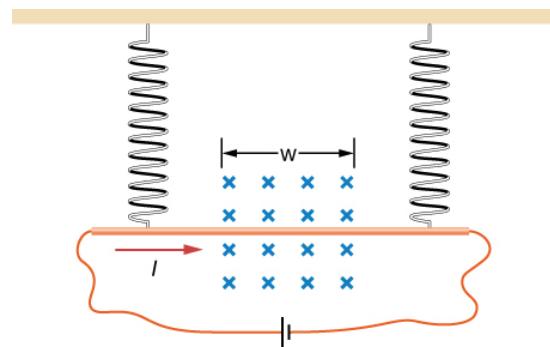
103. A proton of speed $v = 6 \times 10^6 \text{ m/s}$ enters a region of uniform magnetic field of $B = 0.5 \text{ T}$ at an angle of $q = 30^\circ$ to the magnetic field. In the region of magnetic field proton describes a helical path with radius R and pitch p (distance between loops). Find R and p .

104. A particle's path is bent when it passes through a region of non-zero magnetic field although its speed remains unchanged. This is very useful for "beam steering" in particle accelerators. Consider a proton of speed $4 \times 10^6 \text{ m/s}$ entering a region of uniform magnetic field 0.2 T over a 5-cm-wide region. Magnetic field is perpendicular to the velocity of the particle. By how much angle will the path of the proton be bent? (Hint: The particle comes out tangent to a circle.)

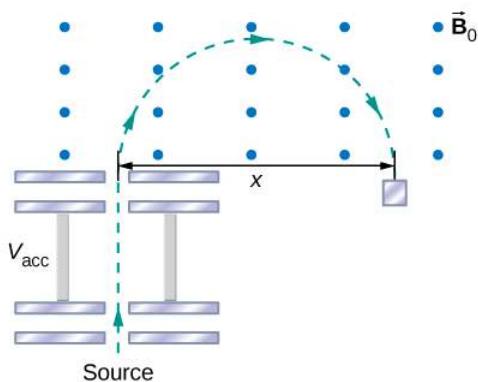


105. In a region a non-uniform magnetic field exists such that $B_x = 0$, $B_y = 0$, and $B_z = ax$, where a is a constant. At some time t , a wire of length L is carrying a current I is located along the x -axis from origin to $x = L$. Find the magnetic force on the wire at this instant in time.

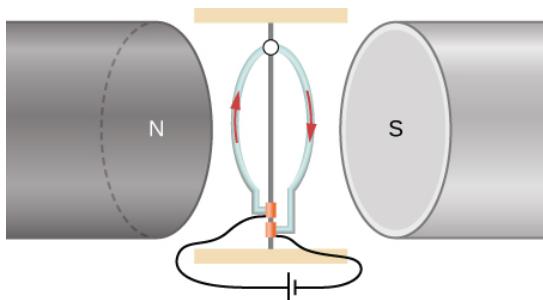
106. A copper rod of mass m and length L is hung from the ceiling using two springs of spring constant k . A uniform magnetic field of magnitude B_0 pointing perpendicular to the rod and spring (coming out of the page in the figure) exists in a region of space covering a length w of the copper rod. The ends of the rod are then connected by flexible copper wire across the terminals of a battery of voltage V . Determine the change in the length of the springs when a current I runs through the copper rod in the direction shown in figure. (Ignore any force by the flexible wire.)



107. The accompanied figure shows an arrangement for measuring mass of ions by an instrument called the mass spectrometer. An ion of mass m and charge $+q$ is produced essentially at rest in source S , a chamber in which a gas discharge is taking place. The ion is accelerated by a potential difference V_{acc} and allowed to enter a region of constant magnetic field \vec{B}_0 . In the uniform magnetic field region, the ion moves in a semicircular path striking a photographic plate at a distance x from the entry point. Derive a formula for mass m in terms of B_0 , q , V_{acc} , and x .



- 108.** A wire is made into a circular shape of radius R and pivoted along a central support. The two ends of the wire are touching a brush that is connected to a dc power source. The structure is between the poles of a magnet such that we can assume there is a uniform magnetic field on the wire. In terms of a coordinate system with origin at the center of the ring, magnetic field is $B_x = B_0, B_y = B_z = 0$, and the ring rotates about the **z-axis**. Find the torque on the ring when it is not in the xz-plane.



- 109.** A long-rigid wire lies along the **x-axis** and carries a current of **2.5 A** in the positive x-direction. Around the wire is the magnetic field $\vec{B} = 2.0\hat{i} + 5.0x^2\hat{j}$, with x in meters and \mathbf{B} in millitesla. Calculate the magnetic force on the segment of wire between $x = 2.0 \text{ m}$ and $x = 4.0 \text{ m}$.

- 110.** A circular loop of wire of area 10 cm^2 carries a current of **25 A**. At a particular instant, the loop lies in the **xy-plane** and is subjected to a magnetic field $\vec{B} = (2.0\hat{i} + 6.0\hat{j} + 8.0\hat{k}) \times 10^{-3}\text{T}$. As viewed from above the **xy-plane**, the current is circulating clockwise.

- (a) What is the magnetic dipole moment of the current loop?
- (b) At this instant, what is the magnetic torque on the loop?

This page titled [18.15: Magnetic Forces and Fields \(Exercise\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [11.10: Magnetic Forces and Fields \(Exercise\)](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.16: Sources of Magnetic Fields (Exercise)

Conceptual Questions

12.2 The Biot-Savart Law

1. For calculating magnetic fields, what are the advantages and disadvantages of the Biot-Savart law?
2. Describe the magnetic field due to the current in two wires connected to the two terminals of a source of emf and twisted tightly around each other.
3. How can you decide if a wire is infinite?
4. Identical currents are carried in two circular loops; however, one loop has twice the diameter as the other loop. Compare the magnetic fields created by the loops at the center of each loop.

12.3 Magnetic Field Due to a Thin Straight Wire

5. How would you orient two long, straight, current-carrying wires so that there is no net magnetic force between them?
(Hint: What orientation would lead to one wire not experiencing a magnetic field from the other?)

12.4 Magnetic Force between Two Parallel Currents

6. Compare and contrast the electric field of an infinite line of charge and the magnetic field of an infinite line of current.
7. Is \vec{B} constant in magnitude for points that lie on a magnetic field line?

12.5 Magnetic Field of a Current Loop

8. Is the magnetic field of a current loop uniform?
9. What happens to the length of a suspended spring when a current passes through it?
10. Two concentric circular wires with different diameters carry currents in the same direction. Describe the force on the inner wire.

12.6 Ampère's Law

11. Is Ampère's law valid for all closed paths? Why isn't it normally useful for calculating a magnetic field?

12.7 Solenoids and Toroids

12. Is the magnetic field inside a toroid completely uniform? Almost uniform?
13. Explain why $\vec{B} = \mathbf{0}$ inside a long, hollow copper pipe that is carrying an electric current parallel to the axis. Is $\vec{B} = \mathbf{0}$ outside the pipe?

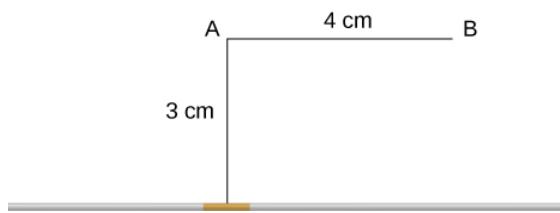
12.8 Magnetism in Matter

14. A diamagnetic material is brought close to a permanent magnet. What happens to the material?
15. If you cut a bar magnet into two pieces, will you end up with one magnet with an isolated north pole and another magnet with an isolated south pole? Explain your answer.

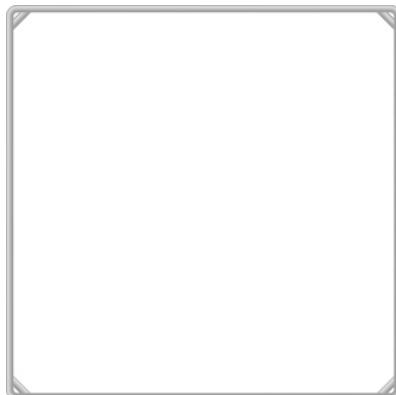
Problems

12.2 The Biot-Savart Law

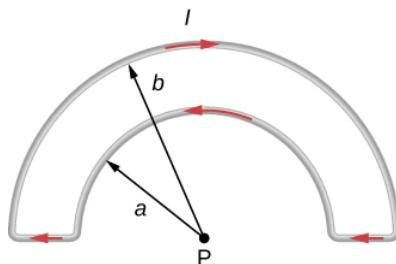
16. A 10-A current flows through the wire shown. What is the magnitude of the magnetic field due to a 0.5-mm segment of wire as measured at (a) point A and (b) point B?



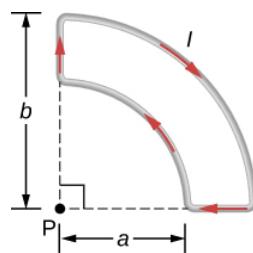
- 17.** Ten amps flow through a square loop where each side is 20 cm in length. At each corner of the loop is a 0.01-cm segment that connects the longer wires as shown. Calculate the magnitude of the magnetic field at the center of the loop.



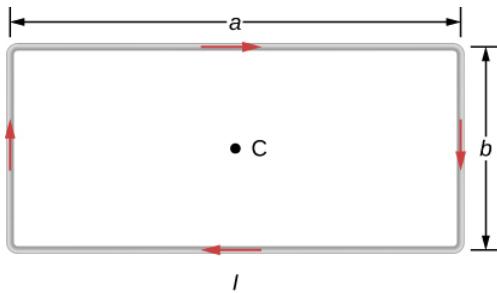
- 18.** What is the magnetic field at P due to the current I in the wire shown?



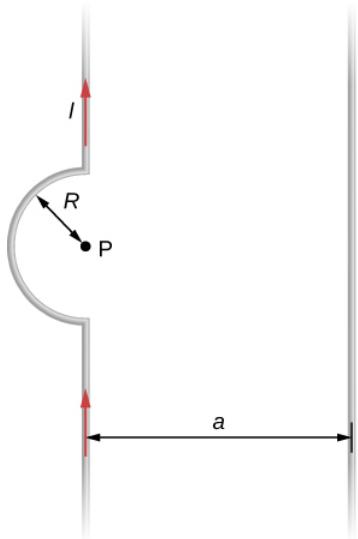
- 19.** The accompanying figure shows a current loop consisting of two concentric circular arcs and two perpendicular radial lines. Determine the magnetic field at point P.



- 20.** Find the magnetic field at the center C of the rectangular loop of wire shown in the accompanying figure.

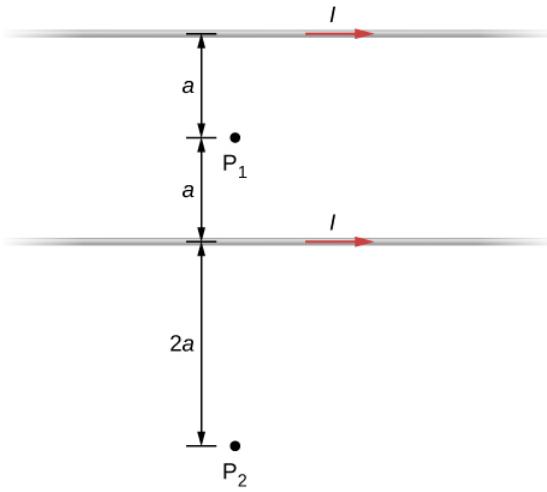


- 21.** Two long wires, one of which has a semicircular bend of radius R , are positioned as shown in the accompanying figure. If both wires carry a current I , how far apart must their parallel sections be so that the net magnetic field at P is zero? Does the current in the straight wire flow up or down?



12.3 Magnetic Field Due to a Thin Straight Wire

- 22.** A typical current in a lightning bolt is 10^4 A. Estimate the magnetic field 1 m from the bolt.
- 23.** The magnitude of the magnetic field 50 cm from a long, thin, straight wire is $8.0 \mu\text{T}$. What is the current through the long wire?
- 24.** A transmission line strung 7.0 m above the ground carries a current of 500 A. What is the magnetic field on the ground directly below the wire? Compare your answer with the magnetic field of Earth.
- 25.** A long, straight, horizontal wire carries a left-to-right current of 20 A. If the wire is placed in a uniform magnetic field of magnitude $4.0 \times 10^{-5} \text{T}$ that is directed vertically downward, what is the resultant magnitude of the magnetic field 20 cm above the wire? 20 cm below the wire?
- 26.** The two long, parallel wires shown in the accompanying figure carry currents in the same direction. If $I_1 = 10\text{A}$ and $I_2 = 20\text{A}$, what is the magnetic field at point P ?
- 27.** The accompanying figure shows two long, straight, horizontal wires that are parallel and a distance $2a$ apart. If both wires carry current I in the same direction, (a) what is the magnetic field at P_1 ? (b) P_2 ?

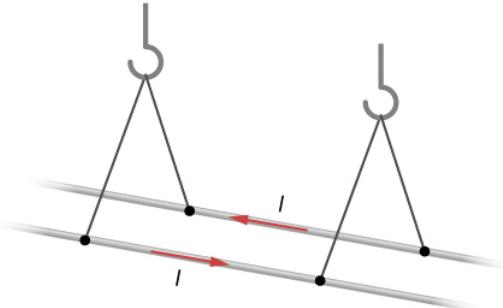


- 28.** Repeat the calculations of the preceding problem with the direction of the current in the lower wire reversed.

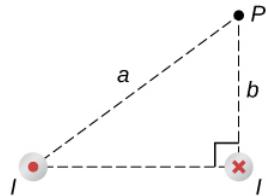
- 29.** Consider the area between the wires of the preceding problem. At what distance from the top wire is the net magnetic field a minimum? Assume that the currents are equal and flow in opposite directions.

12.4 Magnetic Force between Two Parallel Currents

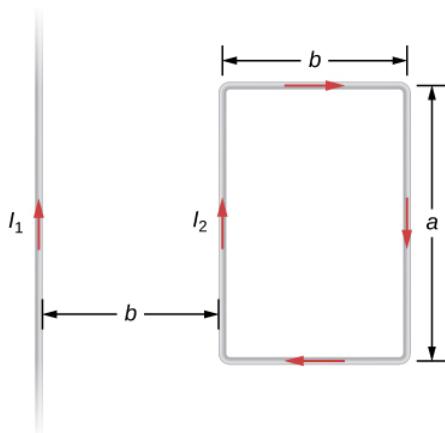
- 30.** Two long, straight wires are parallel and 25 cm apart.
- If each wire carries a current of 50 A in the same direction, what is the magnetic force per meter exerted on each wire?
 - Does the force pull the wires together or push them apart?
 - What happens if the currents flow in opposite directions?
- 31.** Two long, straight wires are parallel and 10 cm apart. One carries a current of 2.0 A, the other a current of 5.0 A.
- If the two currents flow in opposite directions, what is the magnitude and direction of the force per unit length of one wire on the other?
 - What is the magnitude and direction of the force per unit length if the currents flow in the same direction?
- 32.** Two long, parallel wires are hung by cords of length 5.0 cm, as shown in the accompanying figure. Each wire has a mass per unit length of 30 g/m, and they carry the same current in opposite directions. What is the current if the cords hang at 6.0° with respect to the vertical?



- 33.** A circuit with current **I** has two long parallel wire sections that carry current in opposite directions. Find magnetic field at a point **P** near these wires that is a distance **a** from one wire and **b** from the other wire as shown in the figure.



- 34.** The infinite, straight wire shown in the accompanying figure carries a current I_1 . The rectangular loop, whose long sides are parallel to the wire, carries a current I_2 . What are the magnitude and direction of the force on the rectangular loop due to the magnetic field of the wire?

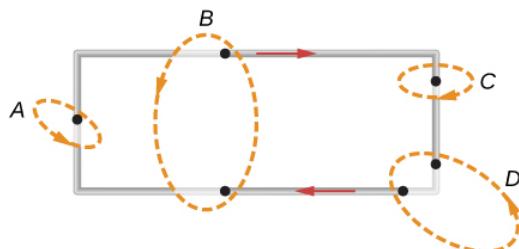


12.5 Magnetic Field of a Current Loop

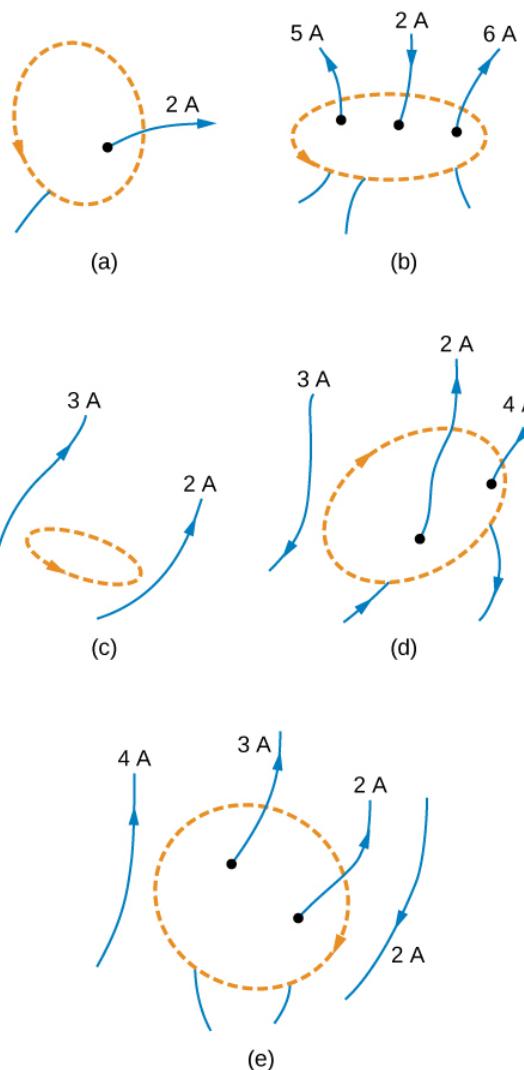
35. When the current through a circular loop is 6.0 A, the magnetic field at its center is $2.0 \times 10^{-4} T$. What is the radius of the loop?
36. How many turns must be wound on a flat, circular coil of radius 20 cm in order to produce a magnetic field of magnitude $4.0 \times 10^{-5} T$ at the center of the coil when the current through it is 0.85 A?
37. A flat, circular loop has 20 turns. The radius of the loop is 10.0 cm and the current through the wire is 0.50 A. Determine the magnitude of the magnetic field at the center of the loop.
38. A circular loop of radius R carries a current I . At what distance along the axis of the loop is the magnetic field one-half its value at the center of the loop?
39. Two flat, circular coils, each with a radius R and wound with N turns, are mounted along the same axis so that they are parallel a distance d apart. What is the magnetic field at the midpoint of the common axis if a current I flows in the same direction through each coil?
40. For the coils in the preceding problem, what is the magnetic field at the center of either coil?

12.6 Ampère's Law

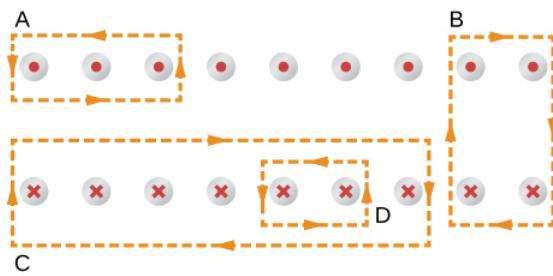
41. A current I flows around the rectangular loop shown in the accompanying figure. Evaluate $\oint \vec{B} \cdot d\vec{l}$ for the paths **A**, **B**, **C**, and **D**.



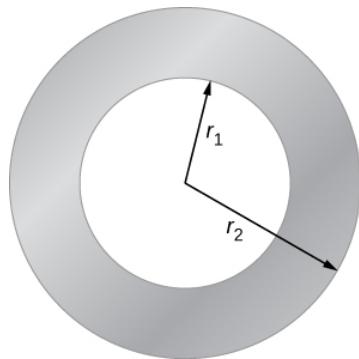
42. Evaluate $\oint \vec{B} \cdot d\vec{l}$ for each of the cases shown in the accompanying figure.



43. The coil whose lengthwise cross section is shown in the accompanying figure carries a current I and has N evenly spaced turns distributed along the length l . Evaluate $\oint \vec{B} \cdot d\vec{l}$ for the paths indicated.

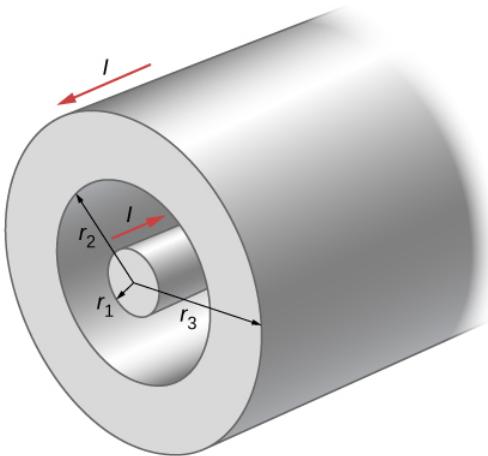


44. A superconducting wire of diameter 0.25 cm carries a current of 1000 A. What is the magnetic field just outside the wire?
45. A long, straight wire of radius R carries a current I that is distributed uniformly over the cross-section of the wire. At what distance from the axis of the wire is the magnitude of the magnetic field a maximum?
46. The accompanying figure shows a cross-section of a long, hollow, cylindrical conductor of inner radius $r_1 = 3.0\text{cm}$ and outer radius $r_2 = 5.0\text{cm}$. A 50-A current distributed uniformly over the cross-section flows into the page. Calculate the magnetic field at $r = 2.0\text{cm}$, $r = 4.0\text{cm}$, and $r = 6.0\text{cm}$.



- 47.** A long, solid, cylindrical conductor of radius 3.0 cm carries a current of 50 A distributed uniformly over its cross-section. Plot the magnetic field as a function of the radial distance r from the center of the conductor.

- 48.** A portion of a long, cylindrical **coaxial cable** is shown in the accompanying figure. A current I flows down the center conductor, and this current is returned in the outer conductor. Determine the magnetic field in the regions (a) $r \leq r_1$, (b) $r_2 \geq r \geq r_1$, (c) $r_3 \geq r \geq r_2$, and (d) $r \geq r_3$. Assume that the current is distributed uniformly over the cross sections of the two parts of the cable.



12.7 Solenoids and Toroids

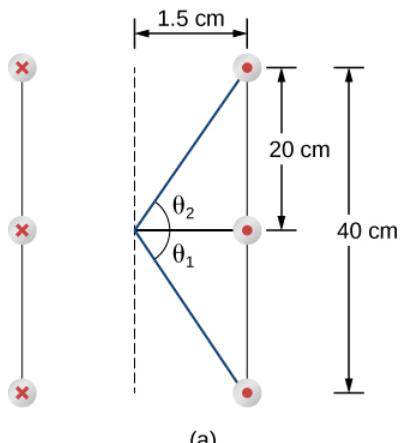
- 49.** A solenoid is wound with 2000 turns per meter. When the current is 5.2 A, what is the magnetic field within the solenoid?

- 50.** A solenoid has 12 turns per centimeter. What current will produce a magnetic field of $2.0 \times 10^{-2} T$ within the solenoid?

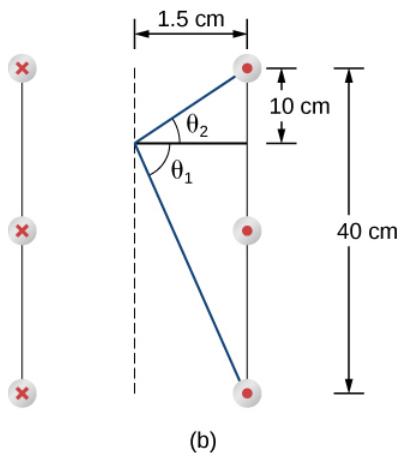
- 51.** If a current is 2.0 A, how many turns per centimeter must be wound on a solenoid in order to produce a magnetic field of $2.0 \times 10^{-3} T$ within it?

- 52.** A solenoid is 40 cm long, has a diameter of 3.0 cm, and is wound with 500 turns. If the current through the windings is 4.0 A, what is the magnetic field at a point on the axis of the solenoid that is

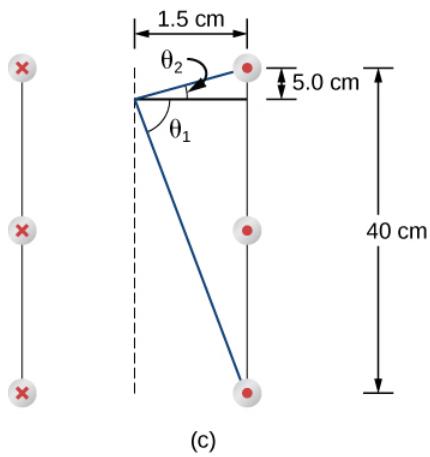
- (a) at the center of the solenoid,
- (b) 10.0 cm from one end of the solenoid, and
- (c) 5.0 cm from one end of the solenoid?
- (d) Compare these answers with the infinite-solenoid case.



(a)



(b)



(c)

53. Determine the magnetic field on the central axis at the opening of a semi-infinite solenoid. (That is, take the opening to be at $x=0$ and the other end to be at $x = \infty$)

54. By how much is the approximation $B = \mu_0 n I$ in error at the center of a solenoid that is 15.0 cm long, has a diameter of 4.0 cm, is wrapped with n turns per meter, and carries a current I ?

55. A solenoid with 25 turns per centimeter carries a current I . An electron moves within the solenoid in a circle that has a radius of 2.0 cm and is perpendicular to the axis of the solenoid. If the speed of the electron is $2.0 \times 10^5 \text{ m/s}$, what is I ?

56. A toroid has 250 turns of wire and carries a current of 20 A. Its inner and outer radii are 8.0 and 9.0 cm. What are the values of its magnetic field at $r=8.1, 8.5$, and 8.9 cm ?

- 57.** A toroid with a square cross section $3.0\text{ cm} \times 3.0\text{ cm}$ has an inner radius of 25.0 cm . It is wound with 500 turns of wire, and it carries a current of 2.0 A . What is the strength of the magnetic field at the center of the square cross section?

12.8 Magnetism in Matter

- 58.** The magnetic field in the core of an air-filled solenoid is 1.50 T . By how much will this magnetic field decrease if the air is pumped out of the core while the current is held constant?

- 59.** A solenoid has a ferromagnetic core, $n = 1000$ turns per meter, and $I = 5.0\text{ A}$. If B inside the solenoid is 2.0 T , what is χ for the core material?

- 60.** A 20-A current flows through a solenoid with 2000 turns per meter. What is the magnetic field inside the solenoid if its core is (a) a vacuum and (b) filled with liquid oxygen at 90 K ?

- 61.** The magnetic dipole moment of the iron atom is about $2.1 \times 10^{-23}\text{ A} \cdot \text{m}^2$.

(a) Calculate the maximum magnetic dipole moment of a domain consisting of 10^{19} iron atoms.

(b) What current would have to flow through a single circular loop of wire of diameter 1.0 cm to produce this magnetic dipole moment?

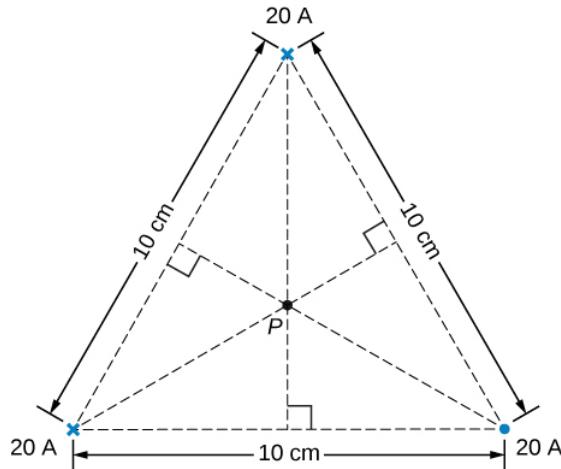
- 62.** Suppose you wish to produce a 1.2-T magnetic field in a toroid with an iron core for which $\chi = 4.0 \times 10^3$. The toroid has a mean radius of 15 cm and is wound with 500 turns. What current is required?

- 63.** A current of 1.5 A flows through the windings of a large, thin toroid with 200 turns per meter and a radius of 1 meter . If the toroid is filled with iron for which $\chi = 3.0 \times 10^3$, what is the magnetic field within it?

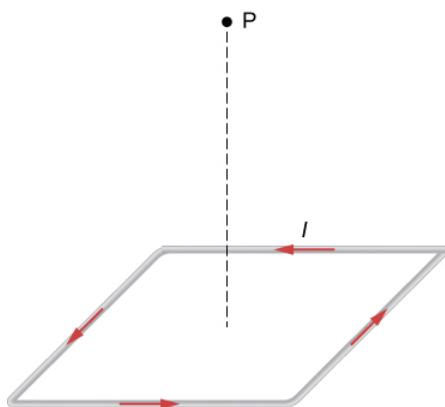
- 64.** A solenoid with an iron core is 25 cm long and is wrapped with 100 turns of wire. When the current through the solenoid is 10 A , the magnetic field inside it is 2.0 T . For this current, what is the permeability of the iron? If the current is turned off and then restored to 10 A , will the magnetic field necessarily return to 2.0 T ?

Additional Problems

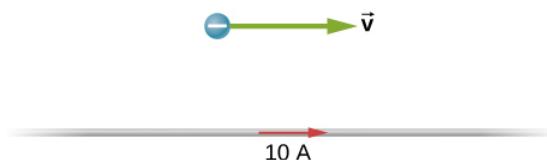
- 65.** Three long, straight, parallel wires, all carrying 20 A , are positioned as shown in the accompanying figure. What is the magnitude of the magnetic field at the point **P**?



- 66.** A current I flows around a wire bent into the shape of a square of side a . What is the magnetic field at the point **P** that is a distance z above the center of the square (see the accompanying figure)?

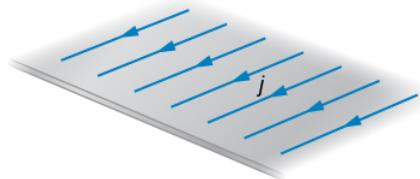


67. The accompanying figure shows a long, straight wire carrying a current of 10 A. What is the magnetic force on an electron at the instant it is 20 cm from the wire, traveling parallel to the wire with a speed of $2.0 \times 10^5 \text{ m/s}$? Describe qualitatively the subsequent motion of the electron.



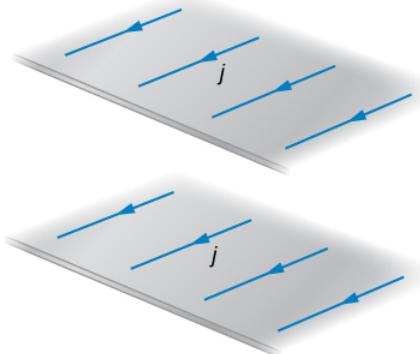
68. Current flows along a thin, infinite sheet as shown in the accompanying figure. The current per unit length along the sheet is J in amperes per meter.

- (a) Use the Biot-Savart law to show that $B = \mu_0 J / 2$ on either side of the sheet. What is the direction of \vec{B} on each side?
- (b) Now use Ampère's law to calculate the field.

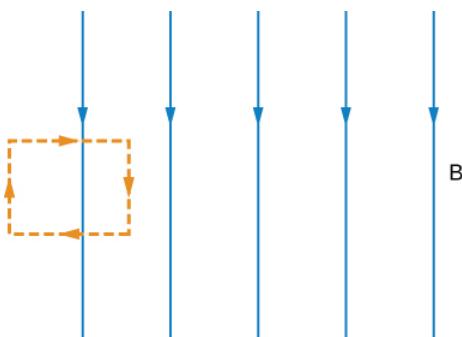


69. (a) Use the result of the previous problem to calculate the magnetic field between, above, and below the pair of infinite sheets shown in the accompanying figure.

- (b) Repeat your calculations if the direction of the current in the lower sheet is reversed.



70. We often assume that the magnetic field is uniform in a region and zero everywhere else. Show that in reality it is impossible for a magnetic field to drop abruptly to zero, as illustrated in the accompanying figure. (**Hint:** Apply Ampère's law over the path shown.)



71. How is the fractional change in the strength of the magnetic field across the face of the toroid related to the fractional change in the radial distance from the axis of the toroid?

72. Show that the expression for the magnetic field of a toroid reduces to that for the field of an infinite solenoid in the limit that the central radius goes to infinity.

73. A toroid with an inner radius of 20 cm and an outer radius of 22 cm is tightly wound with one layer of wire that has a diameter of 0.25 mm.

(a) How many turns are there on the toroid?

(b) If the current through the toroid windings is 2.0 A, what is the strength of the magnetic field at the center of the toroid?

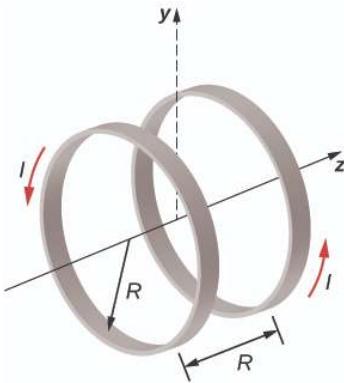
74. A wire element has $\text{vec}dl$, $I\vec{dl} = JAdl = Jdv$, where A and dv are the cross-sectional area and volume of the element, respectively. Use this, the Biot-Savart law, and $\mathbf{J} = nev$ to show that the magnetic field of a moving point charge q is given by:

$$\vec{B} = \frac{\mu_0}{4\pi} \frac{qv \times \hat{r}}{r^2}$$

75. A reasonably uniform magnetic field over a limited region of space can be produced with the Helmholtz coil, which consists of two parallel coils centered on the same axis. The coils are connected so that they carry the same current I. Each coil has N turns and radius R, which is also the distance between the coils.

(a) Find the magnetic field at any point on the z-axis shown in the accompanying figure.

(b) Show that dB/dz and d^2B/dz^2 are both zero at $z = 0$. (These vanishing derivatives demonstrate that the magnetic field varies only slightly near $z = 0$.)



76. A charge of $4.0\mu\text{C}$ is distributed uniformly around a thin ring of insulating material. The ring has a radius of 0.20 m and rotates at $2.0 \times 10^4 \text{ rev/min}$ around the axis that passes through its center and is perpendicular to the plane of the ring. What is the magnetic field at the center of the ring?

77. A thin, nonconducting disk of radius R is free to rotate around the axis that passes through its center and is perpendicular to the face of the disk. The disk is charged uniformly with a total charge q . If the disk rotates at a constant angular velocity

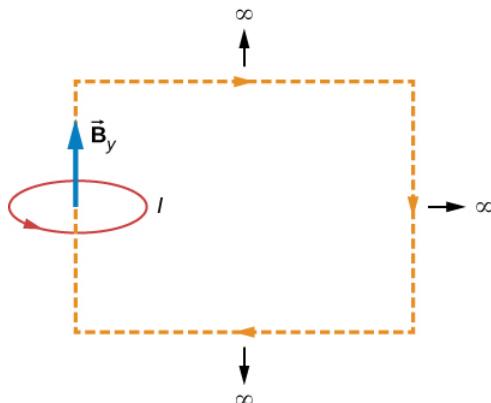
ω, what is the magnetic field at its center?

78. Consider the disk in the previous problem. Calculate the magnetic field at a point on its central axis that is a distance y above the disk.

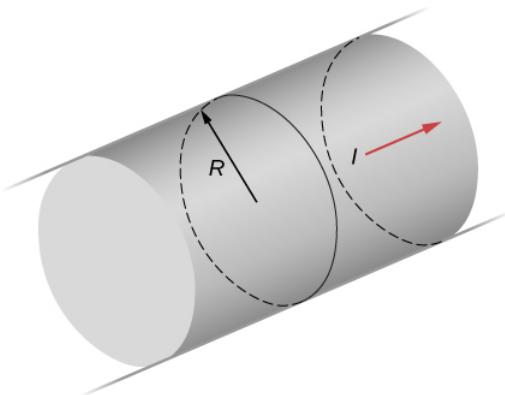
79. Consider the axial magnetic field $B_y = \mu_0 I R^2 / 2(y^2 + R^2)^{3/2}$ of the circular current loop shown below.

(a) Evaluate $\int_{-a}^a B_y dy$. Also show that $\lim_{a \rightarrow \infty} \int_{-a}^a B_y dy = \mu_0 I$.

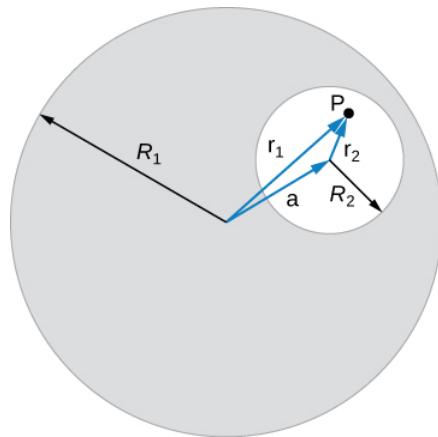
(b) Can you deduce this limit without evaluating the integral? (Hint: See the accompanying figure.)



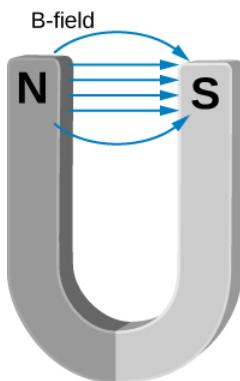
80. The current density in the long, cylindrical wire shown in the accompanying figure varies with distance r from the center of the wire according to $J = cr$, where c is a constant. (a) What is the current through the wire? (b) What is the magnetic field produced by this current for $r \leq R$? For $r \geq R$?



81. A long, straight, cylindrical conductor contains a cylindrical cavity whose axis is displaced by a from the axis of the conductor, as shown in the accompanying figure. The current density in the conductor is given by $\hat{J} = J_0 \hat{k}$, where J_0 is a constant and \hat{k} is along the axis of the conductor. Calculate the magnetic field at an arbitrary point P in the cavity by superimposing the field of a solid cylindrical conductor with radius R_1 and current density \vec{J} onto the field of a solid cylindrical conductor with radius R_2 and current density $-\vec{J}$. Then use the fact that the appropriate azimuthal unit vectors can be expressed as $\hat{\theta}_1 = \hat{k} \times \hat{r}_1$ and $\hat{\theta}_2 = \hat{k} \times \hat{r}_2$ to show that everywhere inside the cavity the magnetic field is given by the constant $\vec{B} = \frac{1}{2} \mu_0 J_0 \hat{k} \times \hat{a}$, where $\hat{a} = \hat{r}_1 - \hat{r}_2$ and $\hat{r}_1 = r_1 \hat{r}_1$ is the position of P relative to the center of the conductor and $\hat{r}_2 = r_2 \hat{r}_2$ is the position of P relative to the center of the cavity.

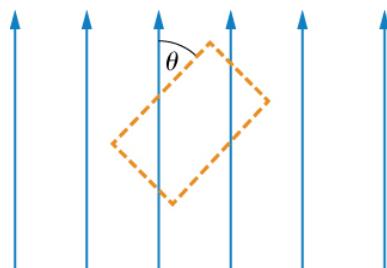


- 82.** Between the two ends of a horseshoe magnet the field is uniform as shown in the diagram. As you move out to outside edges, the field bends. Show by Ampère's law that the field must bend and thereby the field weakens due to these bends.



- 83.** Show that the magnetic field of a thin wire and that of a current loop are zero if you are infinitely far away.

- 84.** An Ampère loop is chosen as shown by dashed lines for a parallel constant magnetic field as shown by solid arrows. Calculate $\oint \vec{B} \cdot d\vec{l}$ for each side of the loop then find the entire $\oint \vec{B} \cdot d\vec{l}$. Can you think of an Ampère loop that would make the problem easier? Do those results match these?



- 85.** A very long, thick cylindrical wire of radius R carries a current density J that varies across its cross-section. The magnitude of the current density at a point a distance r from the center of the wire is given by $J = J_0 \frac{r}{R}$, where J_0 is a constant. Find the magnetic field

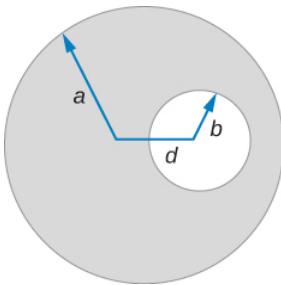
- (a) at a point outside the wire and
- (b) at a point inside the wire. Write your answer in terms of the net current I through the wire.

- 86.** A very long, cylindrical wire of radius a has a circular hole of radius b in it at a distance d from the center. The wire carries a uniform current of magnitude I through it. The direction of the current in the figure is out of the paper. Find the magnetic field

- (a) at a point at the edge of the hole closest to the center of the thick wire,

(b) at an arbitrary point inside the hole, and

(c) at an arbitrary point outside the wire. (Hint: Think of the hole as a sum of two wires carrying current in the opposite directions.)

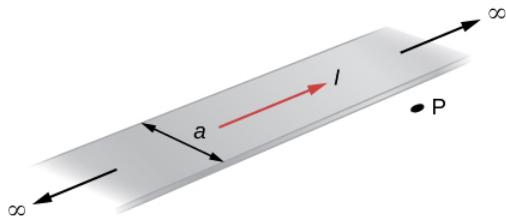


87. Magnetic field inside a torus. Consider a torus of rectangular cross-section with inner radius a and outer radius b . N turns of an insulated thin wire are wound evenly on the torus tightly all around the torus and connected to a battery producing a steady current I in the wire. Assume that the current on the top and bottom surfaces in the figure is radial, and the current on the inner and outer radii surfaces is vertical. Find the magnetic field inside the torus as a function of radial distance \mathbf{r} from the axis.

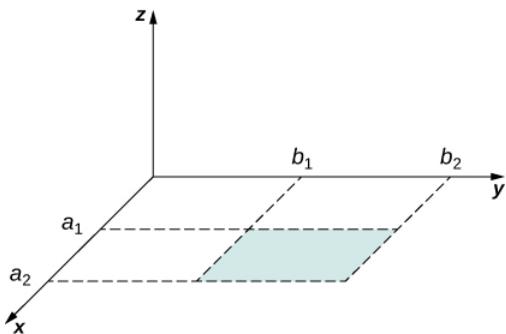
88. Two long coaxial copper tubes, each of length L , are connected to a battery of voltage V . The inner tube has inner radius \mathbf{a} and outer radius \mathbf{b} , and the outer tube has inner radius c and outer radius \mathbf{d} . The tubes are then disconnected from the battery and rotated in the same direction at angular speed of ω radians per second about their common axis. Find the magnetic field (a) at a point inside the space enclosed by the inner tube $\mathbf{r} < \mathbf{a}$, and (b) at a point between the tubes $\mathbf{b} < \mathbf{r} < \mathbf{c}$, and (c) at a point outside the tubes $\mathbf{r} > \mathbf{d}$. (Hint: Think of copper tubes as a capacitor and find the charge density based on the voltage applied, $\mathbf{Q} = \mathbf{V}\mathbf{C}$, $\mathbf{C} = \frac{1}{2\pi\epsilon_0 L} \ln(c/b)$.)

Challenge Problems

89. The accompanying figure shows a flat, infinitely long sheet of width a that carries a current I uniformly distributed across it. Find the magnetic field at the point P , which is in the plane of the sheet and at a distance x from one edge. Test your result for the limit $a \rightarrow 0$.



90. A hypothetical current flowing in the z -direction creates the field $\vec{B} = C[(x/y^2)\hat{i} + (1/y)\hat{j}]$ in the rectangular region of the xy -plane shown in the accompanying figure. Use Ampère's law to find the current through the rectangle.



91. A nonconducting hard rubber circular disk of radius R is painted with a uniform surface charge density σ . It is rotated about its axis with angular speed ω . (a) Find the magnetic field produced at a point on the axis a distance h meters from the center of the disk. (b) Find the numerical value of magnitude of the magnetic field when $\sigma = 1C/m^2$, $R = 20cm$, $h = 2cm$, and $\omega = 400rad/sec$, and compare it with the magnitude of magnetic field of Earth, which is about 1/2 Gauss.

This page titled [18.16: Sources of Magnetic Fields \(Exercise\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.10: Sources of Magnetic Fields \(Exercise\)](#) by OpenStax is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

18.17: Magnetic Forces and Fields (Answers)

Check Your Understanding

11.1. a. 0 N;

b. $2.4 \times 10^{-14} \hat{k}N$;

c. $2.4 \times 10^{-14} \hat{j}N$;

d. $(7.2\hat{j} + 2.2\hat{k}) \times 10^{-15} N$

11.2. a. $9.6 \times 10^{-12} N$ toward the south;

b. $\frac{w}{Fm} = 1.7 \times 10^{-15}$

11.3. a. bends upward;

b. bends downward

11.4. a. aligned or anti-aligned;

b. perpendicular

11.5. a. 1.1 T;

b. 1.6 T

11.6. 0.32 m

Conceptual Questions

1. Both are field dependent. Electrical force is dependent on charge, whereas magnetic force is dependent on current or rate of charge flow.

3. The magnitude of the proton and electron magnetic forces are the same since they have the same amount of charge. The direction of these forces however are opposite of each other. The accelerations are opposite in direction and the electron has a larger acceleration than the proton due to its smaller mass.

5. The magnetic field must point parallel or anti-parallel to the velocity.

7. A compass points toward the north pole of an electromagnet.

9. Velocity and magnetic field can be set together in any direction. If there is a force, the velocity is perpendicular to it. The magnetic field is also perpendicular to the force if it exists.

11. A force on a wire is exerted by an external magnetic field created by a wire or another magnet.

13. Poor conductors have a lower charge carrier density, n , which, based on the Hall effect formula, relates to a higher Hall potential. Good conductors have a higher charge carrier density, thereby a lower Hall potential.

Problems

15. a. left;

b. into the page;

c. up the page;

d. no force;

e. right;

f. down

17. a. right;

b. into the page;

c. down

19. a. into the page;

b. left;

c. out of the page

21. a. $2.64 \times 10^{-8} N$;

b. The force is very small, so this implies that the effect of static charges on airplanes is negligible.

23. $10.1^\circ; 169.9^\circ$

25. 4.27 m

27. a. $4.80 \times 10^{-19} C$;

b. 3;

c. This ratio must be an integer because charges must be integer numbers of the basic charge of an electron. There are no free charges with values less than this basic charge, and all charges are integer multiples of this basic charge.

29. a. $4.09 \times 10^3 m/s$;

b. $7.83 \times 10^3 m$;

c. $1.75 \times 10^5 m/s$, then, $1.83 \times 10^2 m$;

d. $4.27 m$

31. a. $1.8 \times 10^7 m/s$;

b. $6.8 \times 10^6 eV$;

c. $3.4 \times 10^6 V$

33. a. left;

b. into the page;

c. up;

d. no force;

e. right;

f. down

35. a. into the page;

b. left;

c. out of the page

37. a. 2.50 N;

b. This means that the light-rail power lines must be attached in order not to be moved by the force caused by Earth's magnetic field.

39. a. $\tau = NIAB$, so τ decreases by 5.00% if B decreases by 5.00%;

b. 5.26% increase

41. 10.0 A

$$43. A \cdot m^2 \cdot T = A \cdot m^2 \cdot \frac{N}{A \cdot m} = N \cdot m$$

45. $3.48 \times 10^{-26} N \cdot m$

47. $0.666 N \cdot m$

49. $5.8 \times 10^{-7}V$

51. $4.8 \times 10^7 C/kg$

53. a. $4.4 \times 10^{-8}s$;

b. 0.21 m

55. a. $1.92 \times 10^{-12}J$;

b. 12 MeV;

c. 12 MV;

d. $5.2 \times 10^{-8}s$;

e. $1.92 \times 10^{-12}J$, 12 MeV, 12 V, $10.4 \times 10^{-8}s$

57. a. $2.50 \times 10^{-2}m$;

b. Yes, this distance between their paths is clearly big enough to separate the U-235 from the U-238, since it is a distance of 2.5 cm.

Additional Problems

59. $-7.2 \times 10^{-15}N\hat{j}$

61. $9.8 \times 10^{-5}\hat{T}$; the magnetic and gravitational forces must balance to maintain dynamic equilibrium

63. $1.13 \times 10^{-3}T$

65. $1.6\hat{i} - 1.4\hat{j} - 1.1\hat{k}) \times 10^5 V/m$

67. a. circular motion in a north, down plane;

b. $(1.61\hat{j} - 0.58\hat{k}) \times 10^{-14}N$

69. The proton has more mass than the electron; therefore, its radius and period will be larger.

71. $1.3 \times 10^{-25}kg$

73. 1:0.707:1

75. 1/4

77. a. $2.3 \times 10^{-4}m$;

b. $1.37 \times 10^{-4}T$

79. a. 30.0° ;

b. 4.80 N

81. a. 0.283 N;

b. 0.4 N;

c. 0 N;

d. 0 N

83. 0 N and 0.012 Nm

85. a. $0.31Am^2$;

b. 0.16 Nm

87. $0.024Am^2$

89. a. $0.16Am^2$;

b. 0.016 Nm;

c. 0.028 J

91. (Proof)

93. $4.65 \times 10^{-7}V$

95. Since $E = Blv$, where the width is twice the radius, $I = 2r, I = 2r, I = nqAv_d, v_d = \frac{I}{nqA} = \frac{I}{nq\pi r^2}$ so $E = B \times 2r \times \frac{I}{nq\pi r^2}$ $= \frac{2IB}{nq\pi r} \propto \frac{1}{r} \propto \frac{1}{d}$. The Hall voltage is inversely proportional to the diameter of the wire.

97. $6.92 \times 10^7 m/s$; 0.602 m

99. a. $2.4 \times 10^{-19}C$;

b. not an integer multiple of e;

c. need to assume all charges have multiples of e, could be other forces not accounted for

101. a. $B = 5 T$;

b. very large magnet;

c. applying such a large voltage

Challenge Problems

103. $R = (mv \sin \theta) / qB; p = (\frac{2\pi m}{eB}) v \cos \theta$

105. $IaL^2/2$

107. $m = \frac{qB_0^2}{8V_{acc}}x^2$

109. 0.01 N

This page titled [18.17: Magnetic Forces and Fields \(Answers\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [11.11: Magnetic Forces and Fields \(Answers\)](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

18.18: Sources of Magnetic Fields (Answers)

Check Your Understanding

12.1. 1.41 meters

12.2. $\frac{\mu_0 I}{2R}$

12.3. 4 amps flowing out of the page

12.4. Both have a force per unit length of $9.23 \times 10^{-12} N/m$

12.5. 0.608 meters

12.6. In these cases the integrals around the Ampèrian loop are very difficult because there is no symmetry, so this method would not be useful.

12.7. a. 1.00382;

b. 1.00015

12.8. a. $1.0 \times 10^{-4} T$;

b. 0.60 T;

c. 6.0×10^3

Conceptual Questions

1. Biot-Savart law's advantage is that it works with any magnetic field produced by a current loop. The disadvantage is that it can take a long time.

3. If you were to go to the start of a line segment and calculate the angle θ to be approximately 0° , the wire can be considered infinite. This judgment is based also on the precision you need in the result.

5. You would make sure the currents flow perpendicular to one another.

7. A magnetic field line gives the direction of the magnetic field at any point in space. The density of magnetic field lines indicates the strength of the magnetic field.

9. The spring reduces in length since each coil will have a north pole-produced magnetic field next to a south pole of the next coil.

11. Ampère's law is valid for all closed paths, but it is not useful for calculating fields when the magnetic field produced lacks symmetry that can be exploited by a suitable choice of path.

13. If there is no current inside the loop, there is no magnetic field (see Ampère's law). Outside the pipe, there may be an enclosed current through the copper pipe, so the magnetic field may not be zero outside the pipe.

15. The bar magnet will then become two magnets, each with their own north and south poles. There are no magnetic monopoles or single pole magnets.

Problems

17. $5.66 \times 10^{-5} T$

19. $B = \frac{\mu_0 I}{8} \left(\frac{1}{a} - \frac{1}{b} \right)$ out of the page

21. $a = \frac{2R}{\pi}$; the current in the wire to the right must flow up the page.

23. 20 A

25. Both answers have the magnitude of magnetic field of $4.5 \times 10^{-5} T$.

27. At P1, the net magnetic field is zero. At P2, $B = \frac{3\mu_0 I}{8\pi a}$ into the page.

29. The magnetic field is at a minimum at distance a from the top wire, or half-way between the wires.

31. a. $F/l = 8 \times 10^{-6}$ N/m away from the other wire;

b. $F/l = 8 \times 10^{-6}$ N/m toward the other wire

33. $B = \frac{\mu_0 I a}{2\pi b^2}$ into the page

35. 0.019 m

37. $6.28 \times 10^{-5} T$

39. $B = \frac{\mu_0 I R^2}{\left(\left(\frac{d}{2}\right)^2 + R^2\right)^{3/2}}$

41. a. $\mu_0 I$;

b. 0;

c. $\mu_0 I$;

d. 0

43. a. $3\mu_0 I$;

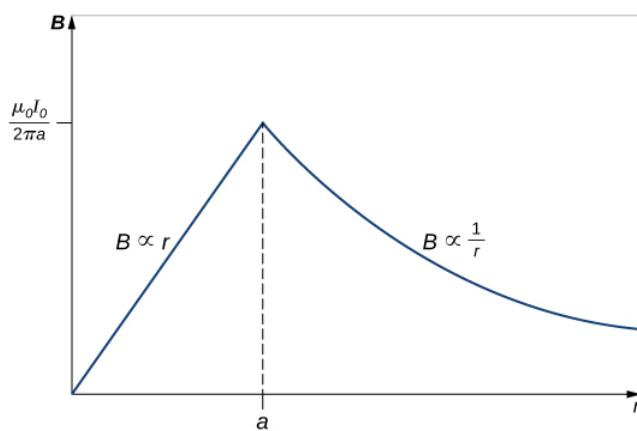
b. 0;

c. $7\mu_0 I$;

d. $-2\mu_0 I$

45. at the radius R

47.



49. $B = 1.3 \times 10^{-2} T$

51. roughly eight turns per cm

53. $B = \frac{1}{2}\mu_0 n I$

55. 0.0181 A

57. 0.0008 T

59. 317.31

61. $2.1 \times 10^{-4} A \cdot m^2$ 2.7 A

63. 0.18 T

Additional Problems

65. $B = 6.93 \times 10^{-5} T$

67. $3.2 \times 10^{-19} N$ in an arc away from the wire

69. a. above and below $B = \mu_0 j$, in the middle $B = 0$;

b. above and below $B = 0$, in the middle $B = \mu_0 j$

71. $\frac{dB}{B} = -\frac{dr}{r}$

73. a. 52778 turns;

b. 0.10 T

75. $B_1(x) = \frac{\mu_0 I R^2}{2(R^2 + z^2)^{3/2}}$

77. $B = \frac{\mu_0 \sigma \omega}{2} R$

79. derivation

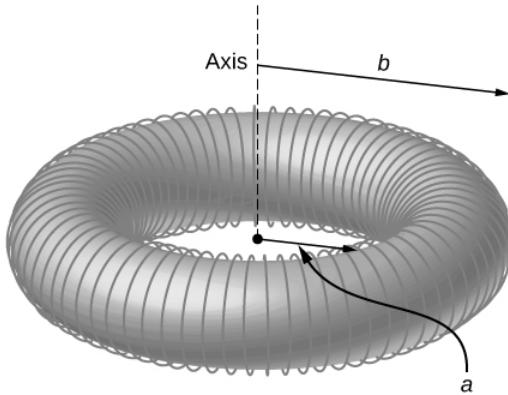
81. derivation

83. As the radial distance goes to infinity, the magnetic fields of each of these formulae go to zero.

85. a. $B = \frac{\mu_0 I}{2\pi r}$;

b. $B = \frac{\mu_0 J_0 r^2}{3R}$

87. $B(r) = \mu_0 NI / 2\pi r$



Challenge Problems

89. $B = \frac{\mu_0 I}{2\pi x}$.

91. a. $B = \frac{\mu_0 \sigma \omega}{2} [\frac{2h^2 + R^2}{\sqrt{R^2 + h^2}}]$;

b. $B = 4.09 \times 10^{-5} T$, 82% of Earth's magnetic field

This page titled [18.18: Sources of Magnetic Fields \(Answers\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [12.11: Sources of Magnetic Fields \(Answers\)](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

CHAPTER OVERVIEW

19: Alternating-Current (AC) Circuits

- [19.1: Introduction](#)
- [19.2: AC Sources](#)
- [19.3: Simple AC Circuits](#)
- [19.4: RLC Series Circuits with AC](#)
- [19.5: Power in an AC Circuit](#)
- [19.6: Resonance in an AC Circuit](#)
- [19.7: AC Safety - Grounding and Bonding](#)
- [19.8: Alternating-Current Circuits \(Summary\)](#)
- [19.9: Alternating-Current Circuits \(Exercise\)](#)
- [19.10: Alternating-Current Circuits \(Answers\)](#)

19: Alternating-Current (AC) Circuits is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

19.1: Introduction

Electric power is delivered to our homes by alternating current (ac) through high-voltage transmission lines. As explained in Transformers, transformers can then change the amplitude of the alternating potential difference to a more useful form. This lets us transmit power at very high voltages, minimizing resistive heating losses in the lines, and then furnish that power to homes at lower, safer voltages. Because constant potential differences are unaffected by transformers, this capability is more difficult to achieve with direct-current transmission.



Figure 19.1.1: The current we draw into our houses is an alternating current (ac). Power lines transmit ac to our neighborhoods, where local power stations and transformers distribute it to our homes. In this chapter, we discuss how a transformer works and how it allows us to transmit power at very high voltages and minimal heating losses across the lines.

In this chapter, we use Kirchhoff's laws to analyze four simple circuits in which ac flows. We have discussed the use of the resistor, capacitor, and inductor in circuits with batteries. These components are also part of ac circuits. However, because ac is required, the constant source of emf supplied by a battery is replaced by an ac voltage source, which produces an oscillating emf.

This page titled [19.1: Introduction](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [15.1: Prelude to Alternating-Current Circuits](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

19.2: AC Sources

Learning Objectives

By the end of the section, you will be able to:

- Explain the differences between direct current (dc) and alternating current (ac)
- Define characteristic features of alternating current and voltage, such as the amplitude or peak and the frequency

Most examples dealt with so far in this book, particularly those using batteries, have constant-voltage sources. Thus, once the current is established, it is constant. **Direct current (dc)** is the flow of electric charge in only one direction. It is the steady state of a constant-voltage circuit.

Most well-known applications, however, use a time-varying voltage source. **Alternating current (ac)** is the flow of electric charge that periodically reverses direction. An ac is produced by an alternating emf, which is generated in a power plant, as described in [Induced Electric Fields](#). If the ac source varies periodically, particularly sinusoidally, the circuit is known as an ac circuit. Examples include the commercial and residential power that serves so many of our needs.

The ac voltages and frequencies commonly used in businesses and homes vary around the world. In a typical house, the potential difference between the two sides of an electrical outlet alternates sinusoidally with a frequency of 60 or 50 Hz and an amplitude of 170 or 311 V, depending on whether you live in the United States or Europe, respectively. Most people know the potential difference for electrical outlets is 120 V or 220 V in the US or Europe, but as explained later in the chapter, these voltages are not the peak values given here but rather are related to the common voltages we see in our electrical outlets. Figure 19.2.1 shows graphs of voltage and current versus time for typical dc and ac power in the United States.

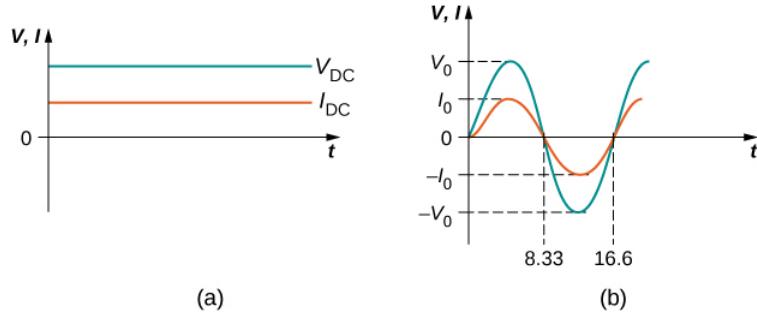


Figure 19.2.1: (a) The dc voltage and current are constant in time, once the current is established. (b) The voltage and current versus time are quite different for ac power. In this example, which shows 60-Hz ac power and time t in milliseconds, voltage and current are sinusoidal and are in phase for a simple resistance circuit. The frequencies and peak voltages of ac sources differ greatly.

Suppose we hook up a resistor to an ac voltage source and determine how the voltage and current vary in time across the resistor. Figure 19.2.2 shows a schematic of a simple circuit with an ac voltage source. The voltage fluctuates sinusoidally with time at a fixed frequency, as shown, on either the battery terminals or the resistor. Therefore, the **ac voltage**, or the “voltage at a plug,” can be given by

$$v(t) = V_0 \sin \omega t,$$

where

- v is the voltage at time t ,
- V_0 is the peak voltage, and
- ω is the angular frequency in radians per second.

For a typical house in the United States, $V_0 = 156\text{ V}$ and $\omega = 120\pi\text{ rad/s}$, whereas in Europe, $V_0 = 311\text{ V}$ and $\omega = 100\pi\text{ rad/s}$.

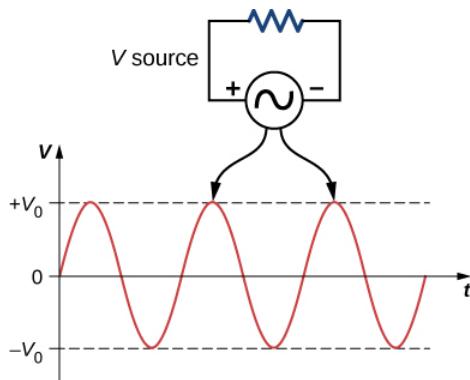


Figure 19.2.2: The potential difference \mathbf{V} between the terminals of an ac voltage source fluctuates, so the source and the resistor have ac sine waves on top of each other. The mathematical expression for \mathbf{v} is given by $\mathbf{v} = \mathbf{V}_0 \sin \omega t$.

For this simple resistance circuit, $I = V/R$, so the **ac current**, meaning the current that fluctuates sinusoidally with time at a fixed frequency, is

$$i(t) = I_0 \sin \omega t,$$

where

- $i(t)$ is the current at time t and
- I_0 is the peak current and is equal to V_0/R .

For this example, the voltage and current are said to be in phase, meaning that their sinusoidal functional forms have peaks, troughs, and nodes in the same place. They oscillate in sync with each other, as shown in Figure 19.2.1b. In these equations, and throughout this chapter, we use lowercase letters (such as i) to indicate instantaneous values and capital letters (such as I) to indicate maximum, or peak, values.

Current in the resistor alternates back and forth just like the driving voltage, since $I = V/R$. If the resistor is a fluorescent light bulb, for example, it brightens and dims 120 times per second as the current repeatedly goes through zero. A 120-Hz flicker is too rapid for your eyes to detect, but if you wave your hand back and forth between your face and a fluorescent light, you will see the stroboscopic effect of ac.

Exercise 19.2.1

If a European ac voltage source is considered, what is the time difference between the zero crossings on an ac voltage-versus-time graph?

Solution

10 ms

This page titled [19.2: AC Sources](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [15.2: AC Sources](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

19.3: Simple AC Circuits

Learning Objectives

By the end of the section, you will be able to:

- Interpret phasor diagrams and apply them to ac circuits with resistors, capacitors, and inductors
- Define the reactance for a resistor, capacitor, and inductor to help understand how current in the circuit behaves compared to each of these devices

In this section, we study simple models of ac voltage sources connected to three circuit components: (1) a resistor, (2) a capacitor, and (3) an inductor. The power furnished by an ac voltage source has an emf given by

$$v(t) = V_0 \sin \omega t,$$

as shown in Figure 19.3.1. This sine function assumes we start recording the voltage when it is $v = 0 \text{ V}$ at a time of $t = 0 \text{ s}$. A phase constant may be involved that shifts the function when we start measuring voltages, similar to the phase constant in the waves we studied in [Waves](#). However, because we are free to choose when we start examining the voltage, we can ignore this phase constant for now. We can measure this voltage across the circuit components using one of two methods: (1) a quantitative approach based on our knowledge of circuits, or (2) a graphical approach that is explained in the coming sections.

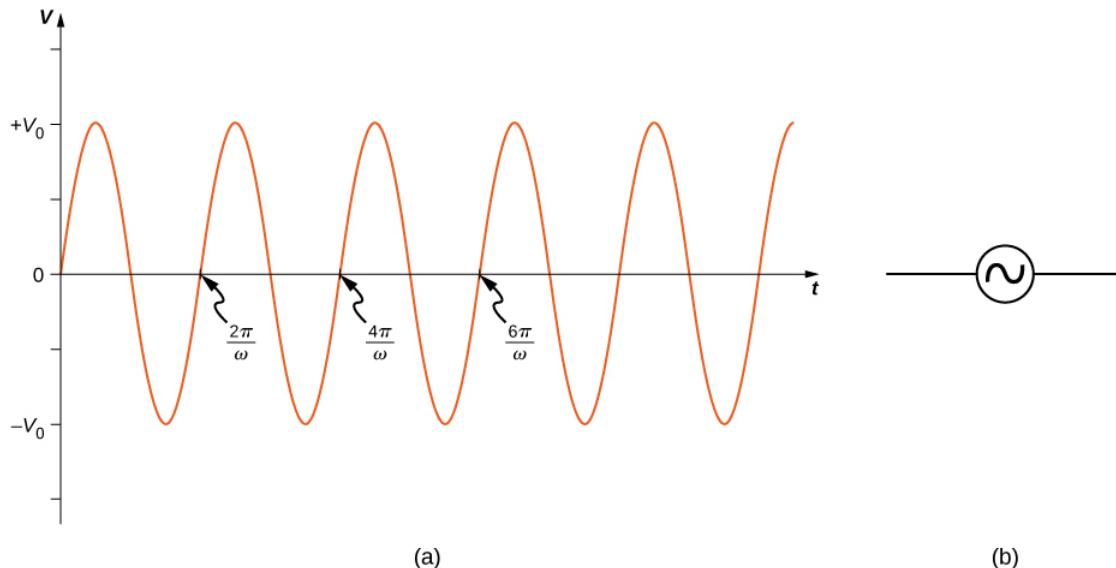


Figure 19.3.1: (a) The output $v(t) = V_0 \sin \omega t$ of an ac generator. (b) Symbol used to represent an ac voltage source in a circuit diagram.

Resistor

First, consider a **resistor** connected across an ac voltage source. From Kirchhoff's loop rule, the instantaneous voltage across the resistor of Figure 19.3.2a is

$$v_R(t) = V_0 \sin \omega t$$

and the instantaneous current through the resistor is

$$i_R(t) = \frac{v_R(t)}{R} = \frac{V_0}{R} \sin \omega t = I_0 \sin \omega t.$$

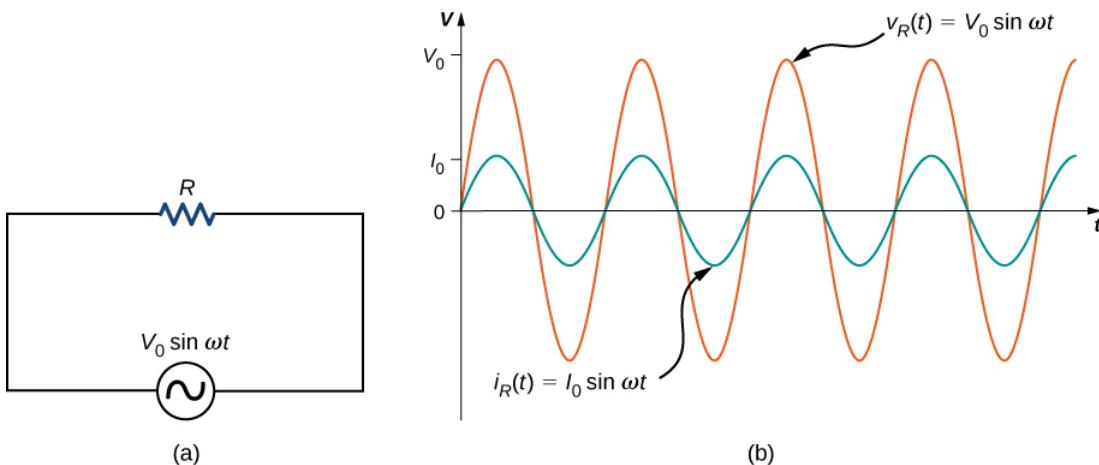


Figure 19.3.2: (a) A resistor connected across an ac voltage source. (b) The current $i_R(t)$ through the resistor and the voltage $v_R(t)$ across the resistor. The two quantities are in phase.

Here, $I_0 = V_0/R$ is the amplitude of the time-varying current. Plots of $i_R(t)$ and $v_R(t)$ are shown in Figure 19.3.2b. Both curves reach their maxima and minima at the same times, that is, the current through and the voltage across the resistor are in phase.

Graphical representations of the phase relationships between current and voltage are often useful in the analysis of ac circuits. Such representations are called **phasor diagrams**. The phasor diagram for $i_R(t)$ is shown in Figure 19.3.3a, with the current on the vertical axis. The arrow (or phasor) is rotating counterclockwise at a constant angular frequency ω , so we are viewing it at one instant in time. If the length of the arrow corresponds to the current amplitude I_0 , the projection of the rotating arrow onto the vertical axis is $i_R(t) = I_0 \sin \omega t$, which is the instantaneous current.

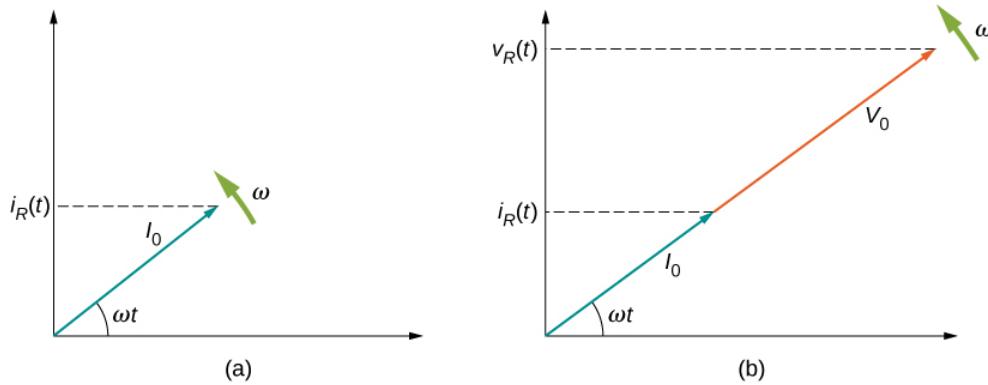


Figure 19.3.3: (a) The phasor diagram representing the current through the resistor of Figure 19.3.2 (b) The phasor diagram representing both $i_R(t)$ and $v_R(t)$.

The vertical axis on a phasor diagram could be either the voltage or the current, depending on the phasor that is being examined. In addition, several quantities can be depicted on the same phasor diagram. For example, both the current $i_R(t)$ and the voltage $v_R(t)$ are shown in the diagram of Figure 19.3.3b. Since they have the same frequency and are in phase, their phasors point in the same direction and rotate together. The relative lengths of the two phasors are arbitrary because they represent different quantities; however, the ratio of the lengths of the two phasors can be represented by the resistance, since one is a voltage phasor and the other is a current phasor.

Capacitor

Now let's consider a **capacitor** connected across an ac voltage source. From Kirchhoff's loop rule, the instantaneous voltage across the capacitor of Figure 19.3.4a is

$$v_C(t) = V_0 \sin \omega t.$$

Recall that the charge in a capacitor is given by $Q = CV$. This is true at any time measured in the ac cycle of voltage. Consequently, the instantaneous charge on the capacitor is

$$q(t) = Cv_C(t) = CV_0 \sin \omega t.$$

Since the current in the circuit is the rate at which charge enters (or leaves) the capacitor,

$$i_C(t) = \frac{dq(t)}{dt} = \omega CV_0 \cos \omega t = I_0 \cos \omega t,$$

where $I_0 = \omega CV_0$ is the current amplitude. Using the trigonometric relationship $\cos \omega t = \sin(\omega t + \pi/2)$, we may express the instantaneous current as

$$i_C(t) = I_0 \sin \left(\omega t + \frac{\pi}{2} \right).$$

Dividing V_0 by I_0 , we obtain an equation that looks similar to Ohm's law:

$$\frac{V_0}{I_0} = \frac{1}{\omega C} = X_C. \quad (19.3.1)$$

The quantity X_C is analogous to resistance in a dc circuit in the sense that both quantities are a ratio of a voltage to a current. As a result, they have the same unit, the ohm. Keep in mind, however, that a capacitor stores and discharges electric energy, whereas a resistor dissipates it. The quantity X_C is known as the capacitive reactance of the capacitor, or the opposition of a capacitor to a change in current. It depends inversely on the frequency of the ac source—high frequency leads to low capacitive reactance.

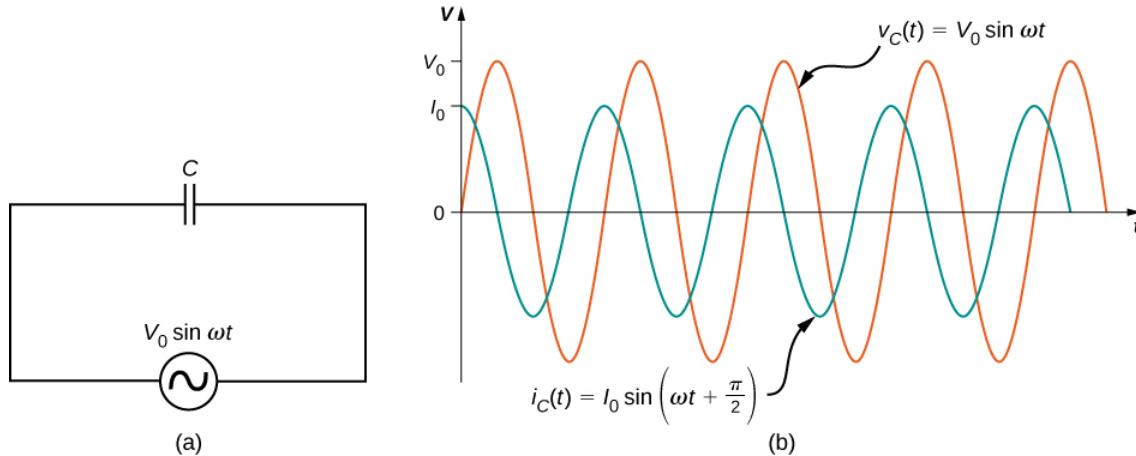


Figure 19.3.4: (a) A capacitor connected across an ac generator. (b) The current $i_C(t)$ through the capacitor and the voltage $v_C(t)$ across the capacitor. Notice that $i_C(t)$ leads $v_C(t)$ by $\pi/2$ rad.

A comparison of the expressions for $v_C(t)$ and $i_C(t)$ shows that there is a phase difference of $\pi/2$ rad between them. When these two quantities are plotted together, the current peaks a quarter cycle (or $\pi/2$ rad) ahead of the voltage, as illustrated in Figure 19.3.4b. The current through a capacitor leads the voltage across a capacitor by $\pi/2$ rad, or a quarter of a cycle.

The corresponding phasor diagram is shown in Figure 19.3.5. Here, the relationship between $i_C(t)$ and $v_C(t)$ is represented by having their phasors rotate at the same angular frequency by $\pi/2$ rad.

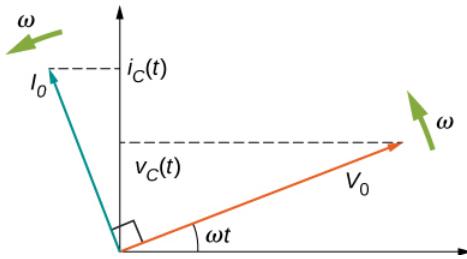


Figure 19.3.5: The current phasor leads the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

To this point, we have exclusively been using peak values of the current or voltage in our discussion, namely, I_0 and V_0 . However, if we average out the values of current or voltage, these values are zero. Therefore, we often use a second convention called the root mean square value, or rms value, in discussions of current and voltage. The rms operates in reverse of the terminology. First, you square the function, next, you take the mean, and then, you find the square root. As a result, the rms values of current and

voltage are not zero. Appliances and devices are commonly quoted with rms values for their operations, rather than peak values. We indicate rms values with a subscript attached to a capital letter (such as I_{rms}).

Although a capacitor is basically an open circuit, an **rms current**, or the root mean square of the current, appears in a circuit with an ac voltage applied to a capacitor. Consider that

 Note

$$I_{rms} = \frac{I_0}{\sqrt{2}},$$

where I_0 is the peak current in an ac system. The **rms voltage**, or the root mean square of the voltage, is

 Note

$$V_{rms} = \frac{V_0}{\sqrt{2}},$$

where V_0 is the peak voltage in an ac system. The rms current appears because the voltage is continually reversing, charging, and discharging the capacitor. If the frequency goes to zero, which would be a dc voltage, X_C tends to infinity, and the current is zero once the capacitor is charged. At very high frequencies, the capacitor's reactance tends to zero—it has a negligible reactance and does not impede the current (it acts like a simple wire).

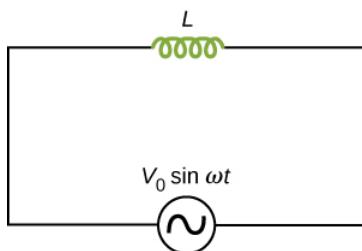
Inductor

Lastly, let's consider an **inductor** connected to an ac voltage source. From Kirchhoff's loop rule, the voltage across the inductor **L** of Figure 19.3.6a is

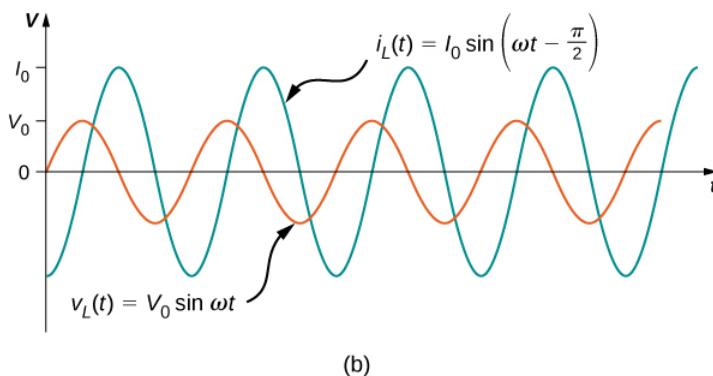
$$v_L(t) = V_0 \sin \omega t. \quad (19.3.2)$$

The emf across an inductor is equal to $\epsilon = -L(dI_L/dt)$; however, the potential difference across the inductor is $v_L(t) = Ldi_L(t)/dt$, because if we consider that the voltage around the loop must equal zero, the voltage gained from the ac source must dissipate through the inductor. Therefore, connecting this with the ac voltage source, we have

$$\frac{di_L(t)}{dt} = \frac{V_0}{L} \sin \omega t.$$



(a)



(b)

Figure 19.3.6: (a) An inductor connected across an ac generator. (b) The current $i_L(t)$ through the inductor and the voltage $v_L(t)$ across the inductor. Here $i_L(t)$ lags $v_L(t)$ by $\pi/2$ rad.

The current $i_L(t)$ is found by integrating this equation. Since the circuit does not contain a source of constant emf, there is no steady current in the circuit. Hence, we can set the constant of integration, which represents the steady current in the circuit, equal to zero, and we have

$$i_L(t) = -\frac{V_0}{\omega L} \cos \omega t = \frac{V_0}{\omega L} \sin \left(\omega t - \frac{\pi}{2} \right) = I_0 \sin \left(\omega t - \frac{\pi}{2} \right), \quad (19.3.3)$$

where $I_0 = V_0/\omega L$. The relationship between V_0 and I_0 may also be written in a form analogous to Ohm's law:

✓ Note

$$\frac{V_0}{I_0} = \omega L = X_L. \quad (19.3.4)$$

The quantity X_L is known as the **inductive reactance** of the inductor, or the opposition of an inductor to a change in current; its unit is also the ohm. Note that X_L varies directly as the frequency of the ac source—high frequency causes high inductive reactance.

A phase difference of $\pi/2$ rad occurs between the current through and the voltage across the inductor. From Equation 19.3.2 and Equation 19.3.3, the current through an inductor lags the potential difference across an inductor by $\pi/2$ rad, or a quarter of a cycle. The phasor diagram for this case is shown in Figure 19.3.7.

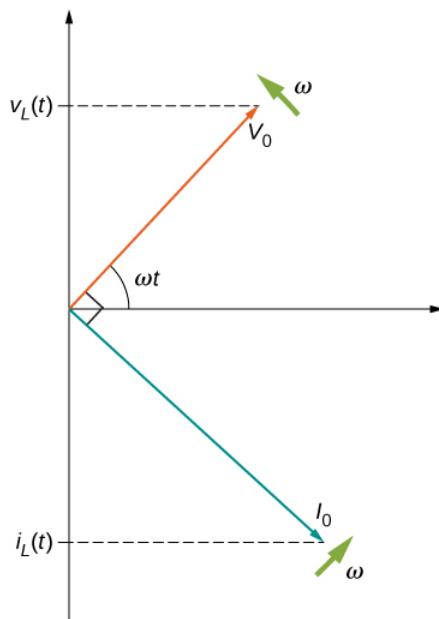


Figure 19.3.7: The current phasor lags the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

✓ Note

An animation from the University of New South Wales [AC Circuits](#) illustrates some of the concepts we discuss in this chapter. They also include wave and phasor diagrams that evolve over time so that you can get a better picture of how each changes over time.

✓ Example 19.3.1: Simple AC Circuits

An ac generator produces an emf of amplitude 10 V at a frequency $f = 60 \text{ Hz}$. Determine the voltages across and the currents through the circuit elements when the generator is connected to (a) a 100Ω resistor, (b) a $10 \mu\text{F}$ capacitor, and (c) a 15-mH inductor.

Strategy

The entire AC voltage across each device is the same as the source voltage. We can find the currents by finding the reactance X of each device and solving for the peak current using $I_0 = V_0/X$.

Solution

The voltage across the terminals of the source is

$$v(t) = V_0 \sin \omega t = (10 \text{ V}) \sin 120\pi t,$$

where $\omega = 2\pi f = 120\pi \text{ rad/s}$ is the angular frequency. Since $v(t)$ is also the voltage across each of the elements, we have

$$v(t) = v_R(t) = v_C(t) = v_L(t) = (10 \text{ V}) \sin 120\pi t.$$

a. When $R = 100 \Omega$, the amplitude of the current through the resistor is

$$I_0 = V_0/R = 10 \text{ V}/100 \Omega = 0.10 \text{ A},$$

so

$$i_R(t) = (0.10 \text{ A}) \sin 120\pi t.$$

b. From Equation 19.3.1, the capacitive reactance is

$$X_C = \frac{1}{\omega C} = \frac{1}{(120\pi \text{ rad/s})(10 \times 10^{-6} \text{ F})} = 265 \Omega,$$

so the maximum value of the current is

$$I_0 = \frac{V_0}{X_C} = \frac{10 \text{ V}}{265 \Omega} = 3.8 \times 10^{-2} \text{ A}$$

and the instantaneous current is given by

$$i_C(t) = (3.8 \times 10^{-2} \text{ A}) \sin \left(120\pi t + \frac{\pi}{2} \right).$$

c. From Equation 19.3.4, the inductive reactance is

$$X_L = \omega L = (120\pi \text{ rad/s})(15 \times 10^{-3} \text{ H}) = 5.7 \Omega.$$

The maximum current is therefore

$$I_0 = \frac{10 \text{ V}}{5.7 \Omega} = 1.8 \text{ A}$$

and the instantaneous current is

$$i_L(t) = (1.8 \text{ A}) \sin \left(120\pi t - \frac{\pi}{2} \right).$$

Significance

Although the voltage across each device is the same, the peak current has different values, depending on the reactance. The reactance for each device depends on the values of resistance, capacitance, or inductance.

Exercise 19.3.1

Repeat Example 19.3.1 for an ac source of amplitude 20 V and frequency 100 Hz.

Answer

- a. $(20 \text{ V}) \sin 200\pi t, (0.20 \text{ A}) \sin 200\pi t;$
- b. $(20 \text{ V}) \sin 200\pi t, (0.13 \text{ A}) \sin (200\pi t + \pi/2);$
- c. $(20 \text{ V}) \sin 200\pi t, (2.1 \text{ A}) \sin (200\pi t - \pi/2)$

This page titled 19.3: Simple AC Circuits is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 15.3: Simple AC Circuits by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

19.4: RLC Series Circuits with AC

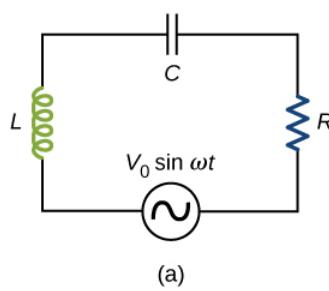
Learning Objectives

By the end of the section, you will be able to:

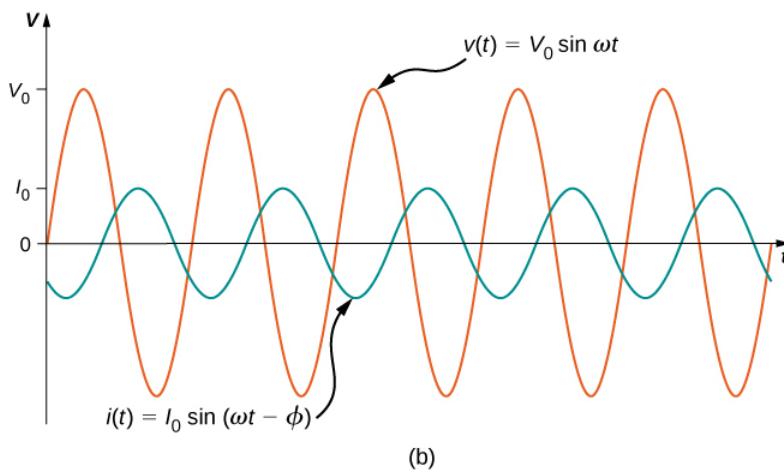
- Describe how the current varies in a resistor, a capacitor, and an inductor while in series with an ac power source
- Use phasors to understand the phase angle of a resistor, capacitor, and inductor ac circuit and to understand what that phase angle means
- Calculate the impedance of a circuit

The ac circuit shown in Figure 19.4.1, called an **RLC** series circuit, is a series combination of a resistor, capacitor, and inductor connected across an ac source. It produces an emf of

$$v(t) = V_0 \sin \omega t.$$



(a)



(b)

Figure 19.4.1: (a) An RLC series circuit. (b) A comparison of the generator output voltage and the current. The value of the phase difference ϕ depends on the values of R, C, and L.

Since the elements are in series, the same current flows through each element at all points in time. The relative phase between the current and the emf is not obvious when all three elements are present. Consequently, we represent the current by the general expression

$$i(t) = I_0 \sin(\omega t - \phi),$$

where I_0 is the current amplitude and ϕ is the phase angle between the current and the applied voltage. The phase angle is thus the amount by which the voltage and current are out of phase with each other in a circuit. Our task is to find I_0 and ϕ .

A phasor diagram involving $i(t)$, $v_R(t)$, $v_C(t)$, and $v_L(t)$ is helpful for analyzing the circuit. As shown in Figure 19.4.2, the phasor representing $v_R(t)$ points in the same direction as the phasor for $i(t)$; its amplitude is $V_R = I_0 R$. The $v_C(t)$ phasor lags the $i(t)$ phasor by $\pi/2$ rad and has the amplitude $V_C = I_0 X_C$. The phasor for $v_L(t)$ leads the $i(t)$ phasor by $\pi/2$ rad and has the amplitude $V_L = I_0 X_L$.

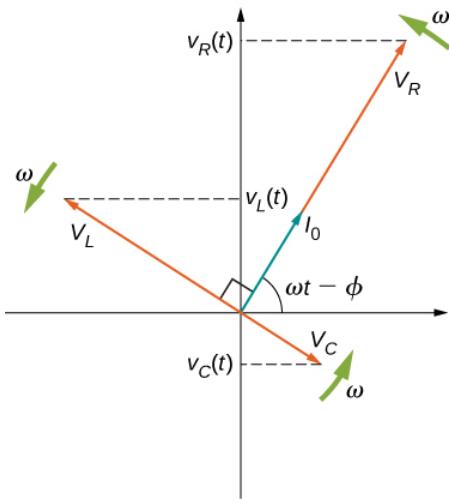


Figure 19.4.1.

At any instant, the voltage across the **RLC** combination is $v_R(t) + v_L(t) + v_C(t) = v(t)$, the emf of the source. Since a component of a sum of vectors is the sum of the components of the individual vectors—for example, $(A + B)_y = A_y + B_y$ —the projection of the vector sum of phasors onto the vertical axis is the sum of the vertical projections of the individual phasors. Hence, if we add vectorially the phasors representing $v_R(t)$, $v_L(t)$, and $v_C(t)$ and then find the projection of the resultant onto the vertical axis, we obtain

$$v_R(t) + v_L(t) + v_C(t) = v(t) = V_0 \sin \omega t.$$

The vector sum of the phasors is shown in Figure 19.4.3. The resultant phasor has an amplitude V_0 and is directed at an angle ϕ with respect to the $v_R(t)$, or $i(t)$, phasor. The projection of this resultant phasor onto the vertical axis is $v(t) = V_0 \sin \omega t$. We can easily determine the unknown quantities I_0 and ϕ from the geometry of the phasor diagram. For the phase angle,

$$\phi = \tan^{-1} \frac{V_L - V_C}{V_R} = \tan^{-1} \frac{I_0 X_L - I_0 X_C}{I_0 R},$$

and after cancellation of I_0 , this becomes

$$\phi = \tan^{-1} \frac{X_L - X_C}{R}. \quad (19.4.1)$$

Furthermore, from the Pythagorean theorem,

$$V_0 = \sqrt{V_R^2 + (V_L - V_C)^2} = \sqrt{(I_0 R)^2 + (I_0 X_L - I_0 X_C)^2} = I_0 \sqrt{R^2 + (X_L - X_C)^2}.$$

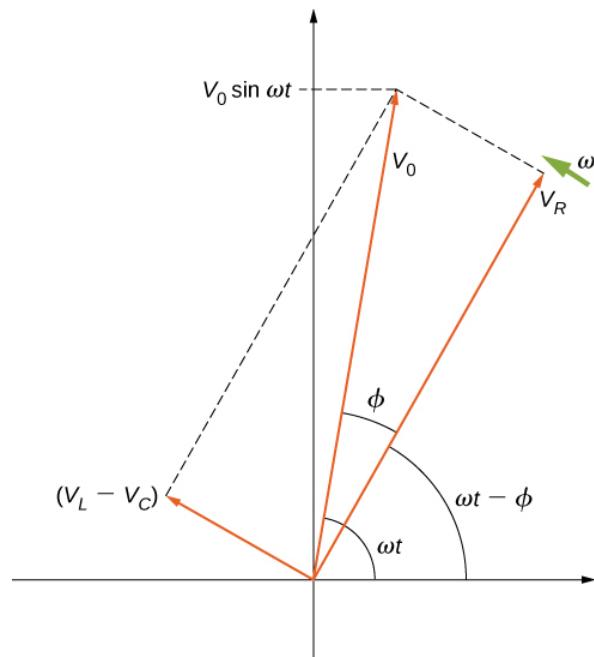


Figure 19.4.3: The resultant of the phasors for $v_L(t)$, $v_C(t)$, and $v_R(t)$ is equal to the phasor for $v_R(t) = V_0 \sin \omega t$. The $i(t)$ phasor (not shown) is aligned with the $v_R(t)$ phasor.

The current amplitude is therefore the ac version of Ohm's law:

$$I_0 = \frac{V_0}{\sqrt{R^2 + (X_L - X_C)^2}} = \frac{V_0}{Z}, \quad (19.4.2)$$

where

$$Z = \sqrt{R^2 + (X_L - X_C)^2} \quad (19.4.3)$$

is known as the impedance of the circuit. Its unit is the ohm, and it is the ac analog to resistance in a dc circuit, which measures the combined effect of resistance, capacitive reactance, and inductive reactance (Figure 19.4.4).



Figure 19.4.4: Power capacitors are used to balance the impedance of the effective inductance in transmission lines.

The **RLC** circuit is analogous to the wheel of a car driven over a corrugated road (Figure 19.4.5). The regularly spaced bumps in the road drive the wheel up and down; in the same way, a voltage source increases and decreases. The shock absorber acts like the resistance of the **RLC** circuit, damping and limiting the amplitude of the oscillation. Energy within the wheel system goes back and forth between kinetic and potential energy stored in the car spring, analogous to the shift between a maximum current, with energy stored in an inductor, and no current, with energy stored in the electric field of a capacitor. The amplitude of the wheel's motion is at a maximum if the bumps in the road are hit at the resonant frequency, which we describe in more detail in [Resonance in an AC Circuit](#).

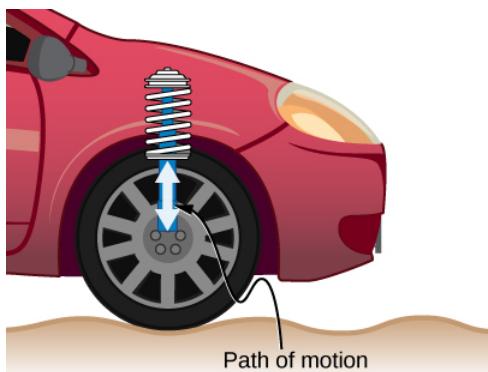


Figure 19.4.5: On a car, the shock absorber damps motion and dissipates energy. This is much like the resistance in an RLC circuit. The mass and spring determine the resonant frequency.

Problem-Solving Strategy: AC Circuits

To analyze an ac circuit containing resistors, capacitors, and inductors, it is helpful to think of each device's reactance and find the equivalent reactance using the rules we used for equivalent resistance in the past. Phasors are a great method to determine whether the emf of the circuit has positive or negative phase (namely, leads or lags other values). A mnemonic device of “ELI the ICE man” is sometimes used to remember that the emf (E) leads the current (I) in an inductor (L) and the current (I) leads the emf (E) in a capacitor (C).

Use the following steps to determine the emf of the circuit by phasors:

1. Draw the phasors for voltage across each device: resistor, capacitor, and inductor, including the phase angle in the circuit.
2. If there is both a capacitor and an inductor, find the net voltage from these two phasors, since they are antiparallel.
3. Find the equivalent phasor from the phasor in step 2 and the resistor's phasor using trigonometry or components of the phasors. The equivalent phasor found is the emf of the circuit.

✓ Example 19.4.1: An RLC Series Circuit

The output of an ac generator connected to an **RLC** series combination has a frequency of 200 Hz and an amplitude of 0.100 V. If $R = 4.00 \Omega$, $L = 3.00 \times 10^{-3} H$, and $C = 8.00 \times 10^{-4} F$, what are (a) the capacitive reactance, (b) the inductive reactance, (c) the impedance, (d) the current amplitude, and (e) the phase difference between the current and the emf of the generator?

Strategy

The reactances and impedance in (a)–(c) are found by substitutions into Equation 15.3.8, Equation 15.3.14, and Equation 19.4.2, respectively. The current amplitude is calculated from the peak voltage and the impedance. The phase difference between the current and the emf is calculated by the inverse tangent of the difference between the reactances divided by the resistance.

Solution

1. From Equation 15.3.8, the capacitive reactance is

$$X_C = \frac{1}{\omega C} = \frac{1}{2\pi(200 \text{ Hz})(8.00 \times 10^{-4} F)} = 0.995 \Omega.$$

2. From Equation 15.3.14, the inductive reactance is

$$X_L = \omega L = 2\pi(200 \text{ Hz})(3.00 \times 10^{-3} H) = 3.77 \Omega.$$

3. Substituting the values of R , X_C , and X_L into Equation 19.4.2, we obtain for the impedance

$$Z = \sqrt{(4.00)^2 + (3.77 \Omega - 0.995 \Omega)^2} = 4.87 \Omega.$$

4. The current amplitude is

$$I_0 = \frac{V_0}{Z} = \frac{0.100 V}{4.87 \Omega} = 2.05 \times 10^{-2} A.$$

5. From Equation 19.4.1, the phase difference between the current and the emf is

$$\phi = \tan^{-1} \frac{X_L - X_C}{R} = \tan^{-1} \frac{2.77 \Omega}{4.00 \Omega} = 0.607 \text{ rad.}$$

Significance

The phase angle is positive because the reactance of the inductor is larger than the reactance of the capacitor.

? Exercise 19.4.1

Find the voltages across the resistor, the capacitor, and the inductor in the circuit of Figure 19.4.1 using $v(t) = V_0 \sin \omega t$ as the output of the ac generator.

Solution

$$v_R = (V_0 R / Z) \sin(\omega t - \phi); v_C = (V_0 X_C / Z) \sin(\omega t - \phi + \pi/2) = -(V_0 X_C / Z) \cos(\omega t - \phi);$$

$$v_L = (V_0 X_L / Z) \sin(\omega t - \phi - \pi/2) = (V_0 X_L / Z) \cos(\omega t - \phi)$$

This page titled [19.4: RLC Series Circuits with AC](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **15.4: RLC Series Circuits with AC** by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

19.5: Power in an AC Circuit

Learning Objectives

By the end of the section, you will be able to:

- Describe how average power from an ac circuit can be written in terms of peak current and voltage and of rms current and voltage
- Determine the relationship between the phase angle of the current and voltage and the average power, known as the power factor

A circuit element dissipates or produces power according to $P = IV$, where I is the current through the element and V is the voltage across it. Since the current and the voltage both depend on time in an ac circuit, the instantaneous power $p(t) = i(t)v(t)$ is also time dependent. A plot of $p(t)$ for various circuit elements is shown in Figure 19.5.1. For a resistor, $i(t)$ and $v(t)$ are [in phase and therefore always have the same sign](#). For a capacitor or inductor, the relative signs of $i(t)$ and $v(t)$ vary over a cycle due to their phase differences. Consequently, $p(t)$ is positive at some times and negative at others, indicating that capacitive and inductive elements produce power at some instants and absorb it at others.

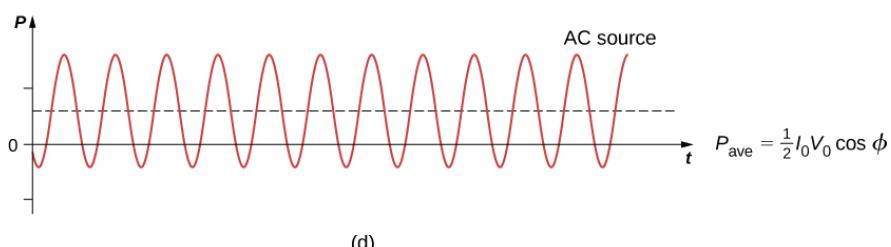
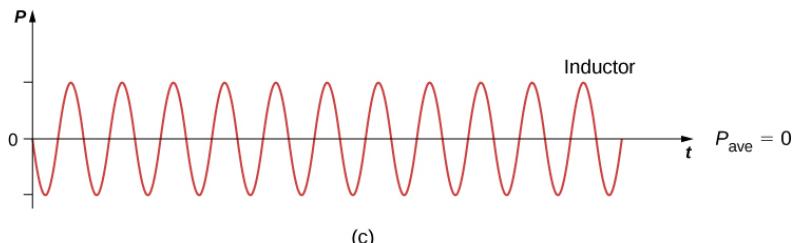
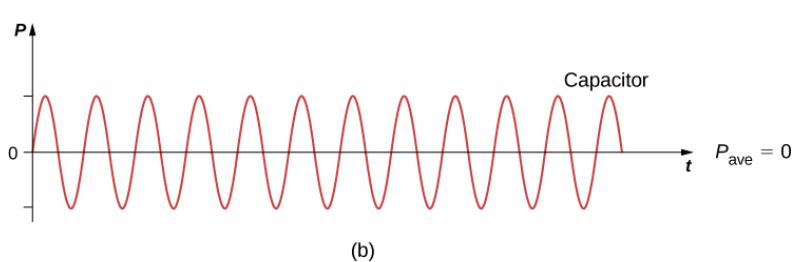
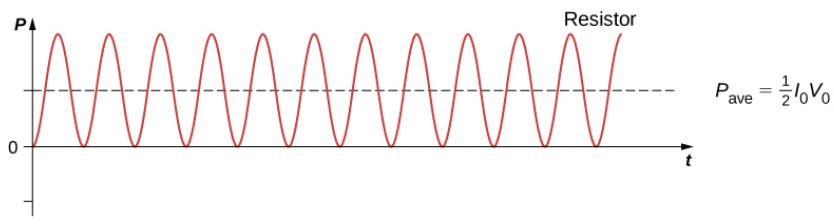


Figure 19.5.1: Graph of instantaneous power for various circuit elements. (a) For the resistor, $P_{ave} = I_0V_0/2$, whereas for (b) the capacitor and (c) the inductor, $P_{ave} = 0$. (d) For the source, $P_{ave} = I_0V_0(\cos \phi)/2$, which may be positive, negative, or zero, depending on ϕ .

Because instantaneous power varies in both magnitude and sign over a cycle, it seldom has any practical importance. What we're almost always concerned with is the power averaged over time, which we refer to as the **average power**. It is defined by the time average of the instantaneous power over one cycle:

$$P_{ave} = \frac{1}{T} \int_0^T p(t) dt, \quad (19.5.1)$$

where $T = 2\pi/\omega$ is the period of the oscillations. With the substitutions $v(t) = V_0 \sin \omega t$ and $i(t) = I_0 \sin(\omega t - \phi)$, Equation 19.5.1 becomes

$$P_{ave} = \frac{I_0V_0}{T} \int_0^T \sin(\omega t - \phi) \sin \omega t dt.$$

Using the [trigonometric difference identity](#)

$$\sin(A - B) = \sin A \cos B - \sin B \cos A$$

we obtain

$$P_{ave} = \frac{I_0 V_0 \cos \phi}{T} \int_0^T \sin^2 \omega t dt - \frac{I_0 V_0 \sin \phi}{T} \int_0^T \sin \omega t \cos \omega t dt.$$

Evaluation of these two integrals yields

$$\frac{1}{T} \int_0^T \sin^2 \omega t dt = \frac{1}{2}$$

and

$$\frac{1}{T} \int_0^T \sin \omega t \cos \omega t dt = 0.$$

Hence, the average power associated with a circuit element is given by

$$P_{ave} = \frac{1}{2} I_0 V_0 \cos \phi. \quad (19.5.2)$$

In engineering applications, $\cos \phi$ is known as the **power factor**, which is the amount by which the power delivered in the circuit is less than the theoretical maximum of the circuit due to voltage and current being out of phase. For a resistor, $\phi = 0$, so the average power dissipated is

$$P_{ave} = \frac{1}{2} I_0 V_0. \quad (19.5.3)$$

A comparison of $p(t)$ and P_{ave} is shown in Figure 19.5.1d. To make Equation 19.5.3 look like its dc counterpart, we use the rms values I_{rms} and V_{rms} of the current and the voltage. By definition, these are

$$I_{rms} = \sqrt{i_{ave}^2}$$

and

$$V_{rms} = \sqrt{v_{ave}^2},$$

where

$$i_{ave}^2 = \frac{1}{T} \int_0^T i^2(t) dt$$

and

$$v_{ave}^2 = \frac{1}{T} \int_0^T v^2(t) dt.$$

With $i(t) = I_0 \sin(\omega t - \phi)$ and $v(t) = V_0 \sin \omega t$, we obtain

$$I_{rms} = \frac{1}{\sqrt{2}} I_0$$

and

$$V_{rms} = \frac{1}{\sqrt{2}} V_0.$$

We may then write for the average power dissipated by a resistor,

$$P_{ave} = \frac{1}{2} I_0 V_0 = I_{rms} V_{rms} = I_{rms}^2 R. \quad (19.5.4)$$

This equation further emphasizes why the rms value is chosen in discussion rather than peak values. Both Equations 19.5.2 and 19.5.4 are correct for average power, but the rms values in the formula give a cleaner representation, so the extra factor of 1/2 is not necessary.

Alternating voltages and currents are usually described in terms of their rms values. For example, the 110 V from a household outlet is an rms value. The amplitude of this source is $110\sqrt{2} \text{ V} = 156 \text{ V}$. Because most ac meters are calibrated in terms of rms values, a typical ac voltmeter placed across a household outlet will read 110 V.

For a capacitor and an inductor, $\phi = \pi/2$ and $-\pi/2 \text{ rad}$, respectively. Since $\cos \pi/2 = \cos(-\pi/2) = 0$, we find from Equation 19.5.2 that the average power dissipated by either of these elements is $P_{ave} = 0$. Capacitors and inductors absorb energy from the circuit during one half-cycle and then discharge it back to the circuit during the other half-cycle. This behavior is illustrated in the plots of Figures 19.5.1b and 19.5.1c which show $p(t)$ oscillating sinusoidally about zero.

The phase angle for an ac generator may have any value. If $\cos \phi > 0$, the generator produces power; if $\cos \phi < 0$, it absorbs power. In terms of rms values, the average power of an ac generator is written as

$$P_{ave} = I_{rms} V_{rms} \cos \phi.$$

For the generator in an RLC circuit,

$$\tan \phi = \frac{X_L - X_C}{R}$$

and

$$\cos \phi = \frac{R}{\sqrt{R^2 + (X_L - X_C)^2}} = \frac{R}{Z}.$$

Hence the average power of the generator is

$$P_{ave} = I_{rms} V_{rms} \cos \phi = \frac{V_{rms}}{Z} V_{rms} \frac{R}{Z} = \frac{V_{rms}^2 R}{Z^2}. \quad (19.5.5)$$

This can also be written as

$$P_{ave} = I_{rms}^2 R,$$

which designates that the power produced by the generator is dissipated in the resistor. As we can see, Ohm's law for the rms ac is found by dividing the rms voltage by the impedance.

✓ Example 19.5.1: Power Output of a Generator

An ac generator whose emf is given by

$$v(t) = (4.00 \text{ V}) \sin [(1.00 \times 10^4 \text{ rad/s})t]$$

is connected to an RLC circuit for which $L = 2.00 \times 10^{-3} \text{ H}$, $C = 4.00 \times 10^{-6} \text{ F}$, and $R = 5.00 \Omega$.

- What is the rms voltage across the generator?
- What is the impedance of the circuit?
- What is the average power output of the generator?

Strategy

The rms voltage is the amplitude of the voltage times $1/\sqrt{2}$. The impedance of the circuit involves the resistance and the reactances of the capacitor and the inductor. The average power is calculated by Equation 19.5.5 because we have the impedance of the circuit Z , the rms voltage V_{rms} , and the resistance R .

Solution

- Since $V_0 = 4.00 \text{ V}$, the rms voltage across the generator is

$$V_{rms} = \frac{1}{\sqrt{2}}(4.00 \text{ V}) = 2.83 \text{ V}.$$

- The impedance of the circuit is

$$\begin{aligned}
 Z &= \sqrt{r^2 + (x_l - x_c)^2} \\
 &= \sqrt{(5.00\Omega)^2 + \left[(1.00 \times 10^4 \text{ rad/s})(2.00 \times 10^{-3} \text{ H}) - \frac{1}{(1.00 \times 10^4 \text{ rad/s})(4.00 \times 10^{-6} \text{ F})} \right]^2} \\
 &= 7.07\Omega.
 \end{aligned}$$

3. From Equation 19.5.5, the average power transferred to the circuit is

$$P_{ave} = \frac{V_{rms}^2 R}{Z^2} = \frac{(2.83\text{ V})^2 (5.00\Omega)}{(7.07\Omega)^2} = 0.801\text{ W}.$$

Significance

If the resistance is much larger than the reactance of the capacitor or inductor, the average power is a dc circuit equation of $P = V^2/R$, where V replaces the rms voltage.

Exercise 19.5.1A

An ac voltmeter attached across the terminals of a 45-Hz ac generator reads 7.07 V. Write an expression for the emf of the generator.

Answer

$$v(t) = (10.0\text{ V}) \sin 90\pi t$$

Exercise 19.5.1B

Show that the rms voltages across a resistor, a capacitor, and an inductor in an ac circuit where the rms current is I_{rms} are given by $I_{rms}R$, $I_{rms}X_C$, and $I_{rms}X_L$, respectively. Determine these values for the components of the RLC circuit of Equation 19.5.2.

Answer

$$2.00\text{ V}; 10.01\text{ V}; 8.01\text{ V}$$

This page titled 19.5: Power in an AC Circuit is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 19.5: Power in an AC Circuit by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

19.6: Resonance in an AC Circuit

Learning Objectives

By the end of the section, you will be able to:

- Determine the peak ac resonant angular frequency for a RLC circuit
- Explain the width of the average power versus angular frequency curve and its significance using terms like bandwidth and quality factor

In the **RLC** series circuit of [Figure 15.4.1](#), the current amplitude is, from [Equation 15.4.7](#),

$$I_0 = \frac{V_0}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}}. \quad (19.6.1)$$

If we can vary the frequency of the ac generator while keeping the amplitude of its output voltage constant, then the current changes accordingly. A plot of I_0 versus ω is shown in Figure 19.6.1.

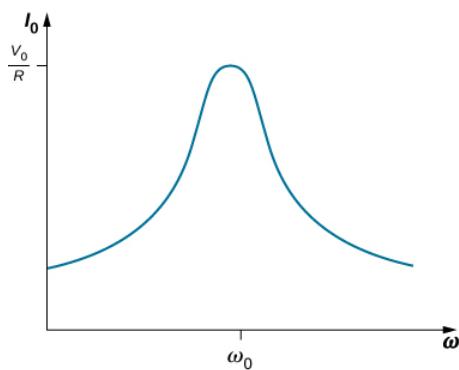


Figure 19.6.1: At an RLC circuit's resonant frequency, $\omega_0 = \sqrt{1/LC}$, the current amplitude is at its maximum value.

In [Oscillations](#), we encountered a similar graph where the amplitude of a damped harmonic oscillator was plotted against the angular frequency of a sinusoidal driving force (see [Forced Oscillations](#)). This similarity is more than just a coincidence, as shown earlier by the application of Kirchhoff's loop rule to the circuit of [Figure 15.4.1](#). This yields

$$L \frac{di}{dt} + iR + \frac{q}{C} = V_0 \sin \omega t, \quad (19.6.2)$$

or

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{1}{C} q = V_0 \sin \omega t,$$

where we substituted $d\mathbf{q}(t)/dt$ for $\mathbf{i}(t)$. A comparison of Equation 19.6.2 and, from [Oscillations](#), [Damped Oscillations](#) for damped harmonic motion clearly demonstrates that the driven **RLC** series circuit is the electrical analog of the driven damped harmonic oscillator.

The **resonant frequency f_0** of the **RLC** circuit is the frequency at which the amplitude of the current is a maximum and the circuit would oscillate if not driven by a voltage source. By inspection, this corresponds to the angular frequency $\omega_0 = 2\pi f_0$ at which the impedance Z in Equation 19.6.1 is a minimum, or when

$$\omega_0 L = \frac{1}{\omega_0 C} \quad (19.6.3)$$

and

$$\omega_0 = \sqrt{\frac{1}{LC}}. \quad (19.6.4)$$

This is the resonant angular frequency of the circuit. Substituting ω_0 into Equation 15.4.5, Equation 15.4.7, and Equation 15.4.8, we find that at resonance,

$$\phi = \tan^{-1}(0) = 0, I_0 = V_0/R, \text{ and } Z = R.$$

Therefore, at resonance, an **RLC** circuit is purely resistive, with the applied emf and current in phase.

What happens to the power at resonance? Equation 15.5.18 tells us how the average power transferred from an ac generator to the **RLC** combination varies with frequency. In addition, P_{ave} reaches a maximum when Z , which depends on the frequency, is a minimum, that is, when $X_L = X_C$ and $Z = R$. Thus, at resonance, the average power output of the source in an **RLC** series circuit is a maximum. From Equation 15.5.18, this maximum is V_{rms}^2/R .

Figure 19.6.2 is a typical plot of P_{ave} versus ω in the region of maximum power output. The **bandwidth** $\Delta\omega$ of the resonance peak is defined as the range of angular frequencies ω over which the average power P_{ave} is greater than one-half the maximum value of P_{ave} . The sharpness of the peak is described by a dimensionless quantity known as the **quality factor** Q of the circuit. By definition,

$$Q = \frac{\omega_0}{\Delta\omega}, \quad (19.6.5)$$

where ω_0 is the resonant angular frequency. A high Q indicates a sharp resonance peak. We can give Q in terms of the circuit parameters as

$$Q = \frac{\omega_0 L}{R}. \quad (19.6.6)$$

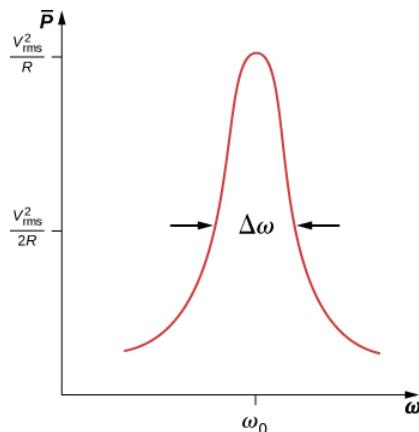


Figure 19.6.2: Like the current, the average power transferred from an ac generator to an **RLC** circuit peaks at the resonant frequency.

Resonant circuits are commonly used to pass or reject selected frequency ranges. This is done by adjusting the value of one of the elements and hence “tuning” the circuit to a particular resonant frequency. For example, in radios, the receiver is tuned to the desired station by adjusting the resonant frequency of its circuitry to match the frequency of the station. If the tuning circuit has a high Q , it will have a small bandwidth, so signals from other stations at frequencies even slightly different from the resonant frequency encounter a high impedance and are not passed by the circuit. Cell phones work in a similar fashion, communicating with signals of around 1 GHz that are tuned by an inductor-capacitor circuit. One of the most common applications of capacitors is their use in ac-timing circuits, based on attaining a resonant frequency. A metal detector also uses a shift in resonance frequency in detecting metals (Figure 19.6.3).



Figure 19.6.3: When a metal detector comes near a piece of metal, the self-inductance of one of its coils changes. This causes a shift in the resonant frequency of a circuit containing the coil. That shift is detected by the circuitry and transmitted to the diver by means of the headphones.

✓ Example 19.6.1: Resonance in an RLC Series Circuit

- What is the resonant frequency of the circuit of [Example 15.3.1](#)?
- If the ac generator is set to this frequency without changing the amplitude of the output voltage, what is the amplitude of the current?

Strategy

The resonant frequency for a **RLC** circuit is calculated from Equation [19.6.4](#), which comes from a balance between the reactances of the capacitor and the inductor. Since the circuit is at resonance, the impedance is equal to the resistor. Then, the peak current is calculated by the voltage divided by the resistance.

Solution

- The resonant frequency is found from Equation [19.6.4](#):

$$\begin{aligned} f_0 &= \frac{1}{2\pi} \sqrt{\frac{1}{LC}} \\ &= \frac{1}{2\pi} \sqrt{\frac{1}{(3.00 \times 10^{-3} H)(8.00 \times 10^{-4} F)}} \\ &= 1.03 \times 10^2 \text{ Hz}. \end{aligned}$$

- At resonance, the impedance of the circuit is purely resistive, and the current amplitude is

$$I_0 = \frac{0.100 V}{4.00 \Omega} = 2.50 \times 10^{-2} A.$$

Significance

If the circuit were not set to the resonant frequency, we would need the impedance of the entire circuit to calculate the current.

✓ Example 19.6.2: Power Transfer in an RLC Series Circuit at Resonance

- What is the resonant angular frequency of an **RLC** circuit with $R = 0.200 \Omega$, $L = 4.00 \times 10^{-3} H$, and $C = 2.00 \times 10^{-6} F$?
- If an ac source of constant amplitude 4.00 V is set to this frequency, what is the average power transferred to the circuit?
- Determine **Q** and the bandwidth of this circuit.

Strategy

The resonant angular frequency is calculated from Equation 19.6.4. The average power is calculated from the rms voltage and the resistance in the circuit. The quality factor is calculated from Equation 19.6.6 and by knowing the resonant frequency. The bandwidth is calculated from Equation 19.6.5 and by knowing the quality factor.

Solution

1. The resonant angular frequency is

$$\begin{aligned}\omega_0 &= \sqrt{\frac{1}{LC}} \\ &= \sqrt{\frac{1}{(4.00 \times 10^{-3}H)(2.00 \times 10^{-6}F)}} \\ &= 1.12 \times 10^4 \text{ rad/s.}\end{aligned}$$

2. At this frequency, the average power transferred to the circuit is a maximum. It is

$$P_{ave} = \frac{V_{rms}^2}{R} = \frac{[(1/\sqrt{2})(4.00 V)]^2}{0.200 \Omega} = 40.0 \text{ W.}$$

3. The quality factor of the circuit is

$$Q = \frac{\omega_0 L}{R} = \frac{(1.12 \times 10^4 \text{ rad/s})(4.00 \times 10^{-3}H)}{0.200 \Omega} = 224.$$

We then find for the bandwidth

$$\Delta\omega = \frac{\omega_0}{Q} = \frac{1.12 \times 10^4 \text{ rad/s}}{224} = 50.0 \text{ rad/s.}$$

Significance

If a narrower bandwidth is desired, a lower resistance or higher inductance would help. However, a lower resistance increases the power transferred to the circuit, which may not be desirable, depending on the maximum power that could possibly be transferred.

Exercise 19.6.1

In the circuit of Figure 15.4.1, $L = 2.0 \times 10^{-3}H$, $C = 5.0 \times 10^{-4}F$, and $R = 40 \Omega$.

- What is the resonant frequency?
- What is the impedance of the circuit at resonance?
- If the voltage amplitude is 10 V, what is $i(t)$ at resonance?
- The frequency of the AC generator is now changed to 200 Hz. Calculate the phase difference between the current and the emf of the generator.

Answer

- 160 Hz; b. 40Ω ; c. $(0.25A) \sin 10^3t$; d. 0.023 rad

Exercise 19.6.2

What happens to the resonant frequency of an RLC series circuit when the following quantities are increased by a factor of 4: (a) the capacitance, (b) the self-inductance, and (c) the resistance?

Answer

- halved; b. halved; c. same

? Exercise 19.6.3

The resonant angular frequency of an **RLC** series circuit is $4.0 \times 10^2 \text{ rad/s}$. An ac source operating at this frequency transfers an average power of $2.0 \times 10^{-2} \text{ W}$ to the circuit. The resistance of the circuit is 0.50Ω . Write an expression for the emf of the source.

Answer

$$v(t) = (0.14 \text{ V}) \sin(4.0 \times 10^2 t)$$

This page titled [19.6: Resonance in an AC Circuit](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [15.6: Resonance in an AC Circuit](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

19.7: AC Safety - Grounding and Bonding

19.7: AC Safety - Grounding and Bonding is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

19.8: Alternating-Current Circuits (Summary)

Key Terms

ac current	current that fluctuates sinusoidally with time at a fixed frequency
ac voltage	voltage that fluctuates sinusoidally with time at a fixed frequency
alternating current (ac)	flow of electric charge that periodically reverses direction
average power	time average of the instantaneous power over one cycle
bandwidth	range of angular frequencies over which the average power is greater than one-half the maximum value of the average power
capacitive reactance	opposition of a capacitor to a change in current
direct current (dc)	flow of electric charge in only one direction
impedance	ac analog to resistance in a dc circuit, which measures the combined effect of resistance, capacitive reactance, and inductive reactance
inductive reactance	opposition of an inductor to a change in current
phase angle	amount by which the voltage and current are out of phase with each other in a circuit
power factor	amount by which the power delivered in the circuit is less than the theoretical maximum of the circuit due to voltage and current being out of phase
quality factor	dimensionless quantity that describes the sharpness of the peak of the bandwidth; a high quality factor is a sharp or narrow resonance peak
resonant frequency	frequency at which the amplitude of the current is a maximum and the circuit would oscillate if not driven by a voltage source
rms current	root mean square of the current
rms voltage	root mean square of the voltage
step-down transformer	transformer that decreases voltage and increases current
step-up transformer	transformer that increases voltage and decreases current
transformer	device that transforms voltages from one value to another using induction
transformer equation	equation showing that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of turns in their windings

Key Equations

AC voltage	$v = V_0 \sin \omega t$
AC current	$i = I_0 \sin \omega t$
capacitive reactance	$\frac{V_0}{I_0} = \frac{1}{\omega C} = X_C$
rms voltage	$V_{rms} = \frac{V_0}{\sqrt{2}}$
rms current	$I_{rms} = \frac{I_0}{\sqrt{2}}$
inductive reactance	$\frac{V_0}{I_0} = \omega L = X_L$
Phase angle of an RLC series circuit	$\phi = \tan^{-1} \frac{X_L - X_C}{R}$
AC version of Ohm's law	$I_0 = \frac{V_0}{Z}$
Impedance of an RLC series circuit	$Z = \sqrt{R^2 + (X_L - X_C)^2}$
Average power associated with a circuit element	$P_{ave} = \frac{1}{2} I_0 V_0 \cos \phi$
Average power dissipated by a resistor	$P_{ave} = \frac{1}{2} I_0 V_0 = I_{rms} V_{rms} = I_{rms}^2 R$
Resonant angular frequency of a circuit	$\omega_0 = \sqrt{\frac{1}{LC}}$
Quality factor of a circuit	$Q = \frac{\omega_0}{\Delta \omega}$
Quality factor of a circuit in terms of the circuit parameters	$Q = \frac{\omega_0 L}{R}$
Transformer equation with voltage	$\frac{V_S}{V_P} = \frac{N_S}{N_P}$
Transformer equation with current	$I_S = \frac{N_P}{N_S} I_P$

Summary

15.2 AC Sources

- Direct current (dc) refers to systems in which the source voltage is constant.
- Alternating current (ac) refers to systems in which the source voltage varies periodically, particularly sinusoidally.
- The voltage source of an ac system puts out a voltage that is calculated from the time, the peak voltage, and the angular frequency.
- In a simple circuit, the current is found by dividing the voltage by the resistance. An ac current is calculated using the peak current (determined by dividing the peak voltage by the resistance), the angular frequency, and the time.

15.3 Simple AC Circuits

- For resistors, the current through and the voltage across are in phase.
- For capacitors, we find that when a sinusoidal voltage is applied to a capacitor, the voltage follows the current by one-fourth of a cycle. Since a capacitor can stop current when fully charged, it limits current and offers another form of ac resistance, called capacitive reactance, which has units of ohms.
- For inductors in ac circuits, we find that when a sinusoidal voltage is applied to an inductor, the voltage leads the current by one-fourth of a cycle.
- The opposition of an inductor to a change in current is expressed as a type of ac reactance. This inductive reactance, which has units of ohms, varies with the frequency of the ac source.

15.4 RLC Series Circuits with AC

- An **RLC** series circuit is a resistor, capacitor, and inductor series combination across an ac source.
- The same current flows through each element of an **RLC** series circuit at all points in time.
- The counterpart of resistance in a dc circuit is impedance, which measures the combined effect of resistors, capacitors, and inductors. The maximum current is defined by the ac version of Ohm's law.
- Impedance has units of ohms and is found using the resistance, the capacitive reactance, and the inductive reactance.

15.5 Power in an AC Circuit

- The average ac power is found by multiplying the rms values of current and voltage.
- Ohm's law for the rms ac is found by dividing the rms voltage by the impedance.
- In an ac circuit, there is a phase angle between the source voltage and the current, which can be found by dividing the resistance by the impedance.
- The average power delivered to an **RLC** circuit is affected by the phase angle.
- The power factor ranges from -1 to 1 .

15.6 Resonance in an AC Circuit

- At the resonant frequency, inductive reactance equals capacitive reactance.
- The average power versus angular frequency plot for a **RLC** circuit has a peak located at the resonant frequency; the sharpness or width of the peak is known as the bandwidth.
- The bandwidth is related to a dimensionless quantity called the quality factor. A high quality factor value is a sharp or narrow peak.

15.7 Transformers

- Power plants transmit high voltages at low currents to achieve lower ohmic losses in their many kilometers of transmission lines.
- Transformers use induction to transform voltages from one value to another.
- For a transformer, the voltages across the primary and secondary coils, or windings, are related by the transformer equation.
- The currents in the primary and secondary windings are related by the number of primary and secondary loops, or turns, in the windings of the transformer.
- A step-up transformer increases voltage and decreases current, whereas a step-down transformer decreases voltage and increases current.

This page titled [19.8: Alternating-Current Circuits \(Summary\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [15.8: Alternating-Current Circuits \(Summary\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/university-physics-volume-2>.

19.9: Alternating-Current Circuits (Exercise)

Conceptual Questions

15.2 AC Sources

1. What is the relationship between frequency and angular frequency?

15.3 Simple AC Circuits

2. Explain why at high frequencies a capacitor acts as an ac short, whereas an inductor acts as an open circuit.

15.4 RLC Series Circuits with AC

3. In an **RLC** series circuit, can the voltage measured across the capacitor be greater than the voltage of the source? Answer the same question for the voltage across the inductor.

15.5 Power in an AC Circuit

4. For what value of the phase angle ϕ between the voltage output of an ac source and the current is the average power output of the source a maximum?
5. Discuss the differences between average power and instantaneous power.
6. The average ac current delivered to a circuit is zero. Despite this, power is dissipated in the circuit. Explain.
7. Can the instantaneous power output of an ac source ever be negative? Can the average power output be negative?
8. The power rating of a resistor used in ac circuits refers to the maximum average power dissipated in the resistor. How does this compare with the maximum instantaneous power dissipated in the resistor?

15.7 Transformers

9. Why do transmission lines operate at very high voltages while household circuits operate at fairly small voltages?
10. How can you distinguish the primary winding from the secondary winding in a step-up transformer?
11. Battery packs in some electronic devices are charged using an adapter connected to a wall socket. Speculate as to the purpose of the adapter.
12. Will a transformer work if the input is a dc voltage?
13. Why are the primary and secondary coils of a transformer wrapped around the same closed loop of iron?

Problems

15.2 AC Sources

14. Write an expression for the output voltage of an ac source that has an amplitude of 12 V and a frequency of 200 Hz.

15.3 Simple AC Circuits

15. Calculate the reactance of a **5.0- μ F** capacitor at

- (a) 60 Hz,
- (b) 600 Hz, and
- (c) 6000 Hz.

16. What is the capacitance of a capacitor whose reactance is **10Ω** at 60 Hz?

17. Calculate the reactance of a 5.0-mH inductor at

- (a) 60 Hz,
- (b) 600 Hz, and

(c) 6000 Hz.

- 18.** What is the self-inductance of a coil whose reactance is **10Ω** at 60 Hz?
- 19.** At what frequency is the reactance of a **20-μF** capacitor equal to that of a 10-mH inductor?
- 20.** At 1000 Hz, the reactance of a 5.0-mH inductor is equal to the reactance of a particular capacitor. What is the capacitance of the capacitor?
- 21.** A **50-Ω** resistor is connected across the emf $v(t) = (160V)\sin(120\pi t)$. Write an expression for the current through the resistor.
- 22.** A **25-μF** capacitor is connected to an emf given by $v(t) = (160V)\sin(120\pi t)$.
- What is the reactance of the capacitor?
 - Write an expression for the current output of the source.
- 23.** A 100-mH inductor is connected across the emf of the preceding problem.
- What is the reactance of the inductor?
 - Write an expression for the current through the inductor.

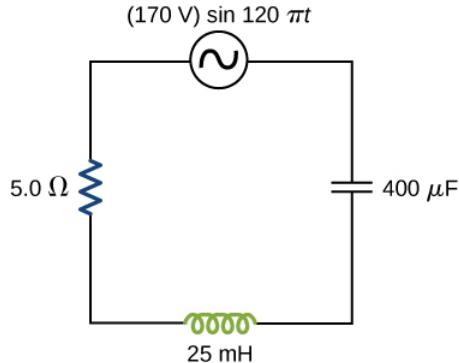
15.4 RLC Series Circuits with AC

- 24.** What is the impedance of a series combination of a **50-Ω** resistor, a **5.0-μF**, **5.0-μF** capacitor, and a **10-μF**, **10-μF** capacitor at a frequency of 2.0 kHz?
- 25.** A resistor and capacitor are connected in series across an ac generator. The emf of the generator is given by $v(t) = V_0\cos\omega t$, where $V_0 = 120V$, $\omega = 120\pi rad/s$, $R = 400\Omega$, and $C = 4.0\mu F$.
- What is the impedance of the circuit?
 - What is the amplitude of the current through the resistor?
 - Write an expression for the current through the resistor.
 - Write expressions representing the voltages across the resistor and across the capacitor.
- 26.** A resistor and inductor are connected in series across an ac generator. The emf of the generator is given by $v(t) = V_0\cos\omega t$, where $V_0 = 120V$ and $\omega = 120\pi rad/s$; also, $R = 400\Omega$ and $L = 1.5H$.
- What is the impedance of the circuit?
 - What is the amplitude of the current through the resistor?
 - Write an expression for the current through the resistor.
 - Write expressions representing the voltages across the resistor and across the inductor.
- 27.** In an **RLC** series circuit, the voltage amplitude and frequency of the source are 100 V and 500 Hz, respectively, an $R = 500\Omega$, $L = 0.20H$, and $C = 2.0\mu F$.
- What is the impedance of the circuit?
 - What is the amplitude of the current from the source?
 - If the emf of the source is given by $v(t) = (100V)\sin 1000\pi t$, how does the current vary with time?
 - Repeat the calculations with i changed to **0.20μF**.
- 28.** An **RLC** series circuit with $R = 600\Omega$, $L = 30mH$, and $C = 0.050\mu F$ is driven by an ac source whose frequency and voltage amplitude are 500 Hz and 50 V, respectively.
- What is the impedance of the circuit?
 - What is the amplitude of the current in the circuit?

(c) What is the phase angle between the emf of the source and the current?

29. For the circuit shown below, what are

- (a) the total impedance and
- (b) the phase angle between the current and the emf?
- (c) Write an expression for $i(t)$.



15.5 Power in an AC Circuit

30. The emf of an ac source is given by $v(t) = V_0 \sin \omega t$, where $V_0 = 100\text{V}$ and $\omega = 200\pi \text{ rad/s}$. Calculate the average power output of the source if it is connected across

- (a) a $20\text{-}\mu\text{F}$ capacitor,
- (b) a 20-mH inductor, and
- (c) a $50\text{-}\Omega$ resistor.

31. Calculate the rms currents for an ac source is given by $v(t) = V_0 \sin \omega t$, where $V_0 = 100\text{V}$ and $\omega = 200\pi \text{ rad/s}$ when connected across

- (a) a $20\text{-}\mu\text{F}$ capacitor,
- (b) a 20-mH inductor, and
- (c) a $50\text{-}\Omega$ resistor.

32. A 40-mH inductor is connected to a 60-Hz AC source whose voltage amplitude is 50 V . If an AC voltmeter is placed across the inductor, what does it read?

33. For an **RLC** series circuit, the voltage amplitude and frequency of the source are 100 V and 500 Hz , respectively; $R=500\Omega$; and $L=0.20\text{H}$. Find the average power dissipated in the resistor for the following values for the capacitance:

- (a) $C=2.0\mu\text{F}$ and
- (b) $C=0.20\mu\text{F}$.

34. An ac source of voltage amplitude 10 V delivers electric energy at a rate of 0.80 W when its current output is 2.5 A . What is the phase angle ϕ between the emf and the current?

35. An **RLC** series circuit has an impedance of 60Ω and a power factor of 0.50 , with the voltage lagging the current. (a) Should a capacitor or an inductor be placed in series with the elements to raise the power factor of the circuit? (b) What is the value of the reactance across the inductor that will raise the power factor to unity?

15.6 Resonance in an AC Circuit

36. (a) Calculate the resonant angular frequency of an **RLC** series circuit for which $R = 20\Omega$, $L = 75\text{mH}$, and $C=4.0\mu\text{F}$. (b) If R is changed to 300Ω , what happens to the resonant angular frequency?

37. The resonant frequency of an **RLC** series circuit is $2.0 \times 10^3 \text{ Hz}$. If the self-inductance in the circuit is 5.0 mH, what is the capacitance in the circuit?

38. (a) What is the resonant frequency of an **RLC** series circuit with **R=20Ω**, **L=2.0mH**, and **C=4.0μF**?

(b) What is the impedance of the circuit at resonance?

39. For an **RLC** series circuit, **R=100Ω**, **L=150mH**, and **C=0.25μF**.

(a) If an ac source of variable frequency is connected to the circuit, at what frequency is maximum power dissipated in the resistor?

(b) What is the quality factor of the circuit?

40. An ac source of voltage amplitude 100 V and variable frequency **f** drives an **RLC** series circuit with **R=10Ω**, **L=2.0mH**, and **C=25μF**.

(a) Plot the current through the resistor as a function of the frequency **f**.

(b) Use the plot to determine the resonant frequency of the circuit.

41. (a) What is the resonant frequency of a resistor, capacitor, and inductor connected in series if **R=100Ω**, **L=2.0H**, and **C=5.0μF**?

(b) If this combination is connected to a 100-V source operating at the constant frequency, what is the power output of the source?

(c) What is the **Q** of the circuit?

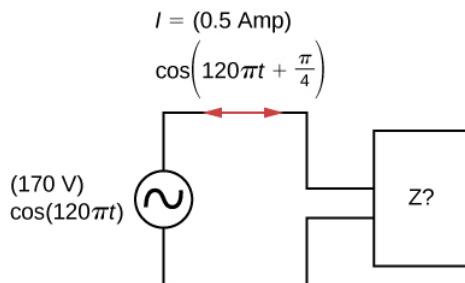
(d) What is the bandwidth of the circuit?

42. Suppose a coil has a self-inductance of 20.0 H and a resistance of **200Ω**. What

(a) capacitance and

(b) resistance must be connected in series with the coil to produce a circuit that has a resonant frequency of 100 Hz and a **Q** of 10?

43. An ac generator is connected to a device whose internal circuits are not known. We only know current and voltage outside the device, as shown below. Based on the information given, what can you infer about the electrical nature of the device and its power usage?



15.7 Transformers

44. A step-up transformer is designed so that the output of its secondary winding is 2000 V (rms) when the primary winding is connected to a 110-V (rms) line voltage.

(a) If there are 100 turns in the primary winding, how many turns are there in the secondary winding?

(b) If a resistor connected across the secondary winding draws an rms current of 0.75 A, what is the current in the primary winding?

45. A step-up transformer connected to a 110-V line is used to supply a hydrogen-gas discharge tube with 5.0 kV (rms). The tube dissipates 75 W of power.

- (a) What is the ratio of the number of turns in the secondary winding to the number of turns in the primary winding?
- (b) What are the rms currents in the primary and secondary windings?
- (c) What is the effective resistance seen by the 110-V source?

46. An ac source of emf delivers 5.0 mW of power at an rms current of 2.0 mA when it is connected to the primary coil of a transformer. The rms voltage across the secondary coil is 20 V.

- (a) What are the voltage across the primary coil and the current through the secondary coil?
- (b) What is the ratio of secondary to primary turns for the transformer?

47. A transformer is used to step down 110 V from a wall socket to 9.0 V for a radio.

- (a) If the primary winding has 500 turns, how many turns does the secondary winding have?
- (b) If the radio operates at a current of 500 mA, what is the current through the primary winding?

48. A transformer is used to supply a 12-V model train with power from a 110-V wall plug. The train operates at 50 W of power.

- (a) What is the rms current in the secondary coil of the transformer?
- (b) What is the rms current in the primary coil?
- (c) What is the ratio of the number of primary to secondary turns?
- (d) What is the resistance of the train?
- (e) What is the resistance seen by the 110-V source?

Additional Problems

49. The emf of an dc source is given by $v(t) = V_0 \sin \omega t$, where $V_0 = 100V$ and $\omega = 200\pi rad/s$. Find an expression that represents the output current of the source if it is connected across

- (a) a **20- μ F** capacitor,
- (b) a 20-mH inductor, and
- (c) a **50- Ω** resistor.

50. A **700- μ F** capacitor is connected across an ac source with a voltage amplitude of 160 V and a frequency of 20 kHz.

- (a) Determine the capacitive reactance of the capacitor and the amplitude of the output current of the source.
- (b) If the frequency is changed to 60 Hz while keeping the voltage amplitude at 160 V, what are the capacitive reactance and the current amplitude?

51. A 20-mH inductor is connected across an AC source with a variable frequency and a constant-voltage amplitude of 9.0 V.

- (a) Determine the reactance of the circuit and the maximum current through the inductor when the frequency is set at 20 kHz.
- (b) Do the same calculations for a frequency of 60 Hz.

52. A **30- μ F** capacitor is connected across a 60-Hz ac source whose voltage amplitude is 50 V.

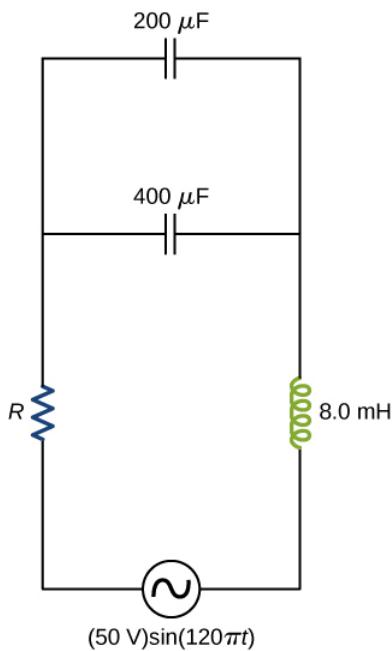
- (a) What is the maximum charge on the capacitor?
- (b) What is the maximum current into the capacitor?
- (c) What is the phase relationship between the capacitor charge and the current in the circuit?

53. A 7.0-mH inductor is connected across a 60-Hz ac source whose voltage amplitude is 50 V.

- (a) What is the maximum current through the inductor?
- (b) What is the phase relationship between the current through and the potential difference across the inductor?

54. What is the impedance of an **RLC** series circuit at the resonant frequency?

55. What is the resistance **R** in the circuit shown below if the amplitude of the ac through the inductor is 4.24 A?



56. An ac source of voltage amplitude 100 V and frequency 1.0 kHz drives an **RLC** series circuit with **R=20Ω**, **L=4.0mH**, and **C=50μF**.

- (a) Determine the rms current through the circuit.
- (b) What are the rms voltages across the three elements?
- (c) What is the phase angle between the emf and the current?
- (d) What is the power output of the source?
- (e) What is the power dissipated in the resistor?

57. In an RLC series circuit, **R = 200Ω**, **L = 1.0H**, **C = 50μF**, **V₀ = 120V**, and **f = 50Hz**. What is the power output of the source?

58. A power plant generator produces 100 A at 15 kV (rms). A transformer is used to step up the transmission line voltage to 150 kV (rms).

- (a) What is rms current in the transmission line?
- (b) If the resistance per unit length of the line is **$8.6 \times 10^{-8} \Omega/m$** , what is the power loss per meter in the line?
- (c) What would the power loss per meter be if the line voltage were 15 kV (rms)?

59. Consider a power plant located 25 km outside a town delivering 50 MW of power to the town. The transmission lines are made of aluminum cables with a **7cm²** cross-sectional area. Find the loss of power in the transmission lines if it is transmitted at

- (a) 200 kV (rms) and
- (b) 120 V (rms).

60. Neon signs require 12-kV for their operation. A transformer is to be used to change the voltage from 220-V (rms) ac to 12-kV (rms) ac.

- (a) What must the ratio be of turns in the secondary winding to the turns in the primary winding?

- (b) What is the maximum rms current the neon lamps can draw if the fuse in the primary winding goes off at 0.5 A?
- (c) How much power is used by the neon sign when it is drawing the maximum current allowed by the fuse in the primary winding?

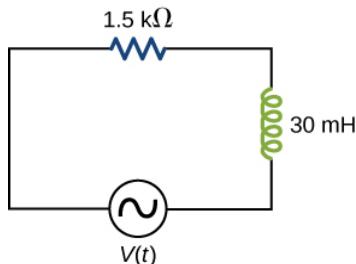
Challenge Problems

61. The 335-kV ac electricity from a power transmission line is fed into the primary winding of a transformer. The ratio of the number of turns in the secondary winding to the number in the primary winding is $N_s/N_p = 1000$.

- (a) What voltage is induced in the secondary winding?
- (b) What is unreasonable about this result?
- (c) Which assumption or premise is responsible?

62. A **1.5-k Ω** resistor and 30-mH inductor are connected in series, as shown below, across a 120-V (rms) ac power source oscillating at 60-Hz frequency.

- (a) Find the current in the circuit.
- (b) Find the voltage drops across the resistor and inductor.
- (c) Find the impedance of the circuit.
- (d) Find the power dissipated in the resistor.
- (e) Find the power dissipated in the inductor.
- (f) Find the power produced by the source.



63. A **20- Ω** resistor, **50- μF** capacitor, and 30-mH inductor are connected in series with an ac source of amplitude 10 V and frequency 125 Hz.

- (a) What is the impedance of the circuit?
- (b) What is the amplitude of the current in the circuit?
- (c) What is the phase constant of the current? Is it leading or lagging the source voltage?
- (d) Write voltage drops across the resistor, capacitor, and inductor and the source voltage as a function of time.
- (e) What is the power factor of the circuit? (f) How much energy is used by the resistor in 2.5 s?

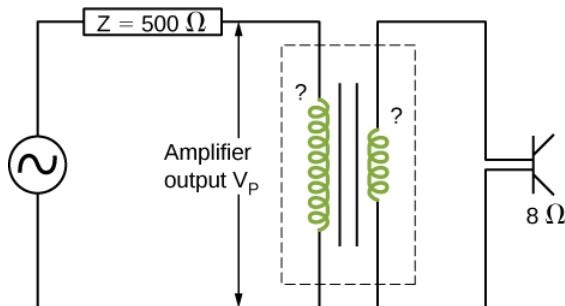
64. A **200- Ω** resistor, **150- μF** capacitor, and 2.5-H inductor are connected in series with an ac source of amplitude 10 V and variable angular frequency ω .

- (a) What is the value of the resonance frequency ω_R ?
- (b) What is the amplitude of the current if $\omega = \omega_R$?
- (c) What is the phase constant of the current when $\omega = \omega_R$? Is it leading or lagging the source voltage, or is it in phase?
- (d) Write an equation for the voltage drop across the resistor as a function of time when $\omega = \omega_R$.
- (e) What is the power factor of the circuit when $\omega = \omega_R$?
- (f) How much energy is used up by the resistor in 2.5 s when $\omega = \omega_R$?

65. Find the reactances of the following capacitors and inductors in ac circuits with the given frequencies in each case:

- (a) 2-mH inductor with a frequency 60-Hz of the ac circuit;
- (b) 2-mH inductor with a frequency 600-Hz of the ac circuit;
- (c) 20-mH inductor with a frequency 6-Hz of the ac circuit;
- (d) 20-mH inductor with a frequency 60-Hz of the ac circuit;
- (e) 2-mF capacitor with a frequency 60-Hz of the ac circuit; and
- (f) 2-mF capacitor with a frequency 600-Hz of the AC circuit.

66. An output impedance of an audio amplifier has an impedance of **500Ω** and has a mismatch with a low-impedance **8-Ω** loudspeaker. You are asked to insert an appropriate transformer to match the impedances. What turns ratio will you use, and why? Use the simplified circuit shown below.



67. Show that the SI unit for capacitive reactance is the ohm. Show that the SI unit for inductive reactance is also the ohm.

68. A coil with a self-inductance of 16 mH and a resistance of **6.0Ω** is connected to an ac source whose frequency can be varied. At what frequency will the voltage across the coil lead the current through the coil by **45°**?

69. An RLC series circuit consists of a **50-Ω** resistor, a **200-μF** capacitor, and a 120-mH inductor whose coil has a resistance of **20Ω**. The source for the circuit has an rms emf of 240 V at a frequency of 60 Hz. Calculate the rms voltages across the

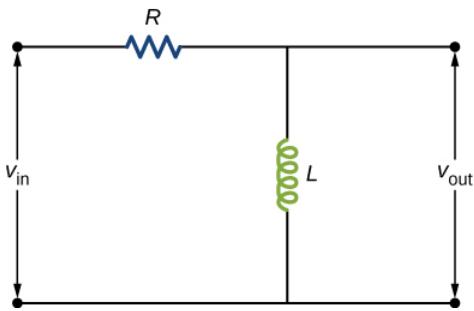
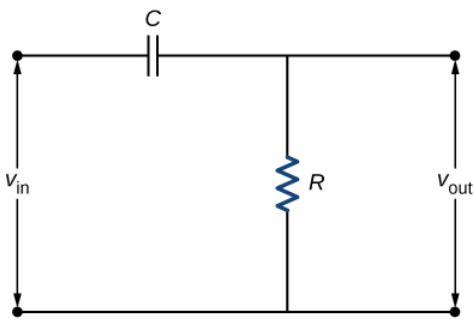
- (a) resistor,
- (b) capacitor, and
- (c) inductor.

70. An RLC series circuit consists of a **10-Ω** resistor, an **8.0-μF** capacitor, and a 50-mH inductor. A 110-V (rms) source of variable frequency is connected across the combination. What is the power output of the source when its frequency is set to one-half the resonant frequency of the circuit?

71. Shown below are two circuits that act as crude high-pass filters. The input voltage to the circuits is v_{in} , and the output voltage is v_{out} .

(a) Show that for the capacitor circuit, $\frac{v_{out}}{v_{in}} = \frac{1}{\sqrt{1 + 1/\omega^2 R^2 C^2}}$, and for the inductor circuit, $\frac{v_{out}}{v_{in}} = \frac{\omega L}{\sqrt{R^2 + \omega^2 L^2}}$.

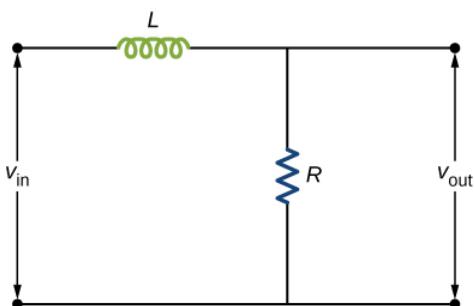
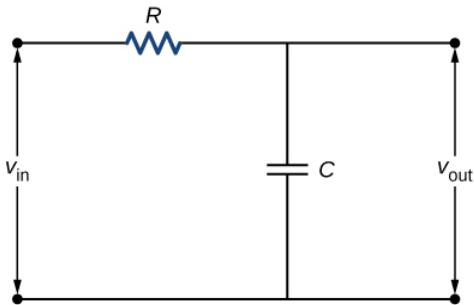
(b) Show that for high frequencies, $v_{out} \approx v_{in}$, but for low frequencies, $v_{out} \approx 0$.



72. The two circuits shown below act as crude low-pass filters. The input voltage to the circuits is v_{in} , and the output voltage is v_{out} .

(a) Show that for the capacitor circuit, $\frac{v_{out}}{v_{in}} = \frac{1}{\sqrt{1 + \omega^2 R^2 C^2}}$, and for the inductor circuit, $\frac{v_{out}}{v_{in}} = \frac{R}{\sqrt{R^2 + \omega^2 L^2}}$.

(b) Show that for low frequencies, $v_{out} \approx v_{in}$, but for high frequencies, $v_{out} \approx 0$.



This page titled [19.9: Alternating-Current Circuits \(Exercise\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [19.9: Alternating-Current Circuits \(Exercise\)](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

19.10: Alternating-Current Circuits (Answers)

Check Your Understanding

15.1. 10 ms

15.2. a. $(20V)\sin 200\pi t$, $(0.20A)\sin 200\pi t$;

b. $(20V)\sin 200\pi t$, $(0.13A)\sin(200\pi t + \pi/2)$;

c. $(20V)\sin 200\pi t$, $(2.14A)\sin(200\pi t - \pi/2)$

15.3. $v_R = (V_0R/Z)\sin(\omega t - \phi)$; $v_C = (V_0X_C/Z)\sin(\omega t - \phi + \pi/2) = -(V_0X_C/Z)\cos(\omega t - \phi)$; $v_L = (V_0X_L/Z)\sin(\omega t - \phi - \pi/2) = (V_0X_L/Z)\cos(\omega t - \phi)$

15.4. $v(t) = (10.0V)\sin 90\pi t$

15.5. 2.00 V; 10.01 V; 8.01 V

15.6. a. 160 Hz;

b. 40Ω ;

c. $(0.25A)\sin 10^3 t$;

d. 0.023 rad

15.7. a. halved;

b. halved;

c. same

15.8. $v(t) = (0.14V)\sin(4.0 \times 10^2 t)$

15.9. a. 12:1;

b. 0.042 A;

c. $2.6 \times 10^3\Omega$

Conceptual Questions

1. Angular frequency is 2π times frequency.

3. yes for both

5. The instantaneous power is the power at a given instant. The average power is the power averaged over a cycle or number of cycles.

7. The instantaneous power can be negative, but the power output can't be negative.

9. There is less thermal loss if the transmission lines operate at low currents and high voltages.

11. The adapter has a step-down transformer to have a lower voltage and possibly higher current at which the device can operate.

13. so each loop can experience the same changing magnetic flux

Problems

15. a. 530Ω ;

b. 53Ω ;

c. 5.3Ω

17. a. 1.9Ω ;

b. 19Ω ;

c. 190Ω

19. 360 Hz

21. $i(t) = (3.2A)\sin(120\pi t)$

23. a. 38Ω ;

b. $i(t) = (4.24A)\sin(120\pi t - \pi/2)$

25. a. 770Ω ;

b. 0.16 A;

c. $I = (0.16A)\cos(120\pi t)$;

d. $v_R = 120\cos(120\pi t)$; $v_C = 120\cos(120\pi t - \pi/2)$

27. a. 690Ω ;

b. 0.15 A;

c. $I = (0.15A)\sin(1000\pi t - 0.753)$;

d. 1100Ω , 0.092 A, $I = (0.092A)\sin(1000\pi t + 1.09)$

29. a. 5.7Ω ;

b. 29° ;

c. $I = (30.4)\cos(120\pi t)$

31. a. 0.89 A;

b. 5.6A;

c. 1.4 A

33. a. 5.3 W;

b. 2.1 W

35. a. inductor;

b. $X_L = 52\Omega$

37. $1.3 \times 10^{-7}F$

39. a. 820 Hz;

b. 7.8

41. a. 50 Hz;

b. 50 W;

c. 13;

d. 25 rad/s

43. The reactance of the capacitor is larger than the reactance of the inductor because the current leads the voltage. The power usage is 30 W.

45. a. 45:1;

b. 0.68 A, 0.015 A;

c. 160Ω

47. a. 41 turns;

b. 40.9 mA

Additional Problems

49. a. $i(t) = (1.26A)\sin(200\pi t + \pi/2)$;

b. $i(t) = (12.6A)\sin(200\pi t - \pi/2)$;

c. $i(t) = (2A)\sin(200\pi t)$

51. a. $2.5 \times 10^3 \Omega, 3.6 \times 10^{-3} A$;

b. $7.5\Omega, 1.2A$

53. a. 19 A;

b. inductor leads by 90°

55. 11.7Ω

57. 36 W

59. a. $5.9 \times 10^4 W$;

b. $1.64 \times 10^{11} W$

Challenge Problems

61. a. 335 MV;

b. the result is way too high, well beyond the breakdown voltage of air over reasonable distances;

c. the input voltage is too high

63. a. 20Ω ;

b. 0.5 A;

c. 5.4° , lagging;

d. $V_R = (9.96V)\cos(250\pi t + 5.4^\circ), V_C = (12.7V)\cos(250\pi t + 5.4^\circ - 90^\circ), V_L = (11.8V)\cos(250\pi t + 5.4^\circ + 90^\circ), V_{source} = (10.0V)\cos(250\pi t);$

e. 0.995;

f. 6.25 J

65. a. 0.75Ω ;

b. 7.5Ω ;

c. 0.75Ω ;

d. 7.5Ω ;

e. 1.3Ω ;

f. 0.13Ω

67. The units as written for inductive reactance Equation 15.8 are $\frac{rad}{s} H$. Radians can be ignored in unit analysis. The Henry can be defined as $H = \frac{V \cdot s}{A} = \Omega \cdot s$. Combining these together results in a unit of Ω for reactance.

69. a. 156 V;

b. 42 V;

c. 154 V

71. a. $\frac{v_{out}}{v_{in}} = \frac{1}{\sqrt{1 + 1/\omega^2 R^2 C^2}}$ and $\left(\frac{v_{out}}{v_{in}}\right)^2 = \frac{1}{R^2 + \omega^2 C^2}$

- b. $v_{out} \approx v_{in}$ and $v_{out} \approx 0$

This page titled [19.10: Alternating-Current Circuits \(Answers\)](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **15.10: Alternating-Current Circuits (Answers)** by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

CHAPTER OVERVIEW

20: Maxwell's Equations

- 20.1: Introduction
- 20.2: Electric Flux
- 20.3: Gauss's Law
- 20.4: Ampère's Law
- 20.5: Maxwell's Equations and Electromagnetic Waves
- 20.6: Plane Electromagnetic Waves
- 20.7: Momentum and Radiation Pressure

20: Maxwell's Equations is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

20.1: Introduction

Maxwell's equations are the fundamental equations that, along with the Lorentz force law, describe classical electrodynamics [1]. It provides a basis for electric and magnetic circuits as well as classical optics. They are named for James Clerk Maxwell [2], who, in 1861 and 1862, published an early form of the equations that included the Lorentz force law. This chapter describes the four equations, which are usually called Gauss's Law, Gauss's Law for Magnetism, Faraday's Law, and the Ampère-Maxwell Law. It then describes some of the implications of these equations, including plane electromagnetic waves and momentum and radiation pressure.

References

1. Wikipedia contributors. [Maxwell's equations](#) [Internet]. Wikipedia, The Free Encyclopedia.
2. Wikipedia contributors. [James Clerk Maxwell](#) [Internet]. Wikipedia, The Free Encyclopedia.

20.1: Introduction is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Ronald Kumon.

20.2: Electric Flux

Learning Objectives

By the end of this section, you will be able to:

- Define the concept of flux
- Describe electric flux
- Calculate electric flux for a given situation

The concept of **flux** describes how much of something goes through a given area. More formally, it is the dot product of a vector field (in this chapter, the electric field) with an area. You may conceptualize the flux of an electric field as a measure of the number of electric field lines passing through an area (Figure 20.2.1). The larger the area, the more field lines go through it and, hence, the greater the flux; similarly, the stronger the electric field is (represented by a greater density of lines), the greater the flux. On the other hand, if the area rotated so that the plane is aligned with the field lines, none will pass through and there will be no flux.

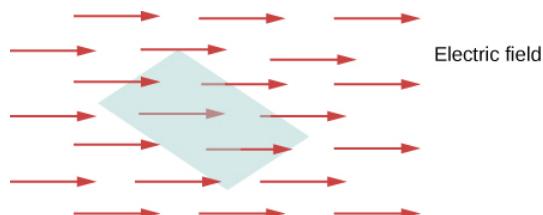


Figure 20.2.1: The flux of an electric field through the shaded area captures information about the “number” of electric field lines passing through the area. The numerical value of the electric flux depends on the magnitudes of the electric field and the area, as well as the relative orientation of the area with respect to the direction of the electric field.

A macroscopic analogy that might help you imagine this is to put a hula hoop in a flowing river. As you change the angle of the hoop relative to the direction of the current, more or less of the flow will go through the hoop. Similarly, the amount of flow through the hoop depends on the strength of the current and the size of the hoop. Again, flux is a general concept; we can also use it to describe the amount of sunlight hitting a solar panel or the amount of energy a telescope receives from a distant star, for example.

To quantify this idea, Figure 20.2.1a shows a planar surface S_1 of area A_1 that is perpendicular to the uniform electric field $\vec{E} = E\hat{j}$. If N field lines pass through S_1 , then we know from the definition of electric field lines (Electric Charges and Fields) that $N/A \propto E$, or $N \propto EA_1$.

The quantity EA_1 is the **electric flux** through S_1 . We represent the electric flux through an open surface like S_1 by the symbol Φ . Electric flux is a scalar quantity and has an SI unit of newton-meters squared per coulomb ($N \cdot m^2/C$). Notice that $N \propto EA_1$ may also be written as $N \propto \Phi$, demonstrating that *electric flux is a measure of the number of field lines crossing a surface*.

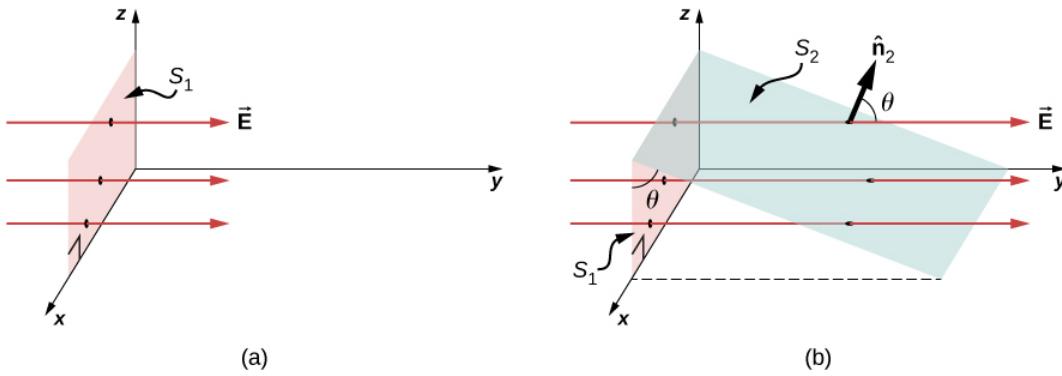


Figure 20.2.2: (a) A planar surface S_1 of area A_1 is perpendicular to the electric field $E\hat{j}$. N field lines cross surface S_1 . (b) A surface S_2 of area A_2 whose projection onto the xz -plane is S_1 . The same number of field lines cross each surface.

Now consider a planar surface that is not perpendicular to the field. How would we represent the electric flux? Figure 20.2.2b shows a surface S_2 of area A_2 that is inclined at an angle θ to the xz -plane and whose projection in that plane is S_1 (area A_1). The areas are

related by $A_2 \cos \theta = A_1$. Because the same number of field lines crosses both S_1 and S_2 , the fluxes through both surfaces must be the same. The flux through S_2 is therefore $\Phi = EA_1 = EA_2 \cos \theta$. Designating \hat{n}_2 as a unit vector normal to S_2 (see Figure 20.2.2b), we obtain

$$\Phi = \vec{E} \cdot \hat{n}_2 A_2.$$

Note

Check out this [video](#) to observe what happens to the flux as the area changes in size and angle, or the electric field changes in strength.

Area Vector

For discussing the flux of a vector field, it is helpful to introduce an area vector \vec{A} . This allows us to write the last equation in a more compact form. What should the magnitude of the area vector be? What should the direction of the area vector be? What are the implications of how you answer the previous question?

The **area vector** of a flat surface of area A has the following magnitude and direction:

- Magnitude is equal to area (A)
- Direction is along the normal to the surface (\hat{n}); that is, perpendicular to the surface.

Since the normal to a flat surface can point in either direction from the surface, the direction of the area vector of an open surface needs to be chosen, as shown in Figure 20.2.3.

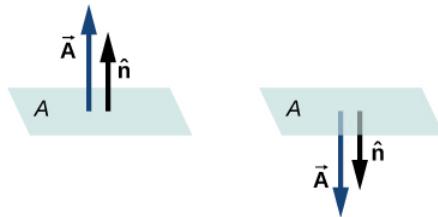


Figure 20.2.3: The direction of the area vector of an open surface needs to be chosen; it could be either of the two cases displayed here. The area vector of a part of a closed surface is defined to point from the inside of the closed space to the outside. This rule gives a unique direction.

Since \hat{n} is a unit normal to a surface, it has two possible directions at every point on that surface (Figure 20.2.1a). For an open surface, we can use either direction, as long as we are consistent over the entire surface. 20.2.1c of the figure shows several cases.

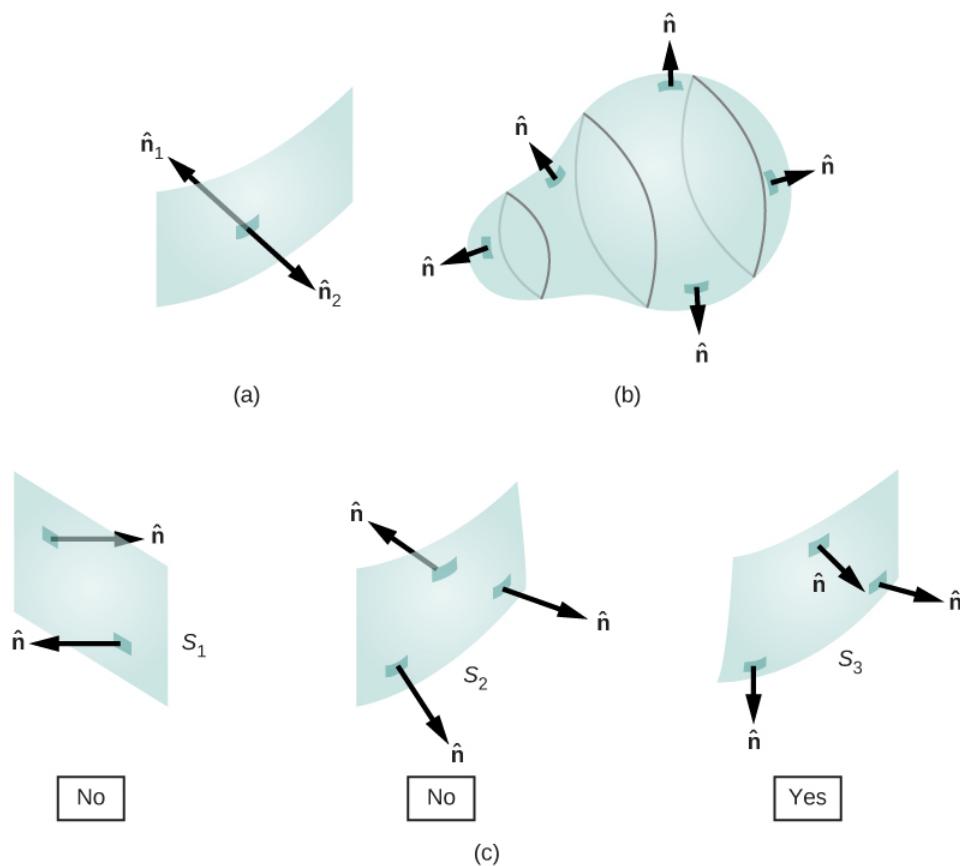


Figure 20.2.4: (a) Two potential normal vectors arise at every point on a surface. (b) The outward normal is used to calculate the flux through a closed surface. (c) Only S_3 has been given a consistent set of normal vectors that allows us to define the flux through the surface.

However, if a surface is closed, then the surface encloses a volume. In that case, the direction of the normal vector at any point on the surface points from the inside to the outside. On a *closed surface* such as that of Figure 20.2.1b, \hat{n} is chosen to be the *outward normal* at every point, to be consistent with the sign convention for electric charge.

Electric Flux

Now that we have defined the area vector of a surface, we can define the electric flux of a uniform electric field through a flat area as the scalar product of the electric field and the area vector:

$$\Phi = \vec{E} \cdot \vec{A} \text{ (uniform } \vec{E}, \text{ flat surface).}$$

Figure 20.2.5 shows the electric field of an oppositely charged, parallel-plate system and an imaginary box between the plates. The electric field between the plates is uniform and points from the positive plate toward the negative plate. A calculation of the flux of this field through various faces of the box shows that the net flux through the box is zero. Why does the flux cancel out here?

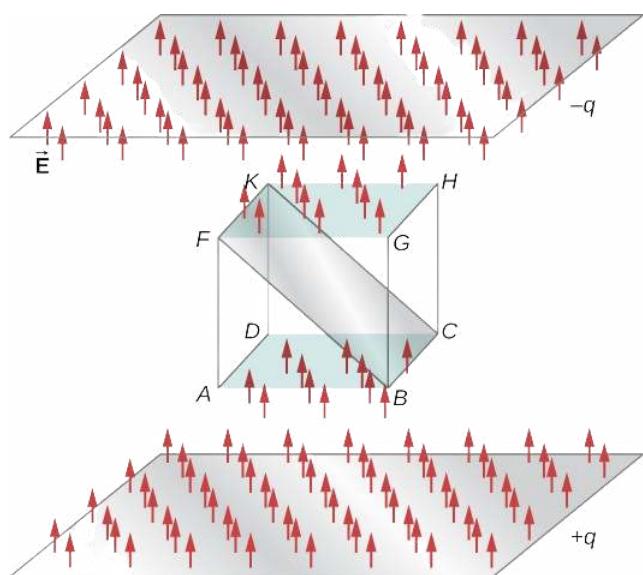


Figure 20.2.5: Electric flux through a cube, placed between two charged plates. Electric flux through the bottom face ($ABCD$) is negative, because \vec{E} is in the opposite direction to the normal to the surface. The electric flux through the top face ($FGHK$) is positive, because the electric field and the normal are in the same direction. The electric flux through the other faces is zero, since the electric field is perpendicular to the normal vectors of those faces. The net electric flux through the cube is the sum of fluxes through the six faces. Here, the net flux through the cube is equal to zero. The magnitude of the flux through rectangle $BCKF$ is equal to the magnitudes of the flux through both the top and bottom faces.

The reason is that the sources of the electric field are outside the box. Therefore, if any electric field line enters the volume of the box, it must also exit somewhere on the surface because there is no charge inside for the lines to land on. Therefore, quite generally, electric flux through a closed surface is zero if there are no sources of electric field, whether positive or negative charges, inside the enclosed volume. In general, when field lines leave (or “flow out of”) a closed surface, Φ is positive; when they enter (or “flow into”) the surface, Φ is negative.

Any smooth, non-flat surface can be replaced by a collection of tiny, approximately flat surfaces, as shown in Figure 20.2.6. If we divide a surface S into small patches, then we notice that, as the patches become smaller, they can be approximated by flat surfaces. This is similar to the way we treat the surface of Earth as locally flat, even though we know that globally, it is approximately spherical.

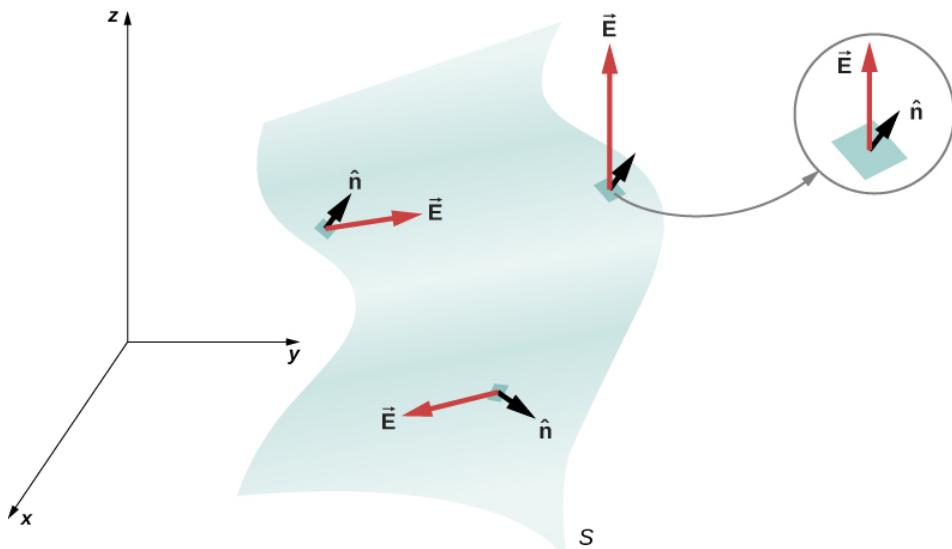


Figure 20.2.6: A surface is divided into patches to find the flux.

To keep track of the patches, we can number them from 1 through N . Now, we define the area vector for each patch as the area of the patch pointed in the direction of the normal. Let us denote the area vector for the i th patch by $\delta \vec{A}_i$. (We have used the symbol δ

to remind us that the area is of an arbitrarily small patch.) With sufficiently small patches, we may approximate the electric field over any given patch as uniform. Let us denote the average electric field at the location of the i th patch by \vec{E}_i .

$$\vec{E}_i = \text{average electric field over the } i\text{th patch}.$$

Therefore, we can write the electric flux Φ through the area of the i th patch as

$$\Phi_i = \vec{E}_i \cdot \delta \vec{A}_i \quad (\text{ith patch}).$$

The flux through each of the individual patches can be constructed in this manner and then added to give us an estimate of the net flux through the entire surface S , which we denote simply as Φ .

$$\Phi = \sum_{i=1}^N \Phi_i = \sum_{i=1}^N \vec{E}_i \cdot \delta \vec{A}_i \quad (N \text{ patch estimate}).$$

This estimate of the flux gets better as we decrease the size of the patches. However, when you use smaller patches, you need more of them to cover the same surface. In the limit of infinitesimally small patches, they may be considered to have area dA and unit normal \hat{n} . Since the elements are infinitesimal, they may be assumed to be planar, and \vec{E}_i may be taken as constant over any element. Then the flux $d\Phi$ through an area dA is given by $d\Phi = \vec{E} \cdot \hat{n} dA$. It is positive when the angle between \vec{E}_i and \hat{n} is less than 90° and negative when the angle is greater than 90° . The net flux is the sum of the infinitesimal flux elements over the entire surface. With infinitesimally small patches, you need infinitely many patches, and the limit of the sum becomes a surface integral. With \int_S representing the integral over S ,

$$\Phi = \int_S \vec{E} \cdot \hat{n} dA = \int_S \vec{E} \cdot d\vec{A} \quad (\text{open surface}).$$

In practical terms, surface integrals are computed by taking the antiderivatives of both dimensions defining the area, with the edges of the surface in question being the bounds of the integral.

To distinguish between the flux through an open surface like that of Figure 20.2.2 and the flux through a closed surface (one that completely bounds some volume), we represent flux through a closed surface by

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \oint_S \vec{E} \cdot d\vec{A} \quad (\text{closed surface})$$

where the circle through the integral symbol simply means that the surface is closed, and we are integrating over the entire thing. If you only integrate over a portion of a closed surface, that means you are treating a subset of it as an open surface.

✓ Example 20.2.1: Flux of a Uniform Electric Field

A constant electric field of magnitude E_0 points in the direction of the positive z -axis (Figure 20.2.7). What is the electric flux through a rectangle with sides a and b in the (a) xy -plane and in the (b) xz -plane?

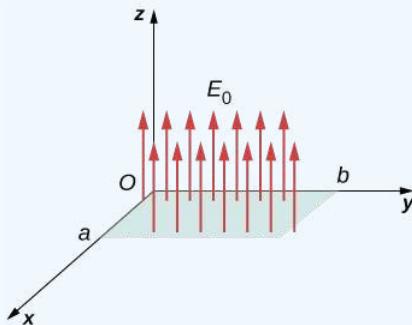


Figure 20.2.7: Calculating the flux of E_0 through a rectangular surface.

Strategy

Apply the definition of flux: $\Phi = \vec{E} \cdot \vec{A}$ (uniform \vec{E}), where the definition of dot product is crucial.

Solution

- In this case, $\Phi = \vec{E}_0 \cdot \vec{A} = E_0 A = E_0 ab$.
- Here, the direction of the area vector is either along the positive y -axis or toward the negative y -axis. Therefore, the scalar product of the electric field with the area vector is zero, giving zero flux.

Significance

The relative directions of the electric field and area can cause the flux through the area to be zero.

✓ Flux of a Uniform Electric Field through a Closed Surface

A constant electric field of magnitude E_0 points in the direction of the positive z -axis (Figure 20.2.8). What is the net electric flux through a cube?

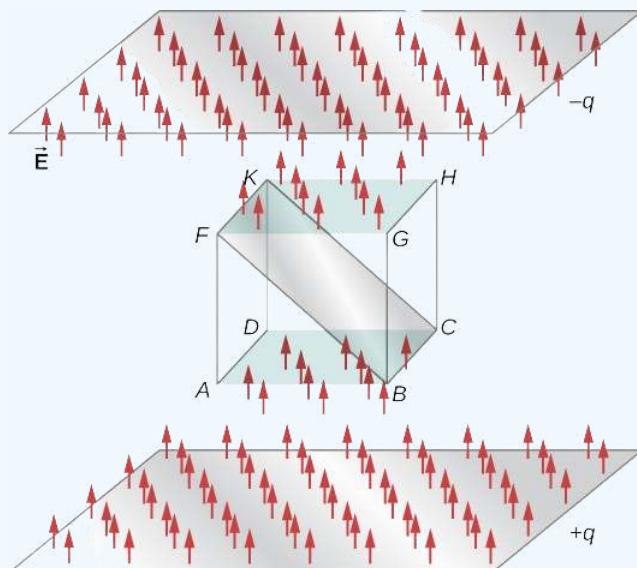


Figure 20.2.8: Calculating the flux of E_0 through a closed cubic surface.

Strategy

Apply the definition of flux: $\Phi = \vec{E} \cdot \vec{A}$ (uniform \vec{E}), noting that a closed surface eliminates the ambiguity in the direction of the area vector.

Solution

Through the top face of the cube $\Phi = \vec{E}_0 \cdot \vec{A} = E_0 A$.

Through the bottom face of the cube, $\Phi = \vec{E}_0 \cdot \vec{A} = -E_0 A$, because the area vector here points downward.

Along the other four sides, the direction of the area vector is perpendicular to the direction of the electric field. Therefore, the scalar product of the electric field with the area vector is zero, giving zero flux.

The net flux is $\Phi_{net} = E_0 A - E_0 A + 0 + 0 + 0 + 0 = 0$.

Significance

The net flux of a uniform electric field through a closed surface is zero.

✓ Example 20.2.3: Electric Flux through a Plane, Integral Method

A uniform electric field \vec{E} of magnitude 10 N/C is directed parallel to the yz -plane at 30° above the xy -plane, as shown in Figure 20.2.9. What is the electric flux through the plane surface of area 6.0 m^2 located in the xz -plane? Assume that \hat{n} points in the positive y -direction.

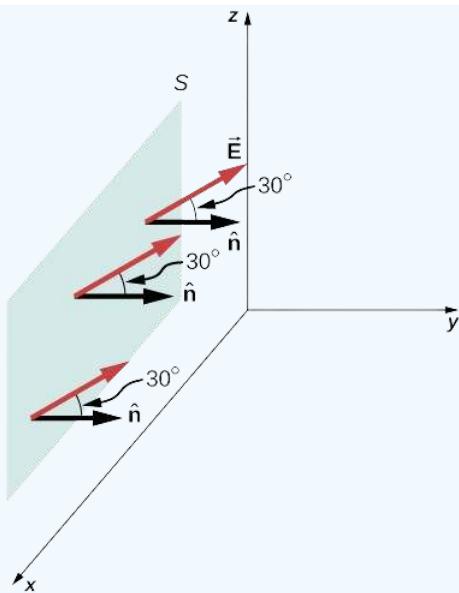


Figure 20.2.9: The electric field produces a net electric flux through the surface S .

Strategy

Apply $\Phi = \int_S \vec{E} \cdot \hat{n} dA$, where the direction and magnitude of the electric field are constant.

Solution

The angle between the uniform electric field \vec{E} and the unit normal \hat{n} to the planar surface is 30° . Since both the direction and magnitude are constant, E comes outside the integral. All that is left is a surface integral over dA , which is A . Therefore, using the open-surface equation, we find that the electric flux through the surface is

$$\begin{aligned}\Phi &= \int_S \vec{E} \cdot \hat{n} dA = EA \cos \theta \\ &= (10 \text{ N/C})(6.0 \text{ m}^2)(\cos 30^\circ) = 52 \text{ N} \cdot \text{m}^2/\text{C}.\end{aligned}$$

Significance

Again, the relative directions of the field and the area matter, and the general equation with the integral will simplify to the simple dot product of area and electric field.

Exercise 20.2.1

What angle should there be between the electric field and the surface shown in Figure 20.2.9 in the previous example so that no electric flux passes through the surface?

Answer

Place it so that its unit normal is perpendicular to \vec{E} .

Example 20.2.4 : Inhomogeneous Electric Field

What is the total flux of the electric field $\vec{E} = cy^2 \hat{k}$ through the rectangular surface shown in Figure 20.2.10?

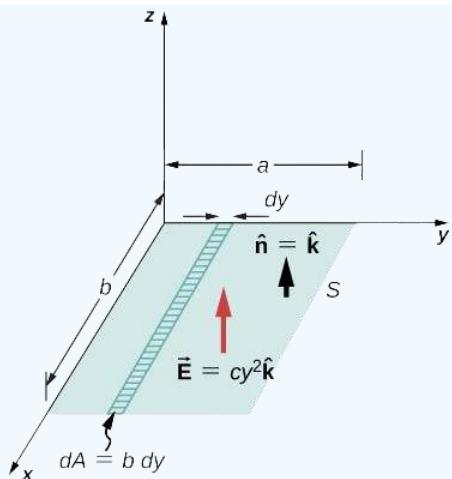


Figure 20.2.10: Since the electric field is not constant over the surface, an integration is necessary to determine the flux.

Strategy

Apply $\Phi = \int_S \vec{E} \cdot \hat{n} dA$. We assume that the unit normal \hat{n} to the given surface points in the positive z -direction, so $\hat{n} = \hat{k}$. Since the electric field is not uniform over the surface, it is necessary to divide the surface into infinitesimal strips along which \vec{E} is essentially constant. As shown in Figure 20.2.10, these strips are parallel to the x -axis, and each strip has an area $dA = b dy$.

Solution

From the open surface integral, we find that the net flux through the rectangular surface is

$$\begin{aligned}\Phi &= \int_S \vec{E} \cdot \hat{n} dA = \int_0^a (cy^2 \hat{k}) \cdot \hat{k} (b dy) \\ &= cb \int_0^a y^2 dy = \frac{1}{3} a^3 bc.\end{aligned}$$

Significance

For a non-constant electric field, the integral method is required.

Exercise 20.2.2

If the electric field in Example 20.2.4 is $\vec{E} = mx\hat{k}$, what is the flux through the rectangular area?

Answer

$$mab^2/2$$

This page titled [20.2: Electric Flux](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.2: Electric Flux](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

20.3: Gauss's Law

Learning Objectives

By the end of this section, you will be able to:

- State Gauss's law
- Explain the conditions under which Gauss's law may be used
- Apply Gauss's law in appropriate systems

We can now determine the electric flux through an arbitrary closed surface due to an arbitrary charge distribution. We found that if a closed surface does not have any charge inside where an electric field line can terminate, then any electric field line entering the surface at one point must necessarily exit at some other point of the surface. Therefore, if a closed surface does not have any charges inside the enclosed volume, then the electric flux through the surface is zero. Now, what happens to the electric flux if there are some charges inside the enclosed volume? Gauss's law gives a quantitative answer to this question.

To get a feel for what to expect, let's calculate the electric flux through a spherical surface around a positive point charge q , since we already know the electric field in such a situation. Recall that when we place the point charge at the origin of a coordinate system, the electric field at a point P that is at a distance r from the charge at the origin is given by

$$\vec{E}_P = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r},$$

where \hat{r} is the radial vector from the charge at the origin to the point P . We can use this electric field to find the flux through the spherical surface of radius r , as shown in Figure 20.3.1.

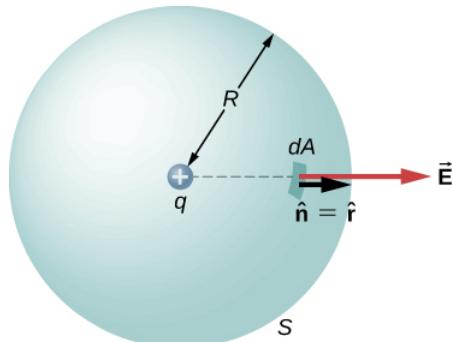


Figure 20.3.1: A closed spherical surface surrounding a point charge q .

Then we apply $\Phi = \int_S \vec{E} \cdot \hat{n} dA$ to this system and substitute known values. On the sphere, \hat{n} and $r = R$ so for an infinitesimal area dA ,

$$\begin{aligned} d\Phi &= \vec{E} \cdot \hat{n} dA \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} \hat{r} \cdot \hat{r} dA \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} dA. \end{aligned}$$

We now find the net flux by integrating this flux over the surface of the sphere:

$$\Phi = \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} \oint_S dA = \frac{1}{4\pi\epsilon_0} \frac{q}{R^2} (4\pi R^2) = \frac{q}{\epsilon_0}.$$

where the total surface area of the spherical surface is $4\pi R^2$. This gives the flux through the closed spherical surface at radius r as

$$\Phi = \frac{q}{\epsilon_0}.$$

A remarkable fact about this equation is that the flux is independent of the size of the spherical surface. This can be directly attributed to the fact that the electric field of a point charge decreases as $1/r^2$ with distance, which just cancels the r^2 rate of increase of the surface area.

Electric Field Lines Picture

An alternative way to see why the flux through a closed spherical surface is independent of the radius of the surface is to look at the electric field lines. Note that every field line from q that pierces the surface at radius R_1 also pierces the surface at R_2 (Figure 20.3.2).

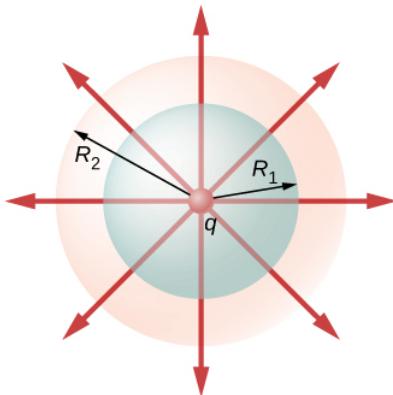


Figure 20.3.2: Flux through spherical surfaces of radii R_1 and R_2 enclosing a charge q are equal, independent of the size of the surface, since all E -field lines that pierce one surface from the inside to outside direction also pierce the other surface in the same direction.

Therefore, the net number of electric field lines passing through the two surfaces from the inside to outside direction is equal. This net number of electric field lines, which is obtained by subtracting the number of lines in the direction from outside to inside from the number of lines in the direction from inside to outside gives a visual measure of the electric flux through the surfaces.

You can see that if no charges are included within a closed surface, then the electric flux through it must be zero. A typical field line enters the surface at dA_1 and leaves at dA_2 . Every line that enters the surface must also leave that surface. Hence the net “flow” of the field lines into or out of the surface is zero (Figure 20.3.3a). The same thing happens if charges of equal and opposite sign are included inside the closed surface, so that the total charge included is zero (Figure 20.3.3b). A surface that includes the same amount of charge has the same number of field lines crossing it, regardless of the shape or size of the surface, as long as the surface encloses the same amount of charge (Figure 20.3.3c).

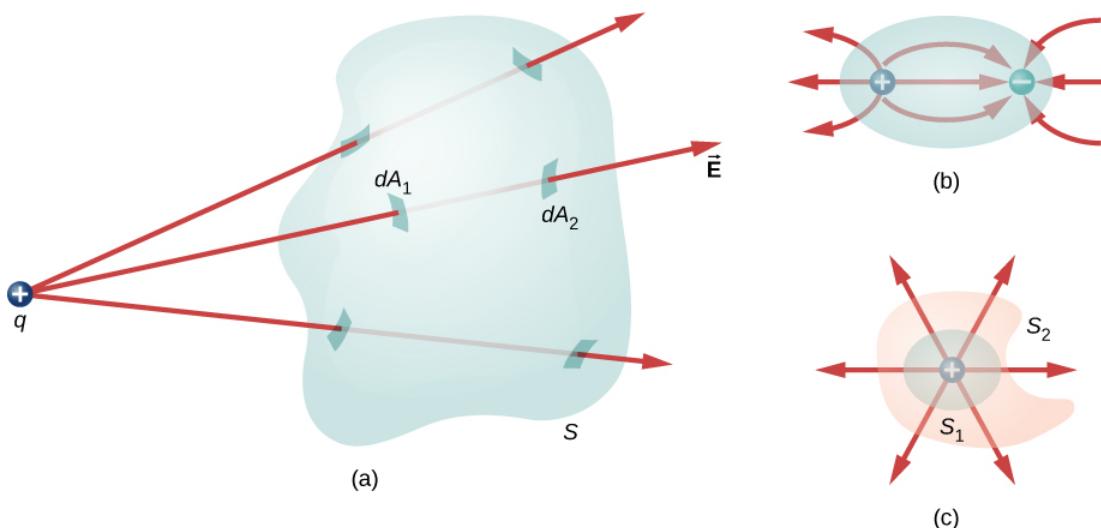


Figure 20.3.3: Understanding the flux in terms of field lines. (a) The electric flux through a closed surface due to a charge outside that surface is zero. (b) Charges are enclosed, but because the net charge included is zero, the net flux through the closed surface is also zero. (c) The shape and size of the surfaces that enclose a charge does not matter because all surfaces enclosing the same charge have the same flux.

Statement of Gauss's Law

Gauss's law generalizes this result to the case of any number of charges and any location of the charges in the space inside the closed surface. According to Gauss's law, the flux of the electric field \vec{E} through any closed surface, also called a **Gaussian surface**, is equal to the net charge enclosed (q_{enc}) divided by the permittivity of free space (ϵ_0):

$$\Phi_{\text{Closed Surface}} = \frac{q_{enc}}{\epsilon_0}.$$

This equation holds for *charges of either sign*, because we define the area vector of a closed surface to point outward. If the enclosed charge is negative (Figure 20.3.4b), then the flux through either S or S' is negative.

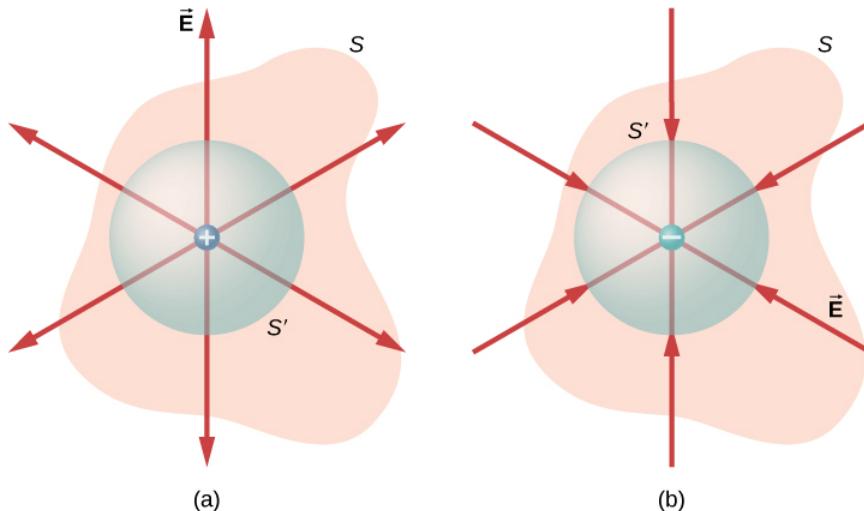


Figure 20.3.4: The electric flux through any closed surface surrounding a point charge q is given by Gauss's law. (a) Enclosed charge is positive. (b) Enclosed charge is negative.

The Gaussian surface does not need to correspond to a real, physical object; indeed, it rarely will. It is a mathematical construct that may be of any shape, provided that it is closed. However, since our goal is to integrate the flux over it, we tend to choose shapes that are highly symmetrical.

If the charges are discrete point charges, then we just add them. If the charge is described by a continuous distribution, then we need to integrate appropriately to find the total charge that resides inside the enclosed volume. For example, the flux through the Gaussian surface S of Figure 20.3.5 is

$$\Phi = (q_1 + q_2 + q_5)/\epsilon_0.$$

Note that q_{enc} is simply the sum of the point charges. If the charge distribution were continuous, we would need to integrate appropriately to compute the total charge within the Gaussian surface.

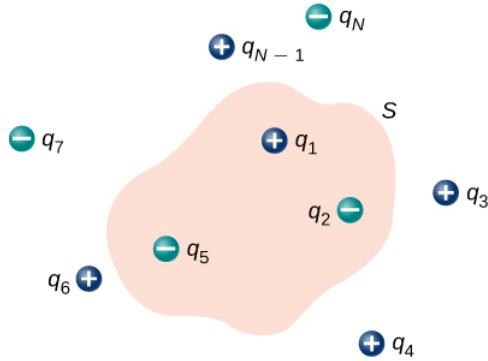


Figure 20.3.5: The flux through the Gaussian surface shown, due to the charge distribution, is $\Phi = (q_1 + q_2 + q_5)/\epsilon_0$.

Recall that the principle of superposition holds for the electric field. Therefore, the total electric field at any point, including those on the chosen Gaussian surface, is the sum of all the electric fields present at this point. This allows us to write Gauss's law in

terms of the total electric field.

Gauss's Law

The flux Φ of the electric field \vec{E} through any closed surface S (a Gaussian surface) is equal to the net charge enclosed (q_{enc}) divided by the permittivity of free space (ϵ_0):

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \frac{q_{enc}}{\epsilon_0}.$$

To use Gauss's law effectively, you must have a clear understanding of what each term in the equation represents. The field \vec{E} is the **total electric field** at every point on the Gaussian surface. This total field includes contributions from charges both inside and outside the Gaussian surface. However, q_{enc} is just the charge **inside** the Gaussian surface. Finally, the Gaussian surface is any closed surface in space. That surface can coincide with the actual surface of a conductor, or it can be an imaginary geometric surface. The only requirement imposed on a Gaussian surface is that it be closed (Figure 20.3.5).



Figure 20.3.6: A **Klein bottle** partially filled with a liquid. Could the Klein bottle be used as a Gaussian surface?

Example 20.3.1: Electric Flux through Gaussian Surfaces

Calculate the electric flux through each Gaussian surface shown in Figure 20.3.7.

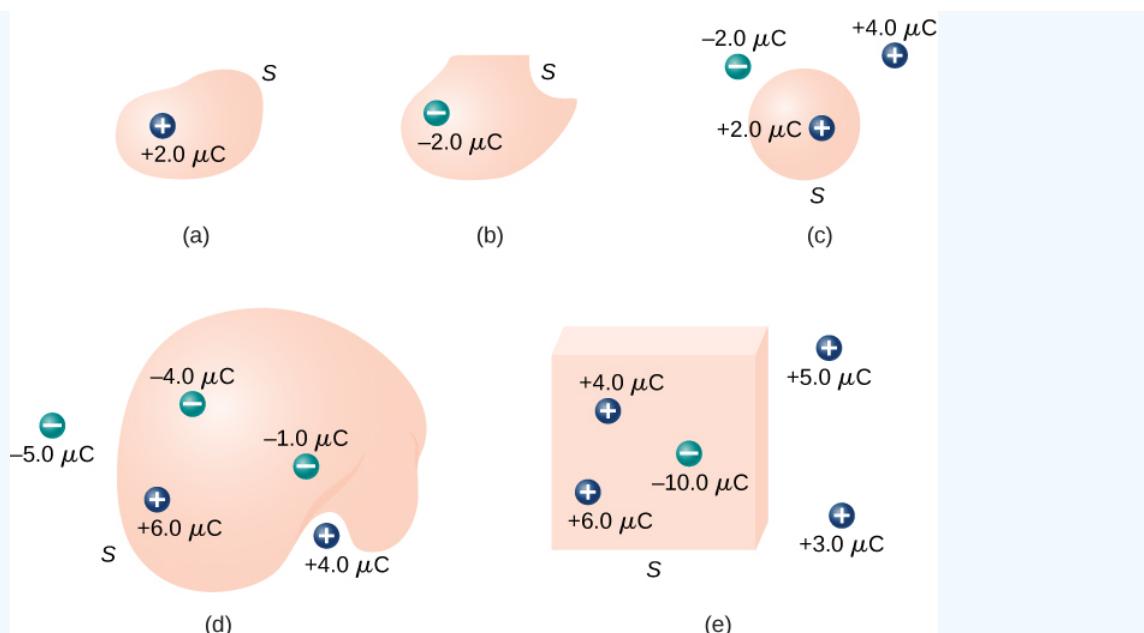


Figure 20.3.7: Various Gaussian surfaces and charges.

Strategy

From Gauss's law, the flux through each surface is given by q_{enc}/ϵ_0 , where q_{enc} is the charge enclosed by that surface.

Solution

For the surfaces and charges shown, we find

a. $\Phi = \frac{2.0 \mu C}{\epsilon_0} = 2.3 \times 10^5 N \cdot m^2/C$.

b. $\Phi = \frac{-2.0 \mu C}{\epsilon_0} = -2.3 \times 10^5 N \cdot m^2/C$.

c. $\Phi = \frac{2.0 \mu C}{\epsilon_0} = 2.3 \times 10^5 N \cdot m^2/C$.

d. $\Phi = \frac{-4.0 \mu C + 6.0 \mu C - 1.0 \mu C}{\epsilon_0} = 1.1 \times 10^5 N \cdot m^2/C$.

e. $\Phi = \frac{4.0 \mu C + 6.0 \mu C - 10.0 \mu C}{\epsilon_0} = 0$.

Significance

In the special case of a closed surface, the flux calculations become a sum of charges. In the next section, this will allow us to work with more complex systems.

Exercise 20.3.1

Calculate the electric flux through the closed cubical surface for each charge distribution shown in Figure 20.3.8.

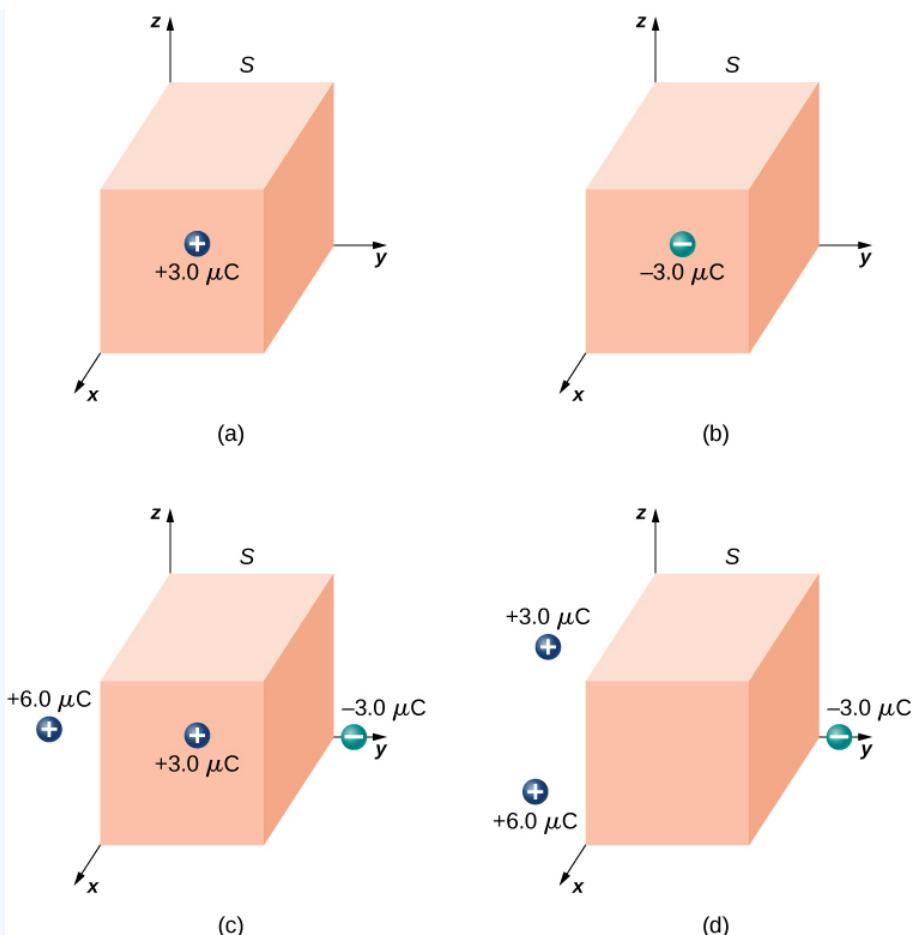


Figure 20.3.8: A cubical Gaussian surface with various charge distributions.

Answer a

$$3.4 \times 10^5 \text{ N} \cdot \text{m}^2/\text{C}$$

Answer b

$$-3.4 \times 10^5 \text{ N} \cdot \text{m}^2/\text{C}$$

Answer c

$$3.4 \times 10^5 \text{ N} \cdot \text{m}^2/\text{C}$$

Answer d

$$0$$

Use this [simulation](#) to adjust the magnitude of the charge and the radius of the Gaussian surface around it. See how this affects the total flux and the magnitude of the electric field at the Gaussian surface.

This page titled [20.3: Gauss's Law](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [6.3: Explaining Gauss's Law](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

20.4: Ampère's Law

Learning Objectives

By the end of this section, you will be able to:

- Explain how Ampère's law relates the magnetic field produced by a current to the value of the current
- Calculate the magnetic field from a long straight wire, either thin or thick, by Ampère's law

A fundamental property of a static magnetic field is that, unlike an electrostatic field, it is not conservative. A conservative field is one that does the same amount of work on a particle moving between two different points regardless of the path chosen. Magnetic fields do not have such a property. Instead, there is a relationship between the magnetic field and its source, electric current. It is expressed in terms of the line integral of \vec{B} and is known as **Ampère's law**. This law can also be derived directly from the Biot-Savart law. We now consider that derivation for the special case of an infinite, straight wire.

Figure 20.4.1 shows an arbitrary plane perpendicular to an infinite, straight wire whose current I is directed out of the page. The magnetic field lines are circles centered on the wire. To begin, let's consider $\oint \vec{B} \cdot d\vec{l}$ over the closed paths **M** and **N**. Notice that one path (**M**) encloses the wire, whereas the other (**N**) does not. Since the field lines are circular, $\vec{B} \cdot d\vec{l}$ is the product of \mathbf{B} and the projection of $d\mathbf{l}$ onto the circle passing through $d\vec{l}$. If the radius of this particular circle is r , the projection is $rd\theta$, and

$$\vec{B} \cdot d\vec{l} = Br d\theta.$$

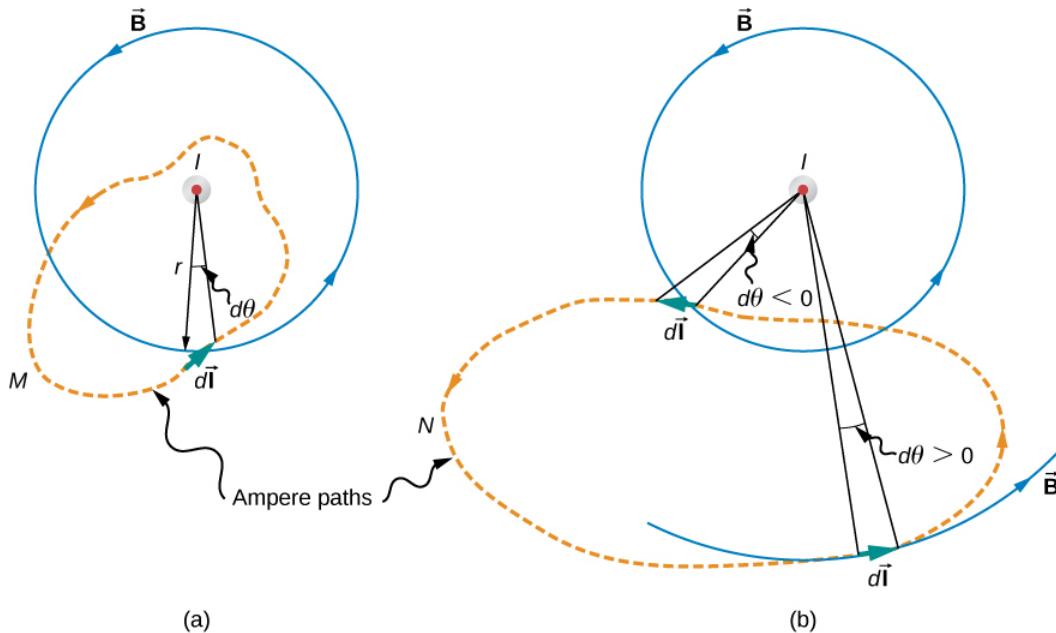


Figure 20.4.1: The current I of a long, straight wire is directed out of the page. The integral $\oint d\theta$ equals 2π and 0, respectively, for paths **M** and **N**.

With \vec{B} given by Equation 12.4.1,

$$\oint \vec{B} \cdot d\vec{l} = \oint \left(\frac{\mu_0 I}{2\pi r} \right) r d\theta = \frac{\mu_0 I}{2\pi} \oint d\theta.$$

For path **M**, which circulates around the wire, $\oint_M d\theta = 2\pi$ and

$$\oint_M \vec{B} \cdot d\vec{l} = \mu_0 I.$$

Path **N**, on the other hand, circulates through both positive (counterclockwise) and negative (clockwise) $d\theta$ (see Figure 20.4.1), and since it is closed, $\oint_N d\theta = 0$. Thus for path **N**,

$$\oint_N \vec{B} \cdot d\vec{l} = 0.$$

The extension of this result to the general case is Ampère's law.

Ampere's Law

Over an arbitrary closed path,

$$\oint \vec{B} \cdot d\vec{l} = \mu_0 I$$

where I is the total current passing through any open surface S whose perimeter is the path of integration. Only currents inside the path of integration need be considered.

To determine whether a specific current I is positive or negative, curl the fingers of your right hand in the direction of the path of integration, as shown in Figure 20.4.1. If I passes through S in the same direction as your extended thumb, I is positive; if I passes through S in the direction opposite to your extended thumb, it is negative.

Problem-Solving Strategy: Ampère's Law

To calculate the magnetic field created from current in wire(s), use the following steps:

1. Identify the symmetry of the current in the wire(s). If there is no symmetry, use the Biot-Savart law to determine the magnetic field.
2. Determine the direction of the magnetic field created by the wire(s) by right-hand rule 2.
3. Choose a path loop where the magnetic field is either constant or zero.
4. Calculate the current inside the loop.
5. Calculate the line integral $\oint \vec{B} \cdot d\vec{l}$ around the closed loop.
6. Equate $\oint \vec{B} \cdot d\vec{l}$ with $\mu_0 I_{enc}$ with $\mu_0 I_{enc}$ and solve for \vec{B} .

Using Ampère's Law to Calculate the Magnetic Field Due to a Wire

Use Ampère's law to calculate the magnetic field due to a steady current I in an infinitely long, thin, straight wire as shown in Figure 20.4.2.

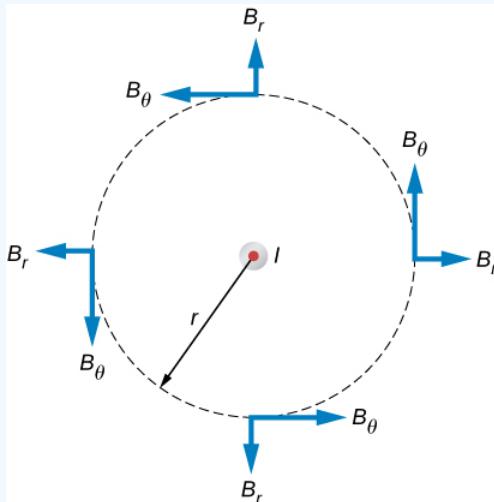


Figure 20.4.2: The possible components of the magnetic field \vec{B} due to a current I , which is directed out of the page. The radial component is zero because the angle between the magnetic field and the path is at a right angle.

Strategy

Consider an arbitrary plane perpendicular to the wire, with the current directed out of the page. The possible magnetic field components in this plane, B_r and B_θ are shown at arbitrary points on a circle of radius r centered on the wire. Since the field is cylindrically symmetric, neither B_r nor B_θ varies with the position on this circle. Also from symmetry, the radial lines, if they exist, must be directed either all inward or all outward from the wire. This means, however, that there must be a net magnetic flux across an arbitrary cylinder concentric with the wire. The radial component of the magnetic field must be zero because $\vec{B} \cdot d\vec{l} = 0$. Therefore, we can apply Ampère's law to the circular path as shown.

Solution

Over this path \vec{B} is constant and parallel to $d\vec{l}$, so

$$\oint \vec{B} \cdot d\vec{l} = B_\theta \oint dl = B_\theta(2\pi r).$$

Thus Ampère's law reduces to

$$B_\theta(2\pi r) = \mu_0 I.$$

Finally, since B_θ is the only component of \vec{B} , we can drop the subscript and write

$$B = \frac{\mu_0 I}{2\pi r}.$$

This agrees with the Biot-Savart calculation above.

Significance

Ampère's law works well if you have a path to integrate over which $\vec{B} \cdot d\vec{l}$ has results that are easy to simplify. For the infinite wire, this works easily with a path that is circular around the wire so that the magnetic field factors out of the integration. If the path dependence looks complicated, you can always go back to the Biot-Savart law and use that to find the magnetic field.

✓ Example 20.4.2: Calculating the Magnetic Field of a Thick Wire with Ampère's Law

The radius of the long, straight wire of Figure 20.4.3 is a , and the wire carries a current I_0 that is distributed uniformly over its cross-section. Find the magnetic field both inside and outside the wire.

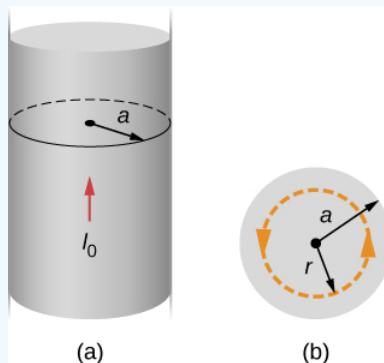


Figure 20.4.3: (a) A model of a current-carrying wire of radius a and current I_0 . (b) A cross-section of the same wire showing the radius a and the Ampère's loop of radius r .

Strategy

This problem has the same geometry as Example 20.4.1, but the enclosed current changes as we move the integration path from outside the wire to inside the wire, where it doesn't capture the entire current enclosed (see Figure 20.4.3).

Solution

For any circular path of radius r that is centered on the wire,

$$\oint \vec{B} \cdot d\vec{l} = \oint B dl = B \oint dl = B(2\pi r).$$

From Ampère's law, this equals the total current passing through any surface bounded by the path of integration.

Consider first a circular path that is inside the wire ($r \leq a$) such as that shown in part (a) of Figure 20.4.3. We need the current I passing through the area enclosed by the path. It's equal to the current density J times the area enclosed. Since the current is uniform, the current density inside the path equals the current density in the whole wire, which is $I_0/\pi a^2$. Therefore the current I passing through the area enclosed by the path is

$$I = \frac{\pi r^2}{\pi a^2} I_0 = \frac{r^2}{a^2} I_0.$$

We can consider this ratio because the current density J is constant over the area of the wire. Therefore, the current density of a part of the wire is equal to the current density in the whole area. Using Ampère's law, we obtain

$$B(2\pi r) = \mu_0 \left(\frac{r^2}{a^2} \right) I_0,$$

and the magnetic field inside the wire is

$$B = \frac{\mu_0 I_0}{2\pi} \frac{r}{a^2} (r \leq a).$$

Outside the wire, the situation is identical to that of the infinite thin wire of the previous example; that is,

$$B = \frac{\mu_0 I_0}{2\pi r} (r \geq a).$$

The variation of \mathbf{B} with r is shown in Figure 20.4.4.

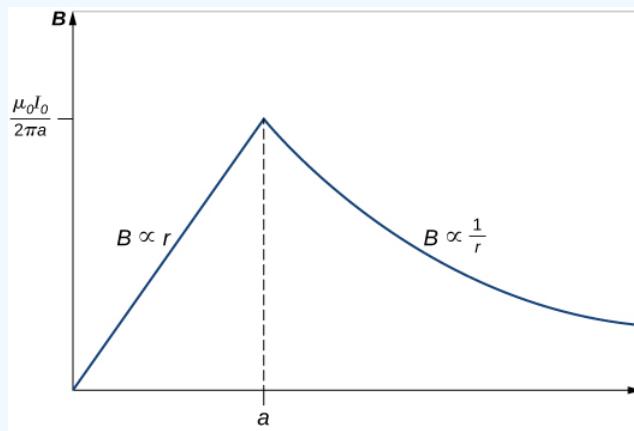


Figure 20.4.4: Variation of the magnetic field produced by a current I_0 in a long, straight wire of radius a .

Significance

The results show that as the radial distance increases inside the thick wire, the magnetic field increases from zero to a familiar value of the magnetic field of a thin wire. Outside the wire, the field drops off regardless of whether it was a thick or thin wire.

This result is similar to how Gauss's law for electrical charges behaves inside a uniform charge distribution, except that Gauss's law for electrical charges has a uniform volume distribution of charge, whereas Ampère's law here has a uniform area of current distribution. Also, the drop-off outside the thick wire is similar to how an electric field drops off outside of a linear charge distribution, since the two cases have the same geometry and neither case depends on the configuration of charges or currents once the loop is outside the distribution.

✓ Using Ampère's Law with Arbitrary Paths

Use Ampère's law to evaluate $\oint \vec{B} \cdot d\vec{l}$ for the current configurations and paths in Figure 20.4.5.

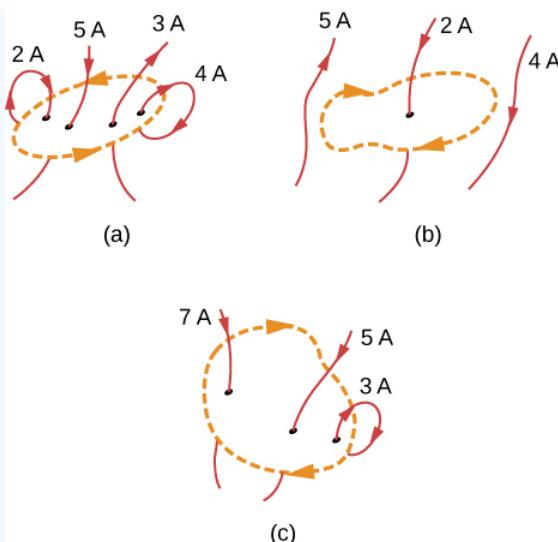


Figure 20.4.5: Current configurations and paths for Example 20.4.3.

Strategy

Ampère's law states that $\oint \vec{B} \cdot d\vec{l} = \mu_0 I$ where I is the total current passing through the enclosed loop. The quickest way to evaluate the integral is to calculate $\mu_0 I$ by finding the net current through the loop. Positive currents flow with your right-hand thumb if your fingers wrap around in the direction of the loop. This will tell us the sign of the answer.

Solution

(a) The current going downward through the loop equals the current going out of the loop, so the net current is zero. Thus, $\oint \vec{B} \cdot d\vec{l} = 0$.

(b) The only current to consider in this problem is 2A because it is the only current inside the loop. The right-hand rule shows us the current going downward through the loop is in the positive direction. Therefore, the answer is $\oint \vec{B} \cdot d\vec{l} = \mu_0(2 \text{ A}) = 2.51 \times 10^{-6} \text{T} \cdot \text{m}$.

(c) The right-hand rule shows us the current going downward through the loop is in the positive direction. There are $7\text{A} + 5\text{A} = 12\text{A}$ of current going downward and -3 A going upward. Therefore, the total current is 9 A and $\oint \vec{B} \cdot d\vec{l} = \mu_0(9 \text{ A}) = 5.65 \times 10^{-6} \text{T} \cdot \text{m}$.

Significance

If the currents all wrapped around so that the same current went into the loop and out of the loop, the net current would be zero and no magnetic field would be present. This is why wires are very close to each other in an electrical cord. The currents flowing toward a device and away from a device in a wire equal zero total current flow through an Ampère loop around these wires. Therefore, no stray magnetic fields can be present from cords carrying current.

Exercise 20.4.1

Consider using Ampère's law to calculate the magnetic fields of a finite straight wire and of a circular loop of wire. Why is it not useful for these calculations?

Answer

In these cases the integrals around the Ampèrean loop are very difficult because there is no symmetry, so this method would not be useful.

This page titled 20.4: Ampère's Law is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- 12.6: Ampère's Law by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

20.5: Maxwell's Equations and Electromagnetic Waves

Learning Objectives

By the end of this section, you will be able to:

- Explain Maxwell's correction of Ampère's law by including the displacement current
- State and apply Maxwell's equations in integral form
- Describe how the symmetry between changing electric and changing magnetic fields explains Maxwell's prediction of electromagnetic waves
- Describe how Hertz confirmed Maxwell's prediction of electromagnetic waves

James Clerk **Maxwell** (1831–1879) was one of the major contributors to physics in the nineteenth century (Figure 20.5.1). Although he died young, he made major contributions to the development of the kinetic theory of gases, to the understanding of color vision, and to the nature of Saturn's rings. He is probably best known for having combined existing knowledge of the laws of electricity and of magnetism with insights of his own into a complete overarching electromagnetic theory, represented by **Maxwell's equations**.



Figure 20.5.1: James Clerk Maxwell, a nineteenth-century physicist, developed a theory that explained the relationship between electricity and magnetism, and correctly predicted that visible light consists of electromagnetic waves.

Maxwell's Correction to the Laws of Electricity and Magnetism

The four basic laws of electricity and magnetism had been discovered experimentally through the work of physicists such as Oersted, Coulomb, Gauss, and Faraday. Maxwell discovered logical inconsistencies in these earlier results and identified the incompleteness of Ampère's law as their cause.

Recall that according to Ampère's law, the integral of the magnetic field around a closed loop **C** is proportional to the current **I** passing through any surface whose boundary is loop **C** itself:

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 I. \quad (20.5.1)$$

There are infinitely many surfaces that can be attached to any loop, and Ampère's law stated in Equation 20.5.1 is independent of the choice of surface.

Consider the set-up in Figure 20.5.2. A source of emf is abruptly connected across a parallel-plate capacitor so that a time-dependent current **I** develops in the wire. Suppose we apply Ampère's law to loop **C** shown at a time before the capacitor is fully charged, so that $I \neq 0$. Surface **S₁** gives a nonzero value for the enclosed current **I**, whereas surface **S₂** gives zero for the enclosed current because no current passes through it:

$$\oint_C \vec{B} \cdot d\vec{s} = \mu_0 I$$

if surface S_1 is used

$$= 0$$

if surface S_2 is used

Clearly, Ampère's law in its usual form does not work here. This may not be surprising, because Ampère's law as applied in earlier chapters required a steady current, whereas the current in this experiment is changing with time and is not steady at all.

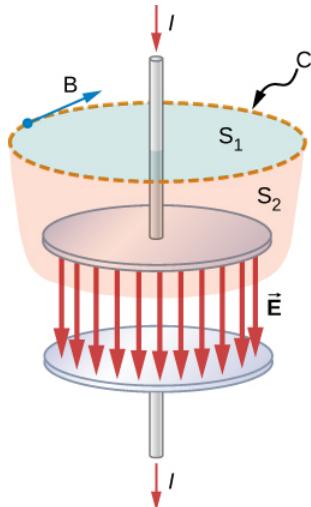


Figure 20.5.2: The currents through surface S_1 and surface S_2 are unequal, despite having the same boundary loop C .

How can Ampère's law be modified so that it works in all situations? Maxwell suggested including an additional contribution, called the displacement current I_d , to the real current \mathbf{I} ,

$$\oint_S \vec{B} \cdot d\vec{s} = \mu_0(I + I_d) \quad (20.5.2)$$

where the displacement current is defined to be

$$I_d = \epsilon_0 \frac{d\Phi_E}{dt}. \quad (20.5.3)$$

Here ϵ_0 is the **permittivity of free space** and Φ_E is the electric flux, defined as

$$\Phi_E = \iint_{\text{Surface } S} \vec{E} \cdot d\vec{A}.$$

The **displacement current** is analogous to a real current in Ampère's law, entering into Ampère's law in the same way. It is produced, however, by a changing electric field. It accounts for a changing electric field producing a magnetic field, just as a real current does, but the displacement current can produce a magnetic field even where no real current is present. When this extra term is included, the modified Ampère's law equation becomes

$$\oint_C \vec{B} \cdot d\vec{s} = \mu_0 I + \epsilon_0 \mu_0 \frac{d\Phi_E}{dt}$$

and is independent of the surface S through which the current \mathbf{I} is measured.

We can now examine this modified version of Ampère's law to confirm that it holds independent of whether the surface S_1 or the surface S_2 in Figure 20.5.2 is chosen. The electric field \vec{E} corresponding to the flux Φ_E in Equation 20.5.3 is between the capacitor plates. Therefore, the \vec{E} field and the displacement current through the surface S_1 are both zero, and Equation 20.5.2 takes the form

$$\oint_C \vec{B} \cdot d\vec{s} = \mu_0 I. \quad (20.5.4)$$

We must now show that for surface S_2 , through which no actual current flows, the displacement current leads to the same value $\mu_0 I$ for the right side of the Ampère's law equation. For surface S_2 the equation becomes

$$\oint_C \vec{B} \cdot d\vec{s} = \mu_0 \frac{d}{dt} \left[\epsilon_0 \iint_{\text{Surface } S_2} \vec{E} \cdot d\vec{A} \right].$$

Gauss's law for electric charge requires a closed surface and cannot ordinarily be applied to a surface like S_1 alone or S_2 alone. But the two surfaces S_1 and S_2 form a closed surface in Figure 20.5.2 and can be used in Gauss's law. Because the electric field is zero on S_1 , the flux contribution through S_1 is zero. This gives us

$$\oint_{\text{Surface } S_1+S_2} \vec{E} \cdot d\vec{A} = \iint_{\text{Surface } S_1} \vec{E} \cdot d\vec{A} + \iint_{\text{Surface } S_2} \vec{E} \cdot d\vec{A} \quad (20.5.5)$$

$$= 0 + \iint_{\text{Surface } S_2} \vec{E} \cdot d\vec{A} \quad (20.5.6)$$

$$= \iint_{\text{Surface } S_2} \vec{E} \cdot d\vec{A}. \quad (20.5.7)$$

Therefore, we can replace the integral over S_2 in Equation 20.5.4 with the closed Gaussian surface $S_1 + S_2$ and apply Gauss's law to obtain

$$\oint_{S_1} \vec{B} \cdot d\vec{s} = \mu_0 \frac{dQ_{in}}{dt} = \mu_0 I.$$

Thus, the modified Ampère's law equation is the same using surface S_2 , where the right-hand side results from the displacement current, as it is for the surface S_1 , where the contribution comes from the actual flow of electric charge.

✓ Displacement current in a charging capacitor

A parallel-plate capacitor with capacitance C whose plates have area A and separation distance d is connected to a resistor R and a battery of voltage V . The current starts to flow at $t = 0$.

- Find the displacement current between the capacitor plates at time t .
- From the properties of the capacitor, find the corresponding real current $I = \frac{dQ}{dt}$, and compare the answer to the expected current in the wires of the corresponding **RC** circuit.

Strategy

We can use the equations from the analysis of an **RC** circuit ([Alternating-Current Circuits](#)) plus Maxwell's version of Ampère's law.

Solution

- The voltage between the plates at time t is given by

$$V_C = \frac{1}{C} Q(t) = V_0 (1 - e^{-t/RC}).$$

Let the **z**-axis point from the positive plate to the negative plate. Then the **z**-component of the electric field between the plates as a function of time t is

$$E_z(t) = \frac{V_0}{d} (1 - e^{-t/RC}).$$

Therefore, the z-component of the displacement current I_d between the plates is

$$I_d(t) = \epsilon_0 A \frac{\partial E_z(t)}{\partial t} = \epsilon_0 A \frac{V_0}{d} \times \frac{1}{RC} e^{-t/RC} = \frac{V_0}{R} e^{-t/RC},$$

where we have used $C = \epsilon_0 \frac{A}{d}$ for the capacitance.

- From the expression for V_C the charge on the capacitor is

$$Q(t) = CV_C = CV_0 (1 - e^{-t/RC}).$$

The current into the capacitor after the circuit is closed, is therefore

$$I = \frac{dQ}{dt} = \frac{V_0}{R} e^{-t/RC}.$$

This current is the same as I_d found in (a).

Maxwell's Equations

With the correction for the displacement current, Maxwell's equations take the form

$$\oint \vec{E} \cdot d\vec{A} = \frac{Q_{in}}{\epsilon_0} \text{ (Gauss's law)} \quad (20.5.8)$$

$$\oint \vec{B} \cdot d\vec{A} = 0 \text{ (Gauss's law for magnetism)} \quad (20.5.9)$$

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_m}{dt} \text{ (Faraday's law)} \quad (20.5.10)$$

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 I + \epsilon_0 I + \epsilon_0 \mu_0 \frac{d\Phi_E}{dt} \text{ (Ampère-Maxwell law).} \quad (20.5.11)$$

Once the fields have been calculated using these four equations, the [Lorentz force equation](#)

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

gives the force that the fields exert on a particle with charge q moving with velocity \vec{v} . The Lorentz force equation combines the force of the electric field and of the magnetic field on the moving charge. The magnetic and electric forces have been examined in earlier modules. These four Maxwell's equations are, respectively:

Maxwell's Equations

1. Gauss's law

The electric flux through any closed surface is equal to the electric charge Q_{in} enclosed by the surface. Gauss's law (Equation 20.5.8) describes the relation between an electric charge and the electric field it produces. This is often pictured in terms of electric field lines originating from positive charges and terminating on negative charges, and indicating the direction of the electric field at each point in space.

2. Gauss's law for magnetism

The magnetic field flux through any closed surface is zero (Equation 20.5.9). This is equivalent to the statement that magnetic field lines are continuous, having no beginning or end. Any magnetic field line entering the region enclosed by the surface must also leave it. No magnetic monopoles, where magnetic field lines would terminate, are known to exist (see section on [Magnetic Fields and Lines](#)).

3. Faraday's law

A changing magnetic field induces an electromotive force (emf) and, hence, an electric field. The direction of the emf opposes the change. Equation 20.5.10 is Faraday's law of induction and includes Lenz's law. The electric field from a changing magnetic field has field lines that form closed loops, without any beginning or end.

4. Ampère-Maxwell law

Magnetic fields are generated by moving charges or by changing electric fields. This fourth of Maxwell's equations, Equation 20.5.11, encompasses Ampère's law and adds another source of magnetic fields, namely changing electric fields.

Maxwell's equations and the Lorentz force law together encompass all the laws of electricity and magnetism. The symmetry that Maxwell introduced into his mathematical framework may not be immediately apparent. Faraday's law describes how changing magnetic fields produce electric fields. The displacement current introduced by Maxwell results instead from a changing electric field and accounts for a changing electric field producing a magnetic field. The equations for the effects of both changing electric fields and changing magnetic fields differ in form only where the absence of magnetic monopoles leads to missing terms. This

symmetry between the effects of changing magnetic and electric fields is essential in explaining the nature of electromagnetic waves.

Later application of Einstein's theory of relativity to Maxwell's complete and symmetric theory showed that electric and magnetic forces are not separate but are different manifestations of the same thing—the electromagnetic force. The electromagnetic force and weak nuclear force are similarly unified as the electroweak force. This unification of forces has been one motivation for attempts to unify all of the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces (see [Particle Physics and Cosmology](#)).

The Mechanism of Electromagnetic Wave Propagation

To see how the symmetry introduced by Maxwell accounts for the existence of combined electric and magnetic waves that propagate through space, imagine a time-varying magnetic field $\vec{B}_0(t)$ produced by the high-frequency alternating current seen in Figure 20.5.3. We represent $\vec{B}_0(t)$ in the diagram by one of its field lines. From Faraday's law, the changing magnetic field through a surface induces a time-varying electric field $\vec{E}_0(t)$ at the boundary of that surface. The displacement current source for the electric field, like the Faraday's law source for the magnetic field, produces only closed loops of field lines, because of the mathematical symmetry involved in the equations for the induced electric and induced magnetic fields. A field line representation of $\vec{E}_0(t)$ is shown. In turn, the changing electric field $\vec{E}_0(t)$ creates a magnetic field $\vec{B}_1(t)$ according to the modified Ampère's law. This changing field induces $\vec{E}_1(t)$ which induces $\vec{B}_2(t)$ and so on. We then have a self-continuing process that leads to the creation of time-varying electric and magnetic fields in regions farther and farther away from \mathbf{O} . This process may be visualized as the propagation of an electromagnetic wave through space.

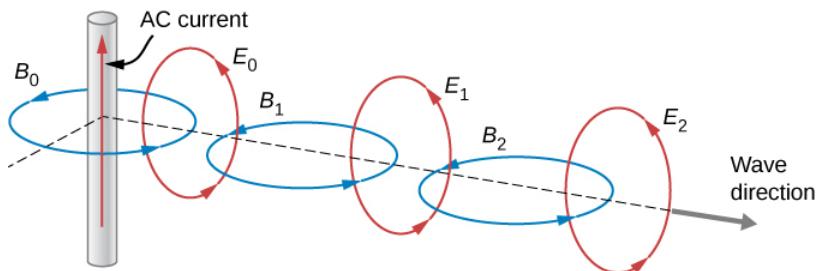


Figure 20.5.3: How changing \vec{E} and \vec{B} fields propagate through space.

In the next section, we show in more precise mathematical terms how Maxwell's equations lead to the prediction of electromagnetic waves that can travel through space without a material medium, implying a speed of electromagnetic waves equal to the speed of light.

Prior to Maxwell's work, experiments had already indicated that light was a wave phenomenon, although the nature of the waves was yet unknown. In 1801, Thomas Young (1773–1829) showed that when a light beam was separated by two narrow slits and then recombined, a pattern made up of bright and dark fringes was formed on a screen. Young explained this behavior by assuming that light was composed of waves that added constructively at some points and destructively at others (see [Interference](#)). Subsequently, Jean Foucault (1819–1868), with measurements of the speed of light in various media, and Augustin Fresnel (1788–1827), with detailed experiments involving interference and diffraction of light, provided further conclusive evidence that light was a wave. So, light was known to be a wave, and Maxwell had predicted the existence of electromagnetic waves that traveled at the speed of light. The conclusion seemed inescapable: Light must be a form of electromagnetic radiation. But Maxwell's theory showed that other wavelengths and frequencies than those of light were possible for electromagnetic waves. He showed that electromagnetic radiation with the same fundamental properties as visible light should exist at any frequency. It remained for others to test, and confirm, this prediction.

Exercise 20.5.1

When the emf across a capacitor is turned on and the capacitor is allowed to charge, when does the magnetic field induced by the displacement current have the greatest magnitude?

Solution

It is greatest immediately after the current is switched on. The displacement current and the magnetic field from it are proportional to the rate of change of electric field between the plates, which is greatest when the plates first begin to charge.

Hertz's Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves but also verified that they travel at the speed of light.

Hertz used an alternating-current **RLC** (resistor-inductor-capacitor) circuit that resonates at a known frequency $f_0 = \frac{1}{2\pi\sqrt{LC}}$ and connected it to a loop of wire, as shown in Figure 20.5.4. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and helped generate electromagnetic waves.

Across the laboratory, Hertz placed another loop attached to another **RLC** circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could thus be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.

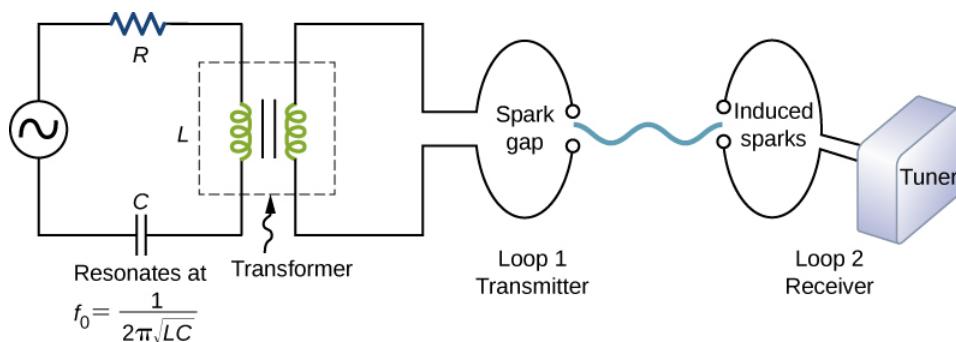


Figure 20.5.4: The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, confirming their wave character. He was able to determine the wavelengths from the interference patterns, and knowing their frequencies, he could calculate the propagation speed using the equation $v = f\lambda$, where v is the speed of a wave, f is its frequency, and λ is its wavelength. Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz (**1 Hz = 1 cycle/second**), is named in his honor.

Exercise 20.5.2

Could a purely electric field propagate as a wave through a vacuum without a magnetic field? Justify your answer.

Solution

No. The changing electric field according to the modified version of Ampère's law would necessarily induce a changing magnetic field.

This page titled [20.5: Maxwell's Equations and Electromagnetic Waves](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- **16.2: Maxwell's Equations and Electromagnetic Waves** by [OpenStax](#) is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

20.6: Plane Electromagnetic Waves

LEARNING OBJECTIVES

By the end of this section, you will be able to:

- Describe how Maxwell's equations predict the relative directions of the electric fields and magnetic fields, and the direction of propagation of plane electromagnetic waves
- Explain how Maxwell's equations predict that the speed of propagation of electromagnetic waves in free space is exactly the speed of light
- Calculate the relative magnitude of the electric and magnetic fields in an electromagnetic plane wave
- Describe how electromagnetic waves are produced and detected

Mechanical waves travel through a medium such as a string, water, or air. Perhaps the most significant prediction of Maxwell's equations is the existence of combined electric and magnetic (or electromagnetic) fields that propagate through space as electromagnetic waves. Because Maxwell's equations hold in free space, the predicted electromagnetic waves, unlike mechanical waves, do not require a medium for their propagation.

A general treatment of the physics of electromagnetic waves is beyond the scope of this textbook. We can, however, investigate the special case of an electromagnetic wave that propagates through free space along the **x**-axis of a given coordinate system.

Electromagnetic Waves in One Direction

An electromagnetic wave consists of an electric field, defined as usual in terms of the force per charge on a stationary charge, and a magnetic field, defined in terms of the force per charge on a moving charge. The electromagnetic field is assumed to be a function of only the **x**-coordinate and time. The **y**-component of the electric field is then written as $E_y(x, t)$, the **z**-component of the magnetic field as $B_z(x, t)$, etc. Because we are assuming free space, there are no free charges or currents, so we can set $Q_{in} = 0$ and $I = 0$ in Maxwell's equations.

The transverse nature of electromagnetic waves

We examine first what Gauss's law for electric fields implies about the relative directions of the electric field and the propagation direction in an electromagnetic wave. Assume the Gaussian surface to be the surface of a rectangular box whose cross-section is a square of side **l** and whose third side has length Δx , as shown in Figure 20.6.1. Because the electric field is a function only of **x** and **t**, the **y**-component of the electric field is the same on both the top (labeled Side 2) and bottom (labeled Side 1) of the box, so that these two contributions to the flux cancel. The corresponding argument also holds for the net flux from the **z**-component of the electric field through Sides 3 and 4. Any net flux through the surface therefore comes entirely from the **x**-component of the electric field. Because the electric field has no **y**- or **z**-dependence, $E_x(x, t)$ is constant over the face of the box with area **A** and has a possibly different value $E_x(x + \Delta x, t)$ that is constant over the opposite face of the box.

Applying [Gauss's law](#) gives

$$\text{Net flux} = -E_x(x, t)A + E_x(x + \Delta x, t)A = \frac{Q_{in}}{\epsilon_0} \quad (20.6.1)$$

where $A = l \times l$ is the area of the front and back faces of the rectangular surface. But the charge enclosed is $Q_{in} = 0$, so this component's net flux is also zero, and Equation 20.6.1 implies $E_x(x, t) = E_x(x + \Delta x, t)$ for any Δx . Therefore, if there is an **x**-component of the electric field, it cannot vary with **x**. A uniform field of that kind would merely be superposed artificially on the traveling wave, for example, by having a pair of parallel-charged plates. Such a component $E_x(x, t)$ would not be part of an electromagnetic wave propagating along the **x**-axis; so $E_x(x, t) = 0$ for this wave. Therefore, the only nonzero components of the electric field are $E_y(x, t)$ and $E_z(x, t)$ perpendicular to the direction of propagation of the wave.

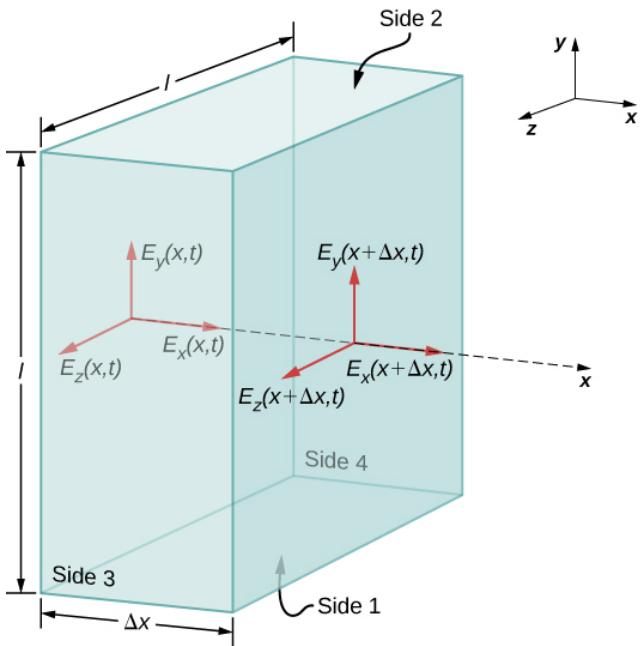


Figure 20.6.1: The surface of a rectangular box of dimensions $l \times l \times \Delta x$ is our Gaussian surface. The electric field shown is from an electromagnetic wave propagating along the x -axis.

A similar argument holds by substituting \mathbf{E} for \mathbf{B} and using Gauss's law for magnetism instead of Gauss's law for electric fields. This shows that the \mathbf{B} field is also perpendicular to the direction of propagation of the wave. The electromagnetic wave is therefore a transverse wave, with its oscillating electric and magnetic fields perpendicular to its direction of propagation.

The speed of propagation of electromagnetic waves

We can next apply Maxwell's equations to the description given in connection with Figure 16.2.3 in the previous section to obtain an equation for the \mathbf{E} field from the changing \mathbf{B} field, and for the \mathbf{B} field from a changing \mathbf{E} field. We then combine the two equations to show how the changing \mathbf{E} and \mathbf{B} fields propagate through space at a speed precisely equal to the speed of light.

First, we apply Faraday's law over Side 3 of the Gaussian surface, using the path shown in Figure 20.6.2. Because $E_x(x, t) = 0$, we have

$$\oint \vec{E} \cdot d\vec{s} = -E_y(x, t)l + E_y(x + \Delta x, t)l.$$

Assuming Δx is small and approximating $E_y(x + \Delta x, t)$ by

$$E_y(x + \Delta x, t) = E_y(x, t) + \frac{\partial E_y(x, t)}{\partial x} \Delta x,$$

we obtain

$$\oint \vec{E} \cdot d\vec{s} = \frac{\partial E_y(x, t)}{\partial x} (l \Delta x).$$

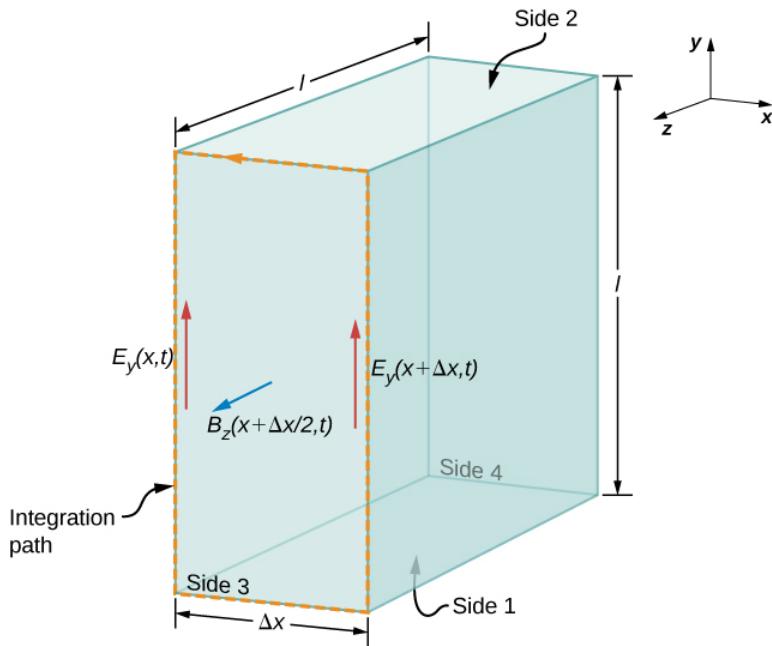


Figure 20.6.2: We apply Faraday's law to the front of the rectangle by evaluating $\oint \vec{E} \cdot d\vec{s}$ along the rectangular edge of Side 3 in the direction indicated, taking the B field crossing the face to be approximately its value in the middle of the area traversed.

Because Δx is small, the magnetic flux through the face can be approximated by its value in the center of the area traversed, namely $B_z\left(x + \frac{\Delta x}{2}, t\right)$. The flux of the **B** field through Face 3 is then the **B** field times the area,

$$\oint_S \vec{B} \cdot \vec{n} dA = B_z\left(x + \frac{\Delta x}{2}, t\right) (l\Delta x). \quad (20.6.2)$$

From [Faraday's law](#),

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d}{dt} \int_S \vec{B} \cdot \vec{n} dA. \quad (20.6.3)$$

Therefore, from Equations 20.6.1 and 20.6.2,

$$\frac{\partial E_y(x, t)}{\partial x} (l\Delta x) = -\frac{\partial}{\partial t} \left[B_z\left(x + \frac{\Delta x}{2}, t\right) \right] (l\Delta x).$$

Canceling $l\Delta x$ and taking the limit as $\Delta x = 0$, we are left with

$$\frac{\partial E_y(x, t)}{\partial x} = -\frac{\partial B_z(x, t)}{\partial t}. \quad (20.6.4)$$

We could have applied Faraday's law instead to the top surface (numbered 2) in Figure 20.6.2, to obtain the resulting equation

$$\frac{\partial B_z(x, t)}{\partial t} = -\frac{\partial E_y(x, t)}{\partial x}. \quad (20.6.5)$$

This is the equation describing the spatially dependent **E** field produced by the time-dependent **B** field.

Next we apply the Ampère-Maxwell law (with $I = 0$) over the same two faces (Surface 3 and then Surface 2) of the rectangular box of Figure 20.6.2. Applying Equation 16.2.16,

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 \epsilon_0 (d/dt) \int_S \vec{E} \cdot \vec{n} da$$

to Surface 3, and then to Surface 2, yields the two equations

$$\frac{\partial E_y(x, t)}{\partial x} = -\epsilon_0 \mu_0 \frac{\partial E_z(x, t)}{\partial t}, \quad (20.6.6)$$

and

$$\frac{\partial B_z(x,t)}{\partial x} = -\epsilon_0 \mu_0 \frac{\partial E_y(x,t)}{\partial t}. \quad (20.6.7)$$

These equations describe the spatially dependent \mathbf{B} field produced by the time-dependent \mathbf{E} field.

We next combine the equations showing the changing \mathbf{B} field producing an \mathbf{E} field with the equation showing the changing \mathbf{E} field producing a \mathbf{B} field. Taking the derivative of Equation 20.6.4 with respect to x and using Equation 20.6.13 gives

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{\partial}{\partial x} \left(\frac{\partial E_y}{\partial x} \right) = -\frac{\partial}{\partial x} \left(\frac{\partial B_z}{\partial t} \right) = -\frac{\partial}{\partial t} \left(\frac{\partial B_z}{\partial x} \right) = \frac{\partial}{\partial t} \left(\epsilon_0 \mu_0 \frac{\partial E_y}{\partial t} \right)$$

or

$$\frac{\partial^2 E_y}{\partial x^2} = \epsilon_0 \mu_0 \frac{\partial^2 E_y}{\partial t^2}$$

This is the form taken by the general wave equation for our plane wave. Because the equations describe a wave traveling at some as-yet-unspecified speed c , we can assume the field components are each functions of $\mathbf{x} - c\mathbf{t}$ for the wave traveling in the $+x$ -direction, that is,

$$E_y(x,t) = f(\xi) \text{ where } \xi = x - ct. \quad (20.6.8)$$

It is left as a mathematical exercise to show, using the chain rule for differentiation, that Equations 20.6.5 and 20.6.6 imply

$$1 = \epsilon_0 \mu_0 c^2.$$

The speed of the electromagnetic wave in free space is therefore given in terms of the permeability and the permittivity of free space by

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}. \quad (20.6.9)$$

We could just as easily have assumed an electromagnetic wave with field components $E_z(x,t)$ and $B_y(x,t)$. The same type of analysis with Equation 20.6.12 and 20.6.11 would also show that the speed of an electromagnetic wave is $c = 1/\sqrt{\epsilon_0 \mu_0}$.

The physics of traveling electromagnetic fields was worked out by Maxwell in 1873. He showed in a more general way than our derivation that electromagnetic waves always travel in free space with a speed given by Equation 20.6.6. If we evaluate the speed $c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$, we find that

$$c = \frac{1}{\sqrt{\left(8.85 \times 10^{-12} \frac{C^2}{N \cdot m^2}\right) \left(4\pi \times 10^{-7} \frac{T \cdot m}{A}\right)}} = 3.00 \times 10^8 m/s,$$

which is the speed of light. Imagine the excitement that Maxwell must have felt when he discovered this equation! He had found a fundamental connection between two seemingly unrelated phenomena: electromagnetic fields and light.

Exercise 20.6.1

The wave equation was obtained by (1) finding the \mathbf{E} field produced by the changing \mathbf{B} field, (2) finding the \mathbf{B} field produced by the changing \mathbf{E} field, and combining the two results. Which of Maxwell's equations was the basis of step (1) and which of step (2)?

Answer (step 1)

Faraday's law

Answer (step 2)

the Ampère-Maxwell law

How the **E** and **B** Fields Are Related

So far, we have seen that the rates of change of different components of the **E** and **B** fields are related, that the electromagnetic wave is transverse, and that the wave propagates at speed **c**. We next show what Maxwell's equations imply about the ratio of the **E** and **B** field magnitudes and the relative directions of the **E** and **B** fields.

We now consider solutions to Equation 20.6.4 in the form of plane waves for the electric field:

$$E_y(x, t) = E_0 \cos(kx - \omega t). \quad (20.6.10)$$

We have arbitrarily taken the wave to be traveling in the $+x$ -direction and chosen its phase so that the maximum field strength occurs at the origin at time $t = 0$. We are justified in considering only sines and cosines in this way, and generalizing the results, because Fourier's theorem implies we can express any wave, including even square step functions, as a superposition of sines and cosines.

At any one specific point in space, the **E** field oscillates sinusoidally at angular frequency ω between $+E_0$ and $-E_0$ and similarly, the **B** field oscillates between $+B_0$ and $-B_0$. The amplitude of the wave is the maximum value of $E_y(x, t)$. The period of oscillation **T** is the time required for a complete oscillation. The frequency **f** is the number of complete oscillations per unit of time, and is related to the angular frequency ω by $\omega = 2\pi f$. The wavelength λ is the distance covered by one complete cycle of the wave, and the wavenumber **k** is the number of wavelengths that fit into a distance of 2π in the units being used. These quantities are related in the same way as for a mechanical wave:

$$\omega = 2\pi f, \quad f = \frac{1}{T}, \quad k = \frac{2\pi}{\lambda}, \quad \text{and} \quad c = f\lambda = \omega/k.$$

Given that the solution of E_y has the form shown in Equation ???, we need to determine the **B** field that accompanies it. From Equation 20.6.11, the magnetic field component B_z must obey

$$\begin{aligned} \frac{\partial B_z}{\partial t} &= -\frac{\partial E_y}{\partial x} \\ \frac{\partial B_z}{\partial t} &= -\frac{\partial}{\partial x} E_0 \cos(kx - \omega t) = kE_0 \sin(kx - \omega t). \end{aligned} \quad (20.6.11)$$

Because the solution for the **B**-field pattern of the wave propagates in the $+x$ -direction at the same speed **c** as the **E**-field pattern, it must be a function of $k(x - ct) = kx - \omega t$. Thus, we conclude from Equation 20.6.8 that B_z is

$$B_z(x, t) = \frac{k}{\omega} E_0 \cos(kx - \omega t) = \frac{1}{c} E_0 \cos(kx - \omega t).$$

These results may be written as

$$\begin{aligned} E_y(x, t) &= E_0 \cos(kx - \omega t) \\ B_z(x, t) &= B_0 \cos(kx - \omega t) \end{aligned} \quad (20.6.12)$$

$$\frac{E_y}{B_z} = \frac{E_0}{B_0} = c. \quad (20.6.13)$$

Therefore, the peaks of the **E** and **B** fields coincide, as do the troughs of the wave, and at each point, the **E** and **B** fields are in the same ratio equal to the speed of light **c**. The plane wave has the form shown in Figure 20.6.3.

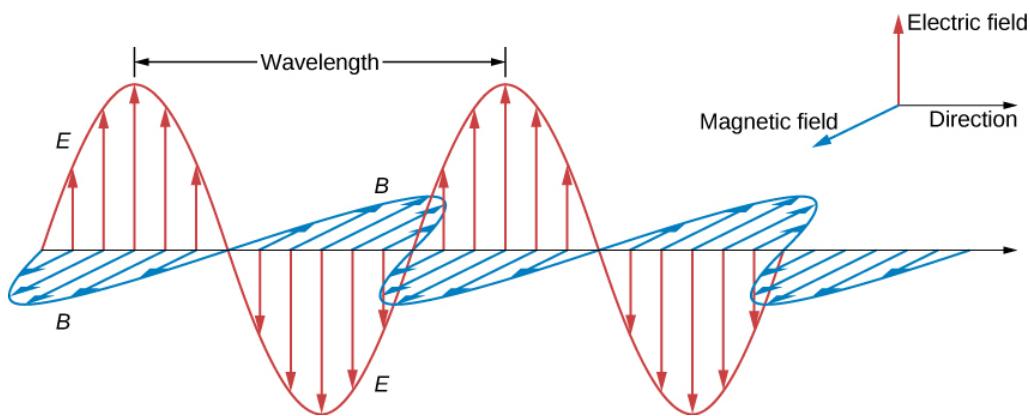


Figure 20.6.3: The plane wave solution of Maxwell's equations has the **B** field directly proportional to the **E** field at each point, with the relative directions shown.

✓ Example 20.6.1: Calculating B-Field Strength in an Electromagnetic Wave

What is the maximum strength of the **B** field in an electromagnetic wave that has a maximum **E**-field strength of 1000 V/m?

Strategy

To find the **B**-field strength, we rearrange Equation 20.6.10 to solve for **B**, yielding

$$B = \frac{E}{c}.$$

Solution We are given **E**, and **c** is the speed of light. Entering these into the expression for **B** yields

$$B = \frac{1000 \text{ V/m}}{3.00 \times 10^8 \text{ m/s}} = 3.33 \times 10^{-6} \text{ T}.$$

Significance

The **B**-field strength is less than a tenth of Earth's admittedly weak magnetic field. This means that a relatively strong electric field of 1000 V/m is accompanied by a relatively weak magnetic field.

Changing electric fields create relatively weak magnetic fields. The combined electric and magnetic fields can be detected in electromagnetic waves, however, by taking advantage of the phenomenon of resonance, as Hertz did. A system with the same natural frequency as the electromagnetic wave can be made to oscillate. All radio and TV receivers use this principle to pick up and then amplify weak electromagnetic waves, while rejecting all others not at their resonant frequency.

? Exercise 20.6.2

What conclusions did our analysis of Maxwell's equations lead to about these properties of a plane electromagnetic wave:

- the relative directions of wave propagation, of the **E** field, and of **B** field,
- the speed of travel of the wave and how the speed depends on frequency, and
- the relative magnitudes of the **E** and **B** fields.

Answer a

The directions of wave propagation, of the **E** field, and of **B** field are all mutually perpendicular.

Answer b

The speed of the electromagnetic wave is the speed of light $c = 1/\sqrt{\epsilon_0\mu_0}$ independent of frequency.

Answer c

The ratio of electric and magnetic field amplitudes is $E/B = c$.

Production and Detection of Electromagnetic Waves

A steady electric current produces a magnetic field that is constant in time and which does not propagate as a wave. Accelerating charges, however, produce electromagnetic waves. An electric charge oscillating up and down, or an alternating current or flow of charge in a conductor, emit radiation at the frequencies of their oscillations. The electromagnetic field of a **dipole antenna** is shown in Figure 20.6.4. The positive and negative charges on the two conductors are made to reverse at the desired frequency by the output of a transmitter as the power source. The continually changing current accelerates charge in the antenna, and this results in an oscillating electric field a distance away from the antenna. The changing electric fields produce changing magnetic fields that in turn produce changing electric fields, which thereby propagate as electromagnetic waves. The frequency of this radiation is the same as the frequency of the ac source that is accelerating the electrons in the antenna. The two conducting elements of the dipole antenna are commonly straight wires. The total length of the two wires is typically about one-half of the desired wavelength (hence, the alternative name **half-wave antenna**), because this allows standing waves to be set up and enhances the effectiveness of the radiation.

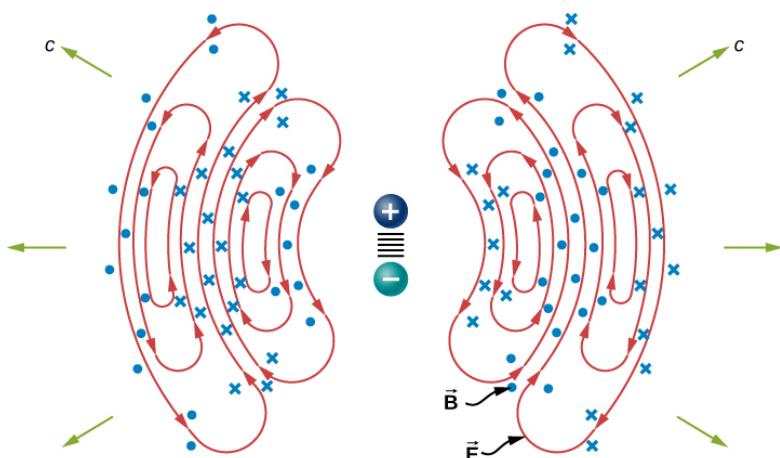


Figure 20.6.4: The oscillatory motion of the charges in a dipole antenna produces electromagnetic radiation.

The electric field lines in one plane are shown. The magnetic field is perpendicular to this plane. This radiation field has cylindrical symmetry around the axis of the dipole. Field lines near the dipole are not shown. The pattern is not at all uniform in all directions. The strongest signal is in directions perpendicular to the axis of the antenna, which would be horizontal if the antenna is mounted vertically. There is zero intensity along the axis of the antenna. The fields detected far from the antenna are from the changing electric and magnetic fields inducing each other and traveling as electromagnetic waves. Far from the antenna, the wave fronts, or surfaces of equal phase for the electromagnetic wave, are almost spherical. Even farther from the antenna, the radiation propagates like electromagnetic plane waves.

The electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving electromagnetic signals works in reverse. Incoming electromagnetic waves induce oscillating currents in the antenna, each at its own frequency. The radio receiver includes a tuner circuit, whose resonant frequency can be adjusted. The tuner responds strongly to the desired frequency but not others, allowing the user to tune to the desired broadcast. Electrical components amplify the signal formed by the moving electrons. The signal is then converted into an audio and/or video format.

Note

Use this [simulation](#) to broadcast radio waves. Wiggle the transmitter electron manually or have it oscillate automatically. Display the field as a curve or vectors. The strip chart shows the electron positions at the transmitter and at the receiver.

This page titled [20.6: Plane Electromagnetic Waves](#) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [16.3: Plane Electromagnetic Waves](#) by OpenStax is licensed CC BY 4.0. Original source: <https://openstax.org/details/books/university-physics-volume-2>.

20.7: Momentum and Radiation Pressure

Learning Objectives

By the end of this section, you will be able to:

- Describe the relationship of the radiation pressure and the energy density of an electromagnetic wave
- Explain how the radiation pressure of light, while small, can produce observable astronomical effects

Material objects consist of charged particles. An electromagnetic wave incident on the object exerts forces on the charged particles, in accordance with the [Lorentz force](#). These forces do work on the particles of the object, increasing its energy, as discussed in the previous section. The energy that sunlight carries is a familiar part of every warm sunny day. A much less familiar feature of electromagnetic radiation is the extremely weak pressure that electromagnetic radiation produces by exerting a force in the direction of the wave. This force occurs because electromagnetic waves contain and transport momentum.

To understand the direction of the force for a very specific case, consider a plane electromagnetic wave incident on a metal in which electron motion, as part of a current, is damped by the resistance of the metal, so that the average electron motion is in phase with the force causing it. This is comparable to an object moving against friction and stopping as soon as the force pushing it stops (Figure 20.7.1). When the electric field is in the direction of the positive y -axis, electrons move in the negative y -direction, with the magnetic field in the direction of the positive z -axis. By applying the right-hand rule, and accounting for the negative charge of the electron, we can see that the force on the electron from the magnetic field is in the direction of the positive x -axis, which is the direction of wave propagation. When the \vec{E} field reverses, the \vec{B} field does too, and the force is again in the same direction. Maxwell's equations together with the Lorentz force equation imply the existence of radiation pressure much more generally than this specific example, however.

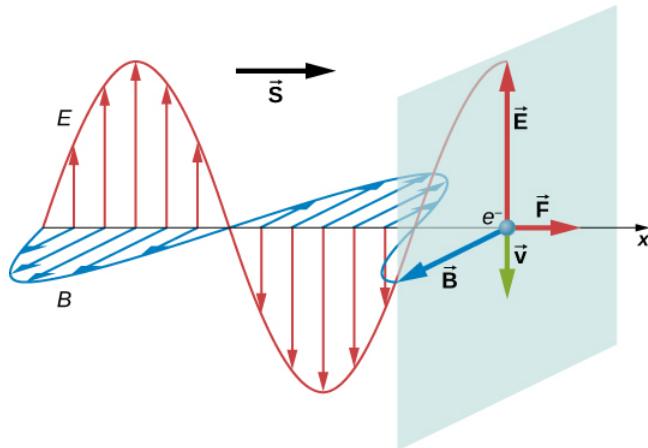


Figure 20.7.1: Electric and magnetic fields of an electromagnetic wave can combine to produce a force in the direction of propagation, as illustrated for the special case of electrons whose motion is highly damped by the resistance of a metal.

Maxwell predicted that an electromagnetic wave carries momentum. An object absorbing an electromagnetic wave would experience a force in the direction of propagation of the wave. The force corresponds to radiation pressure exerted on the object by the wave. The force would be twice as great if the radiation were reflected rather than absorbed.

Maxwell's prediction was confirmed in 1903 by Nichols and Hull by precisely measuring radiation pressures with a torsion balance. The schematic arrangement is shown in Figure 20.7.2. The mirrors suspended from a fiber were housed inside a glass container. Nichols and Hull were able to obtain a small measurable deflection of the mirrors from shining light on one of them. From the measured deflection, they could calculate the unbalanced force on the mirror, and obtained agreement with the predicted value of the force.

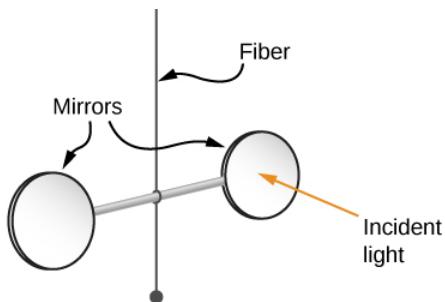


Figure 20.7.2: Simplified diagram of the central part of the apparatus Nichols and Hull used to precisely measure radiation pressure and confirm Maxwell's prediction.

The **radiation pressure** p_{rad} applied by an electromagnetic wave on a perfectly absorbing surface turns out to be equal to the energy density of the wave:

$$\underbrace{p_{rad} = u}_{\text{Perfect absorber}} . \quad (20.7.1)$$

If the material is perfectly reflecting, such as a metal surface, and if the incidence is along the normal to the surface, then the pressure exerted is twice as much because the momentum direction reverses upon reflection:

$$\underbrace{p_{rad} = 2u}_{\text{Perfect reflector}} . \quad (20.7.2)$$

We can confirm that the units are right:

$$[u] = \frac{J}{m^3} = \frac{N \cdot m}{m^3} = \frac{N}{m^2} = \text{units of pressure.}$$

Equations 20.7.1 and 20.7.2 give the **instantaneous** pressure, but because the energy density oscillates rapidly, we are usually interested in the time-averaged radiation pressure, which can be written in terms of intensity:

$$p = \langle p_{rad} \rangle = \begin{cases} I/c & \text{Perfect absorber} \\ 2I/c & \text{Perfect reflector} \end{cases} \quad (20.7.3)$$

Radiation pressure plays a role in explaining many observed astronomical phenomena, including the appearance of **comets**. Comets are basically chunks of icy material in which frozen gases and particles of rock and dust are embedded. When a comet approaches the Sun, it warms up and its surface begins to evaporate. The **coma** of the comet is the hazy area around it from the gases and dust. Some of the gases and dust form tails when they leave the comet. Notice in Figure 20.7.3 that a comet has **two tails**. The **ion tail** (or **gas tail**) is composed mainly of ionized gases. These ions interact electromagnetically with the solar wind, which is a continuous stream of charged particles emitted by the Sun. The force of the solar wind on the ionized gases is strong enough that the ion tail almost always points directly away from the Sun. The second tail is composed of dust particles. Because the **dust tail** is electrically neutral, it does not interact with the solar wind. However, this tail is affected by the radiation pressure produced by the light from the Sun. Although quite small, this pressure is strong enough to cause the dust tail to be displaced from the path of the comet.

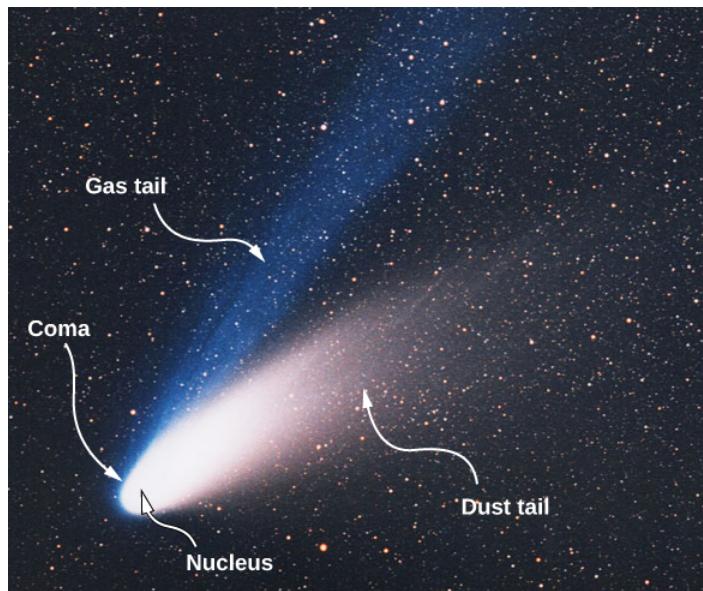


Figure 20.7.3: Evaporation of material being warmed by the Sun forms two tails, as shown in this photo of Comet Ison. (credit: modification of work by E. Slawik—ESO)

✓ Example 20.7.1: Halley's Comet

On February 9, 1986, Comet Halley was at its closest point to the Sun, about $9.0 \times 10^{10} m$ from the center of the Sun. The average power output of the Sun is $3.8 \times 10^{26} W$.

- Calculate the radiation pressure on the comet at this point in its orbit. Assume that the comet reflects all the incident light.
- Suppose that a 10-kg chunk of material of cross-sectional area $4.0 \times 10^{-2} m^2$ breaks loose from the comet. Calculate the force on this chunk due to the solar radiation. Compare this force with the gravitational force of the Sun.

Strategy

Calculate the intensity of solar radiation at the given distance from the Sun and use that to calculate the radiation pressure. From the pressure and area, calculate the force.

Solution

a. The intensity of the solar radiation is the average solar power per unit area. Hence, at $9.0 \times 10^{10} m$ from the center of the Sun, we have

$$\begin{aligned} I &= S_{avg} \\ &= \frac{3.8 \times 10^{26} W}{4\pi(9.0 \times 10^{10} m)^2} \\ &= 3.7 \times 10^3 W/m^2. \end{aligned}$$

Assuming the comet reflects all the incident radiation, we obtain from Equation 20.7.3

$$\begin{aligned} p &= \frac{2I}{c} \\ &= \frac{2(3.7 \times 10^3 W/m^2)}{3.00 \times 10^8 m/s} \\ &= 2.5 \times 10^{-5} N/m^2. \end{aligned}$$

b. The force on the chunk due to the radiation is

$$\begin{aligned} F &= pA \\ &= (2.5 \times 10^{-5} N/m^2)(4.0 \times 10^{-2} m^2) \\ &= 1.0 \times 10^{-6} N, \end{aligned}$$

whereas the gravitational force of the Sun is

$$\begin{aligned}
 F_g &= \frac{GMm}{r^2} \\
 &= \frac{(6.67 \times 10^{-11} N \cdot m^2/kg^2)(2.0 \times 10^{30} kg)(10 kg)}{(9.0 \times 10^{10} m)^2} \\
 &= 0.16 N.
 \end{aligned}$$

Significance

The gravitational force of the Sun on the chunk is therefore much greater than the force of the radiation.

After Maxwell showed that light carried momentum as well as energy, a novel idea eventually emerged, initially only as science fiction. Perhaps a spacecraft with a large reflecting **light sail** could use radiation pressure for propulsion. Such a vehicle would not have to carry fuel. It would experience a constant but small force from solar radiation, instead of the short bursts from rocket propulsion. It would accelerate slowly, but by being accelerated continuously, it would eventually reach great speeds. A spacecraft with small total mass and a sail with a large area would be necessary to obtain a usable acceleration.

When the space program began in the 1960s, the idea started to receive serious attention from NASA. The most recent development in light propelled spacecraft has come from a citizen-funded group, the Planetary Society. It is currently testing the use of light sails to propel a small vehicle built from **CubeSats**, tiny satellites that NASA places in orbit for various research projects during space launches intended mainly for other purposes.

The **LightSail** spacecraft shown below (Figure 20.7.4) consists of three **CubeSats** bundled together. It has a total mass of only about 5 kg and is about the size as a loaf of bread. Its sails are made of very thin Mylar and open after launch to have a surface area of $32 m^2$.

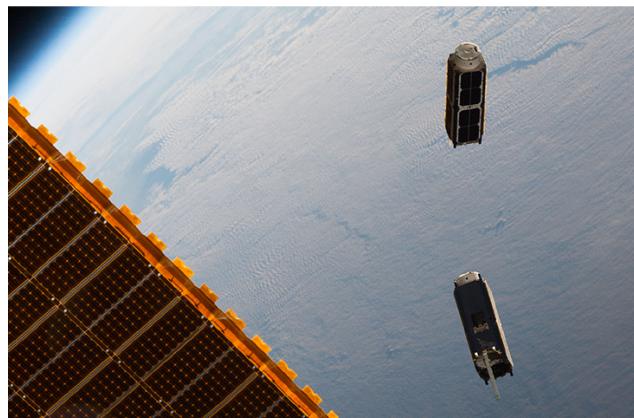
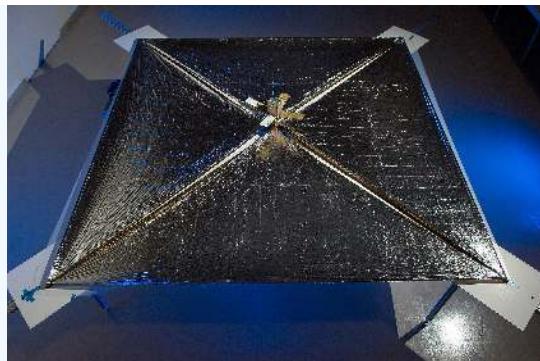


Figure 20.7.3: Two small CubeSat satellites deployed from the International Space Station in May, 2016. The solar sails open out when the CubeSats are far enough away from the Station.

✓ Example 20.7.2: LightSail Acceleration

The first **LightSail** spacecraft was launched in 2015 to test the sail deployment system. It was placed in low-earth orbit in 2015 by hitching a ride on an Atlas 5 rocket launched for an unrelated mission. The test was successful, but the low-earth orbit allowed too much drag on the spacecraft to accelerate it by sunlight. Eventually, it burned in the atmosphere, as expected. The next Planetary Society's **LightSail** solar sailing spacecraft is scheduled for 2018.



The **LightSail** is based on the on NASA's NanoSail-D project. (Public domain; NASA).

LightSail Acceleration

The intensity of energy from sunlight at a distance of 1 AU from the Sun is 1370 W/m^2 . The **LightSail** spacecraft has sails with total area of 32 m^2 and a total mass of 5.0 kg. Calculate the maximum acceleration LightSail spacecraft could achieve from radiation pressure when it is about 1 AU from the Sun.

Strategy

The maximum acceleration can be expected when the sail is opened directly facing the Sun. Use the light intensity to calculate the radiation pressure and from it, the force on the sails. Then use Newton's second law to calculate the acceleration.

Solution

The radiation pressure is

$$F = pA = 2uA = \frac{2I}{c}A = \frac{2(1370 \text{ W/m}^2)(32 \text{ m}^2)}{(3.00 \times 10^8 \text{ m/s})} = 2.92 \times 10^{-4} \text{ N}.$$

The resulting acceleration is

$$a = \frac{F}{m} = \frac{2.92 \times 10^{-4} \text{ N}}{5.0 \text{ kg}} = 5.8 \times 10^{-5} \text{ m/s}^2.$$

Significance

If this small acceleration continued for a year, the craft would attain a speed of 1829 m/s, or 6600 km/h.

Exercise 20.7.1

How would the speed and acceleration of a radiation-propelled spacecraft be affected as it moved farther from the Sun on an interplanetary space flight?

Solution

Its acceleration would decrease because the radiation force is proportional to the intensity of light from the Sun, which decreases with distance. Its speed, however, would not change except for the effects of gravity from the Sun and planets.

This page titled [20.7: Momentum and Radiation Pressure](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by OpenStax via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [16.5: Momentum and Radiation Pressure](#) by OpenStax is licensed [CC BY 4.0](#). Original source: <https://openstax.org/details/books/university-physics-volume-2>.

CHAPTER OVERVIEW

21: Electrical Transmission Lines

- 21.1: Introduction
- 21.2: Phasors
- 21.3: Introduction to Transmission Lines
- 21.4: Types of Transmission Lines
- 21.5: Transmission Lines as Two-Port Devices
- 21.6: Lumped-Element Model
- 21.7: Telegrapher's Equations
- 21.8: Wave Equation for a Transmission Line
- 21.9: Characteristic Impedance of a Transmission Line
- 21.10: Wave Propagation on a Transmission Line
- 21.11: Lossless and Low-Loss Transmission Lines
- 21.12: Voltage Reflection Coefficient
- 21.13: Standing Waves
- 21.14: Standing Wave Ratio
- 21.15: Parallel Wire Transmission Line
- 21.16: Attenuation in Coaxial Cable
- 21.17: Power Handling Capability of Coaxial Cable
- 21.18: Why 50 Ohms?
- 21.19: Conclusion

21: Electrical Transmission Lines is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

21.1: Introduction

Transmission lines are cables that can carry electromagnetic signals from one location to another. This chapter develops a theory using a lumped element approach that describes how the signals are propagated through the transmission line.

In the context of amateur radio, transmission lines are an important part of the overall radio system. Antennas are often mounted in locations remote from the radio to maximize their ability to transmit and receive. As such, it is the transmission line which must carry the signal to or from the antenna, preferably with no reflection and minimal losses. This chapter defines and discusses factors associated with propagation of electromagnetic signals in a transmission line like standing wave ratio. In particular, the chapter discusses the coaxial cable, which is one of the most commonly used transmission lines.

21.1: Introduction is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Ronald Kumon.

21.2: Phasors

In many areas of engineering, signals are well-modeled as sinusoids. Also, devices that process these signals are often well-modeled as *linear time-invariant (LTI)* systems. The response of an LTI system to any linear combination of sinusoids is another linear combination of sinusoids having the same frequencies. In other words, (1) sinusoidal signals processed by LTI systems remain sinusoids and are not somehow transformed into square waves or some other waveform; and (2) we may calculate the response of the system for one sinusoid at a time, and then add the results to find the response of the system when multiple sinusoids are applied simultaneously. This property of LTI systems is known as *superposition*.

The analysis of systems that process sinusoidal waveforms is greatly simplified when the sinusoids are represented as phasors. Here is the key idea:

Definition: phasor

A *phasor* is a complex-valued number that represents a real-valued sinusoidal waveform. Specifically, a phasor has the magnitude and phase of the sinusoid it represents

Figure 21.2.1 and 21.2.2 show some examples of phasors and the associated sinusoids. It is important to note that a phasor by itself is not the signal. A phasor is merely a simplified mathematical representation in which the actual, real-valued physical signal is represented as a complex-valued constant.

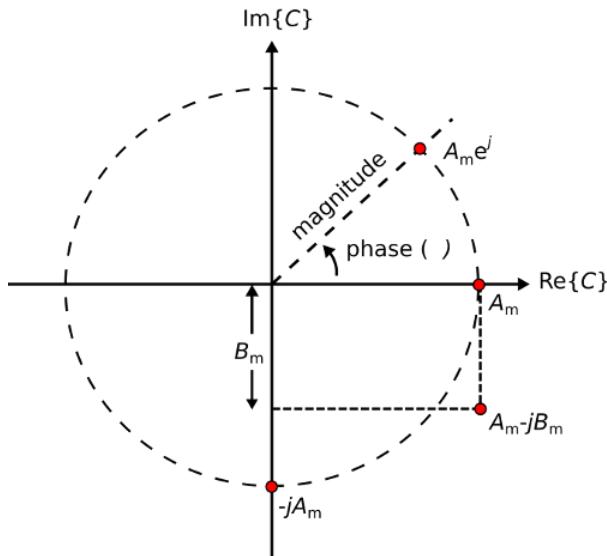


Figure 21.2.1: Examples of phasors, displayed here as points in the real-imaginary plane.

Equation 21.2.1 is a completely general form for a physical (hence, real-valued) quantity varying sinusoidally with angular frequency $\omega = 2\pi f$

$$A(t; \omega) = A_m(\omega) \cos(\omega t + \psi(\omega)) \quad (21.2.1)$$

where $A_m(\omega)$ is magnitude at the specified frequency, $\psi(\omega)$ is phase at the specified frequency, and t is time. Also, we require $\partial A_m / \partial t = 0$; that is, that the time variation of $A(t)$ is completely represented by the cosine function alone. Now we can equivalently express $A(t; \omega)$ as a phasor $C(\omega)$:

$$C(\omega) = A_m(\omega) e^{j\psi(\omega)} \quad (21.2.2)$$

To convert this phasor back to the physical signal it represents, we (1) restore the time dependence by multiplying by $e^{j\omega t}$, and then (2) take the real part of the result. In mathematical notation:

$$A(t; \omega) = \operatorname{Re} \{ C(\omega) e^{j\omega t} \} \quad (21.2.3)$$

To see why this works, simply substitute the right hand side of Equation 21.2.2 into Equation 21.2.3. Then

$$\begin{aligned}
 A(t) &= \operatorname{Re} \{ A_m(\omega) e^{j\psi(\omega)} e^{j\omega t} \} \\
 &= \operatorname{Re} \{ A_m(\omega) e^{j(\omega t + \psi(\omega))} \} \\
 &= \operatorname{Re} \{ A_m(\omega) [\cos(\omega t + \psi(\omega)) + j \sin(\omega t + \psi(\omega))] \} \\
 &= A_m(\omega) \cos(\omega t + \psi(\omega))
 \end{aligned}$$

as expected.

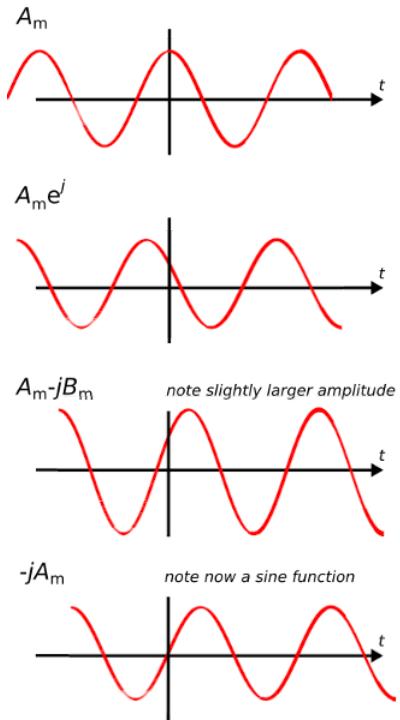


Figure 21.2.2: Sinusoids corresponding to the phasors shown in Figure 21.2.1

It is common to write Equation 21.2.3 as follows, dropping the explicit indication of frequency dependence:

$$C = A_m e^{j\psi}$$

This does not normally cause any confusion since the definition of a phasor requires that values of C and ψ are those that apply at whatever frequency is represented by the suppressed sinusoidal dependence $e^{j\omega t}$.

Table 21.2.1 shows mathematical representations of the same phasors demonstrated in Figure 21.2.1 (and their associated sinusoidal waveforms in Figure 21.2.2). It is a good exercise is to confirm each row in the table, transforming from left to right and vice-versa.

Table 21.2.1: Some examples of physical (real-valued) sinusoidal signals and the corresponding phasors. A_m and B_m are real-valued and constant with respect to t

$A(t)$	C
$A_m \cos(\omega t)$	A_m
$A_m \cos(\omega t + \psi)$	$A_m e^{j\psi}$
$A_m \sin(\omega t) = A_m (\cos(\omega t - \frac{\pi}{2}))$	$-jA_m$
$A_m \cos(\omega t) + B_m \sin(\omega t) = A_m \cos(\omega t) + B_m \cos(\omega t - \frac{\pi}{2})$	$A_m - jB_m$

It is not necessary to use a phasor to represent a sinusoidal signal. We choose to do so because phasor representation leads to dramatic simplifications. For example:

- Calculation of the peak value from data representing $A(t; \omega)$ requires a time-domain search over one period of the sinusoid. However, if you know C , the peak value of $A(t)$ is simply $|C|$, and no search is required.
- Calculation of ψ from data representing $A(t; \omega)$ requires correlation (essentially, integration) over one period of the sinusoid. However, if you know C , then ψ is simply the phase of C , and no integration is required.

Furthermore, mathematical operations applied to $A(t; \omega)$ can be equivalently performed as operations on C , and the latter are typically much easier than the former. To demonstrate this, we first make two important claims and show that they are true.

✓ Example 21.2.1: Claim 1

Let C_1 and C_2 be two complex-valued constants (independent of t). Also, $\operatorname{Re}\{C_1 e^{j\omega t}\} = \operatorname{Re}\{C_2 e^{j\omega t}\}$ for all t . Then, $C_1 = C_2$.

Proof

Evaluating at $t = 0$ we find $\operatorname{Re}\{C_1\} = \operatorname{Re}\{C_2\}$. Since C_1 and C_2 are constant with respect to time, this must be true for all t . At $t = \pi/(2\omega)$ we find

$$\operatorname{Re}\{C_1 e^{j\omega t}\} = \operatorname{Re}\{C_1 \cdot j\} = -\operatorname{Im}\{C_1\}$$

and similarly

$$\operatorname{Re}\{C_2 e^{j\omega t}\} = \operatorname{Re}\{C_2 \cdot j\} = -\operatorname{Im}\{C_2\}$$

therefore $\operatorname{Im}\{C_1\} = \operatorname{Im}\{C_2\}$. Once again: Since C_1 and C_2 are constant with respect to time, this must be true for all t . Since the real and imaginary parts of C_1 and C_2 are equal, $C_1 = C_2$.

What does this mean?

We have just shown that if two phasors are equal, then the sinusoidal waveforms that they represent are also equal.

✓ Example 21.2.2: Claim 2

For any real-valued linear operator \mathcal{T} and complex-valued quantity C ,

$$\mathcal{T}(\operatorname{Re}\{C\}) = \operatorname{Re}\{\mathcal{T}(C)\}. \quad (21.2.4)$$

Proof

Let $C = c_r + jc_i$ where c_r and c_i are real-valued quantities, and evaluate the right side of Equation 21.2.4:

$$\begin{aligned} \operatorname{Re}\{\mathcal{T}(C)\} &= \operatorname{Re}\{\mathcal{T}(c_r + jc_i)\} \\ &= \operatorname{Re}\{\mathcal{T}(c_r) + j\mathcal{T}(c_i)\} \\ &= \mathcal{T}(c_r) \\ &= \mathcal{T}(\operatorname{Re}\{C\}) \end{aligned}$$

What does this mean?

The operators that we have in mind for \mathcal{T} include addition, multiplication by a constant, differentiation, integration, and so on. Here's an example with differentiation:

$$\begin{aligned} \operatorname{Re}\left\{\frac{\partial}{\partial\omega}C\right\} &= \operatorname{Re}\left\{\frac{\partial}{\partial\omega}(c_r + jc_i)\right\} = \frac{\partial}{\partial\omega}c_r \\ \frac{\partial}{\partial\omega}\operatorname{Re}\{C\} &= \frac{\partial}{\partial\omega}\operatorname{Re}\{(c_r + jc_i)\} = \frac{\partial}{\partial\omega}c_r \end{aligned}$$

In other words, differentiation of a sinusoidal signal can be accomplished by differentiating the associated phasor, so there is no need to transform a phasor back into its associated real-valued signal in order to perform this operation.

Summary

Claims 1 and 2 together entitle us to perform operations on phasors as surrogates for the physical, real-valued, sinusoidal waveforms they represent. Once we are done, we can transform the resulting phasor back into the physical waveform it represents using Equation 21.2.3, if desired

However, a final transformation back to the time domain is usually *not* desired, since the phasor tells us everything we can know about the corresponding sinusoid

A skeptical student might question the value of phasor analysis on the basis that signals of practical interest are sometimes not sinusoidally-varying, and therefore phasor analysis seems not to apply generally. It is certainly true that many signals of practical interest are not sinusoidal, and many are far from it. Nevertheless, phasor analysis is broadly applicable. There are basically two reasons why this is so:

- Many signals, although not strictly sinusoidal, are “narrowband” and therefore well-modeled as sinusoidal. For example, a cellular telecommunications signal might have a bandwidth on the order of 10 MHz and a center frequency of about 2 GHz. This means the difference in frequency between the band edges of this signal is just 0.5% of the center frequency. The frequency response associated with signal propagation or with hardware can often be assumed to be constant over this range of frequencies. With some caveats, doing phasor analysis at the center frequency and assuming the results apply equally well over the bandwidth of interest is often a pretty good approximation.
- It turns out that phasor analysis is easily extensible to any physical signal, regardless of bandwidth. This is so because any physical signal can be decomposed into a linear combination of sinusoids – this is known as [Fourier analysis](#). The way to find this linear combination of sinusoids is by computing the Fourier series, if the signal is periodic, or the Fourier Transform, otherwise. Phasor analysis applies to each frequency independently, and (invoking superposition) the results can be added together to obtain the result for the complete signal. The process of combining results after phasor analysis results is nothing more than integration over frequency; i.e.:

$$\int_{-\infty}^{+\infty} A(t; \omega) d\omega$$

Using Equation 21.2.3, this can be rewritten:

$$\int_{-\infty}^{+\infty} \operatorname{Re} \{C(\omega) e^{j\omega t}\} d\omega$$

We can go one step further using Claim 2:

$$\operatorname{Re} \left\{ \int_{-\infty}^{+\infty} C(\omega) e^{j\omega t} d\omega \right\}$$

The quantity in the curly braces is simply the Fourier transform of $C(\omega)$. Thus, we see that we can analyze a signal of arbitrarily-large bandwidth simply by keeping ω as an independent variable while we are doing phasor analysis, and if we ever need the physical signal, we just take the real part of the Fourier transform of the phasor. So not only is it possible to analyze any time-domain signal using phasor analysis, it is also often far easier than doing the same analysis on the time-domain signal direct

summary

Phasor analysis does not limit us to sinusoidal waveforms. Phasor analysis is not only applicable to sinusoids and signals that are sufficiently narrowband, but is also applicable to signals of arbitrary bandwidth via Fourier analysis.

This page titled [21.2: Phasors](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson](#) ([Virginia Tech Libraries' Open Education Initiative](#)) .

- [1.5: Phasors](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.3: Introduction to Transmission Lines

A transmission line is a structure intended to transport electromagnetic signals or power.

A rudimentary transmission line is simply a pair of wires with one wire serving as a datum (i.e., a reference; e.g., “ground”) and the other wire bearing an electrical potential that is defined relative to that datum. Transmission lines having random geometry, such as the test leads shown in Figure 21.3.1, are useful only at very low frequencies and when loss, reactance, and immunity to electromagnetic interference (EMI) are not a concern.



Figure 21.3.1: These leads used to connect test equipment to circuits in a laboratory are a very rudimentary form of transmission line, suitable only for very low frequencies. (Public Domain; Dmitry G)

However, many circuits and systems operate at frequencies where the length or cross-sectional dimensions of the transmission line may be a significant fraction of a wavelength. In this case, the transmission line is no longer “transparent” to the circuits at either end. Furthermore, loss, reactance, and EMI are significant problems in many applications. These concerns motivate the use of particular types of transmission lines, and make it necessary to understand how to properly connect the transmission line to the rest of the system.

In electromagnetics, the term “transmission line” refers to a structure which is intended to support a *guided wave*. A guided wave is an electromagnetic wave that is contained within or bound to the line, and which does not radiate away from the line. This condition is normally met if the length and cross-sectional dimensions of the transmission line are small relative to a wavelength – say $\lambda/100$ (i.e., 1% of the wavelength). For example, two randomly-arranged wires might serve well enough to carry a signal at $f = 10\text{ MHz}$ over a length $l = 3\text{ cm}$, since l is only 0.1% of the wavelength $\lambda = c/f = 30\text{ m}$. However, if l is increased to 3 m , or if f is increased to 1 GHz , then l is now 10% of the wavelength. In this case, one should consider using a transmission line that forms a proper guided wave.

Preventing unintended radiation is not the only concern. Once we have established a guided wave on a transmission line, it is important that power applied to the transmission line be delivered to the circuit or device at the other end and not reflected back into the source. For the random wire $f = 10\text{ MHz}$, $l = 3\text{ cm}$ example above, there is little need for concern, since we expect a phase shift of roughly $0.001 \cdot 360^\circ = 0.36^\circ$ over the length of the transmission line, which is about 0.72° for a round trip. So, to a good approximation, the entire transmission line is at the same electrical potential and thus transparent to the source and destination. However, if l is increased to 3 m , or if f is increased to 1 GHz , then the associated round-trip phase shift becomes 72° . In this case, a reflected signal traveling in the opposite direction will add to create a total electrical potential, which varies in both magnitude and phase with position along the line. Thus, the impedance looking toward the destination via the transmission line will be different than the impedance looking toward the destination directly (Section 3.15 gives the details). The modified impedance will depend on the cross-sectional geometry, materials, and length of the line.

Cross-sectional geometry and materials also determine the loss and EMI immunity of the transmission line.

Summarizing:

Transmission lines are designed to support guided waves with controlled impedance, low loss, and a degree of immunity from EMI.

This page titled [21.3: Introduction to Transmission Lines](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- **3.1: Introduction to Transmission Lines** by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.4: Types of Transmission Lines

Two common types of transmission line are **coaxial line** (Figure 21.4.1) and **microstrip line** (Figure 21.4.2). Both are examples of *transverse electromagnetic* (TEM) transmission lines. A TEM line employs a single electromagnetic wave “mode” having electric and magnetic field vectors in directions perpendicular to the axis of the line, as shown in Figures 21.4.3 and 21.4.4. TEM transmission lines appear primarily in radio frequency applications.

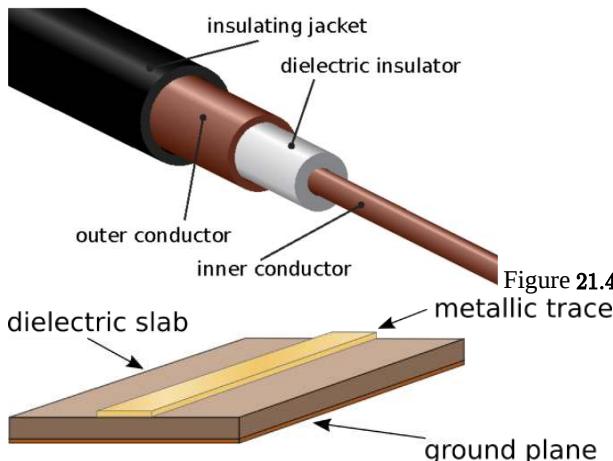


Figure 21.4.1: Structure of a coaxial transmission line. (CC BY 3.0 (modified)).

Figure 21.4.2: Structure of a microstrip transmission line. (CC BY SA 3.0 (modified))

TEM transmission lines such as coaxial lines and microstrip lines are designed to support a single electromagnetic wave that propagates along the length of the transmission line with electric and magnetic field vectors perpendicular to the direction of propagation.

Not all transmission lines exhibit TEM field structure. In non-TEM transmission lines, the electric and magnetic field vectors that are not necessarily perpendicular to the axis of the line, and the structure of the fields is complex relative to the field structure of TEM lines. An example of a transmission line that exhibits non-TEM field structure is the waveguide (see example in Figure 21.4.5). Waveguides are most prevalent at radio frequencies, and tend to appear in applications where it is important to achieve very low loss or where power levels are very high. Another example is common “multimode” optical fiber (Figure 21.4.6). Optical fiber exhibits complex field structure because the wavelength of light is very small compared to the cross-section of the fiber, making the excitation and propagation of non-TEM waves difficult to avoid. (This issue is overcome in a different type of optical fiber, known as “single mode” fiber, which is much more difficult and expensive to manufacture.)

Higher-order transmission lines, including radio-frequency waveguides and multimode optical fiber, are designed to guide waves that have relatively complex structure.

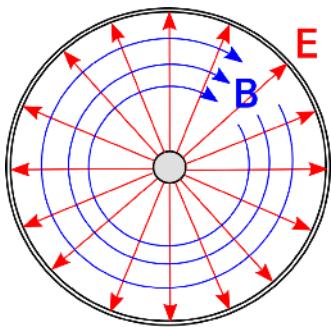
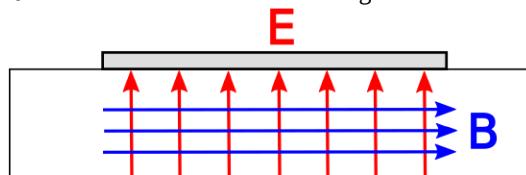


Figure 21.4.3: Structure of the electric and magnetic fields within coaxial line. In this case, the



wave is propagating away from the viewer.

Figure 21.4.4: Structure of the



shown.) In this case, the wave is propagating away from the viewer. ([CC BY SA 3.0 Unported](#)).

Figure 21.4.5: A network of radio frequency waveguides in an air traffic control radar. ([CC BY SA 2.0 Germany](#))



Figure 21.4.6: Strands of optical fiber.

This page titled [21.4: Types of Transmission Lines](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.2: Types of Transmission Lines](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source:
<https://doi.org/10.21061/electromagnetics-vol-1>.

21.5: Transmission Lines as Two-Port Devices

Figure 21.5.1 shows common ways to represent transmission lines in circuit diagrams. In each case, the source is represented using a Thévenin equivalent circuit consisting of a voltage source V_S in series with an impedance Z_S .¹ In transmission line analysis, the source may also be referred to as the *generator*. The termination on the receiving end of the transmission line is represented, without loss of generality, as an impedance Z_L . This termination is often referred to as the *load*, although in practice it can be any circuit that exhibits an input impedance of Z_L .

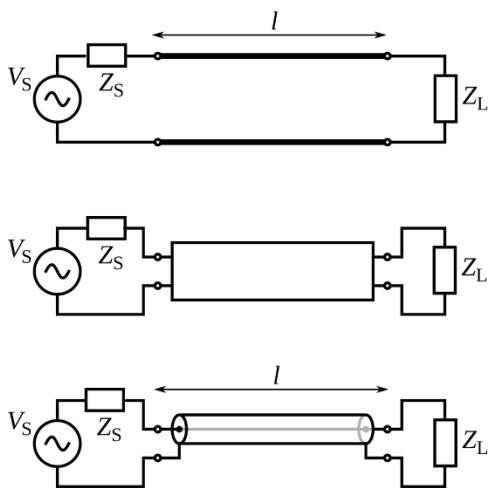


Figure 21.5.1: Symbols representing transmission lines: Top: As a generic two-conductor direct connection. Middle: As a generic two-port “black box.” Bottom: As a coaxial cable. © CC BY SA 3.0 Unported (modified)

The two-port representation of a transmission line is completely described by its length l along with some combination of the following parameters:

- Phase propagation constant β , having units of rad/m. This parameter also represents the wavelength in the line through the relationship $\lambda = 2\pi/\beta$. (See Sections 1.3 and 3.8 for details.)
- Attenuation constant α , having units of 1/m. This parameter quantifies the effect of loss in the line. (See Section 3.8 for details.)
- Characteristic impedance Z_0 , having units of Ω . This is the ratio of potential (“voltage”) to current when the line is perfectly impedance-matched at both ends. (See Section 3.7 for details.)

These parameters depend on the materials and geometry of the line.

Note that a transmission line is typically not transparent to the source and load. In particular, the load impedance may be Z_L , but the impedance presented to the source may or may not be equal to Z_L . (See Section 3.15 for more on this concept.) Similarly, the source impedance may be Z_S , but the impedance presented to the load may or may not be equal to Z_S . The effect of the transmission line on the source and load impedances will depend on the parameters identified above.

-
1. For a refresher on this concept, see “Additional Reading” at the end of this section. ↩

This page titled [21.5: Transmission Lines as Two-Port Devices](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.3: Transmission Lines as Two-Port Devices](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.6: Lumped-Element Model

It is possible to ascertain the relevant behaviors of a transmission line using elementary circuit theory applied to a differential-length lumped-element model of the transmission line. The concept is illustrated in Figure 21.6.1, which shows a generic transmission line aligned with its length along the z axis. The transmission line is divided into segments having small but finite length Δz . Each segment is modeled as an identical two-port having the equivalent circuit representation shown in Figure 21.6.2. The equivalent circuit consists of 4 components as follows:

- The resistance $R'\Delta z$ represents the series-combined ohmic resistance of the two conductors. This should account for *both* conductors since the current in the actual transmission line must flow through both conductors. The prime notation reminds us that R' is resistance *per unit length*; i.e., Ω/m , and it is only after multiplying by length that we get a resistance in Ω .
- The conductance $G'\Delta z$ represents the leakage of current directly from one conductor to the other. When $G'\Delta z > 0$, the resistance between the conductors is less than infinite, and therefore, current may flow between the conductors. This amounts to a loss of power separate from the loss associated with R' above. G' has units of S/m . Further note that G' is *not* equal to $1/R'$ as defined above. G' and R' are describing entirely different physical mechanisms (and in principle *either* could be defined as either a resistance or a conductance).
- The capacitance $C'\Delta z$ represents the capacitance of the transmission line structure. Capacitance is the tendency to store energy in electric fields and depends on the cross-sectional geometry and the media separating the conductors. C' has units of F/m .
- The inductance $L'\Delta z$ represents the inductance of the transmission line structure. Inductance is the tendency to store energy in magnetic fields, and (like capacitance) depends on the cross-sectional geometry and the media separating the conductors. L' has units of H/m .

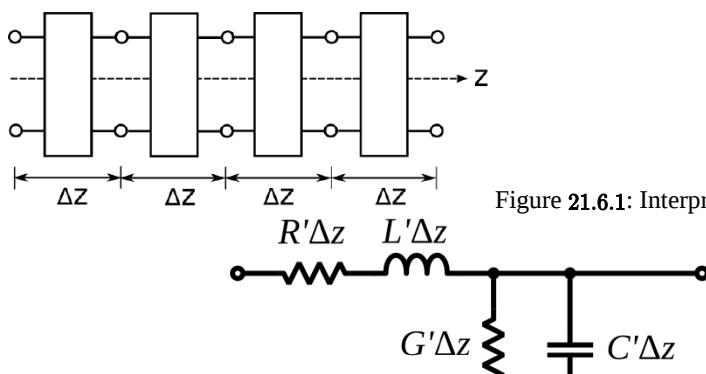


Figure 21.6.1: Interpretation of a transmission line as a cascade of discrete series-

connected two-ports.

Figure 21.6.1: Lumped-element equivalent circuit model for each of the two-ports in Figure 21.6.2. (CC BY SA 3.0 Unported (modified))

In order to use the model, one must have values for R' , G' , C' , and L' . Methods for computing these parameters are addressed elsewhere in this book.

This page titled 21.6: Lumped-Element Model is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Steven W. Ellingson (Virginia Tech Libraries' Open Education Initiative) .

- 3.4: Lumped-Element Model by Steven W. Ellingson is licensed CC BY-SA 4.0. Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.7: Telegrapher's Equations

In this section, we derive the equations that govern the potential $v(z, t)$ and current $i(z, t)$ along a transmission line that is oriented along the z axis. For this, we will employ the lumped-element model developed in Section 3.4.

To begin, we define voltages and currents as shown in Figure 21.7.1.

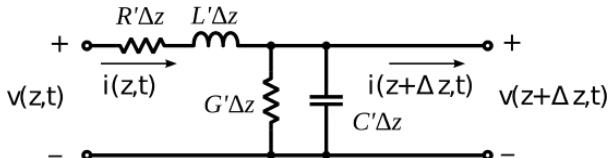


Figure 21.7.1: Lumped-element equivalent circuit transmission line

model, annotated with sign conventions for potentials and currents. (© CC BY SA 3.0 Unported (modified))

We assign the variables $v(z, t)$ and $i(z, t)$ to represent the potential and current on the left side of the segment, with reference polarity and direction as shown in the figure. Similarly we assign the variables $v(z + \Delta z, t)$ and $i(z + \Delta z, t)$ to represent the potential and current on the right side of the segment, again with reference polarity and direction as shown in the figure. Applying Kirchoff's voltage law from the left port, through $R'\Delta z$ and $L'\Delta z$, and returning via the right port, we obtain:

$$v(z, t) - (R'\Delta z) i(z, t) - (L'\Delta z) \frac{\partial}{\partial t} i(z, t) - v(z + \Delta z, t) = 0$$

Moving terms referring to current to the right side of the equation and then dividing through by Δz , we obtain

$$\begin{aligned} -\frac{v(z + \Delta z, t) - v(z, t)}{\Delta z} &= \\ R' i(z, t) + L' \frac{\partial}{\partial t} i(z, t) & \end{aligned}$$

Then taking the limit as $\Delta z \rightarrow 0$:

$$-\frac{\partial}{\partial z} v(z, t) = R' i(z, t) + L' \frac{\partial}{\partial t} i(z, t) \quad (21.7.1)$$

Applying Kirchoff's current law at the right port, we obtain:

$$\begin{aligned} i(z, t) - (G'\Delta z) v(z + \Delta z, t) - (C'\Delta z) \frac{\partial}{\partial t} v(z + \Delta z, t) \\ - i(z + \Delta z, t) = 0 \end{aligned}$$

Moving terms referring to potential to the right side of the equation and then dividing through by Δz , we obtain

$$\begin{aligned} -\frac{i(z + \Delta z, t) - i(z, t)}{\Delta z} &= \\ G' v(z + \Delta z, t) + C' \frac{\partial}{\partial t} v(z + \Delta z, t) & \end{aligned}$$

Taking the limit as $\Delta z \rightarrow 0$:

$$-\frac{\partial}{\partial z} i(z, t) = G' v(z, t) + C' \frac{\partial}{\partial t} v(z, t) \quad (21.7.2)$$

Equations 21.7.1 and 21.7.2 are the *telegrapher's equations*. These coupled (simultaneous) differential equations can be solved for $v(z, t)$ and $i(z, t)$ given R' , G' , L' , C' and suitable boundary conditions.

The time-domain telegrapher's equations are usually more than we need or want. If we are only interested in the response to a sinusoidal stimulus, then considerable simplification is possible using phasor representation.¹ First we define phasors $\tilde{V}(z)$ and $\tilde{I}(z)$ through the usual relationship:

$$v(z, t) = \text{Re} \left\{ \tilde{V}(z) e^{j\omega t} \right\}$$

$$i(z, t) = \operatorname{Re} \{ \tilde{I}(z) e^{j\omega t} \}$$

Now we see:

$$\begin{aligned}\frac{\partial}{\partial z} v(z, t) &= \frac{\partial}{\partial z} \operatorname{Re} \{ \tilde{V}(z) e^{j\omega t} \} \\ &= \operatorname{Re} \left\{ \left[\frac{\partial}{\partial z} \tilde{V}(z) \right] e^{j\omega t} \right\}\end{aligned}$$

In other words, $\partial v(z, t)/\partial z$ expressed in phasor representation is simply $\partial \tilde{V}(z)/\partial z$; and

$$\begin{aligned}\frac{\partial}{\partial t} i(z, t) &= \frac{\partial}{\partial t} \operatorname{Re} \{ \tilde{I}(z) e^{j\omega t} \} \\ &= \operatorname{Re} \left\{ \frac{\partial}{\partial t} [\tilde{I}(z) e^{j\omega t}] \right\} \\ &= \operatorname{Re} \{ [j\omega \tilde{I}(z)] e^{j\omega t} \}\end{aligned}$$

In other words, $\partial i(z, t)/\partial t$ expressed in phasor representation is $j\omega \tilde{I}(z)$. Therefore, Equation 21.7.1 expressed in phasor representation is:

$$\boxed{-\frac{\partial}{\partial z} \tilde{V}(z) = [R' + j\omega L'] \tilde{I}(z)} \quad (21.7.3)$$

Following the same procedure, Equation 21.7.2 expressed in phasor representation is found to be:

$$\boxed{-\frac{\partial}{\partial z} \tilde{I}(z) = [G' + j\omega C'] \tilde{V}(z)} \quad (21.7.4)$$

Equations 21.7.3 and 21.7.4 are the telegrapher's equations in phasor representation.

The principal advantage of these equations over the time-domain versions is that we no longer need to contend with derivatives with respect to time – only derivatives with respect to distance remain. This considerably simplifies the equations.

1. For a refresher on phasor analysis, see Section 1.5. ↪

This page titled [21.7: Telegrapher's Equations](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.5: Telegrapher's Equations](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.8: Wave Equation for a Transmission Line

Consider a TEM transmission line aligned along the z axis. The phasor form of the Telegrapher's Equations (Section 3.5) relate the potential phasor $\tilde{V}(z)$ and the current phasor $\tilde{I}(z)$ to each other and to the lumped-element model equivalent circuit parameters R' , G' , C' , and L' . These equations are

$$-\frac{\partial}{\partial z} \tilde{V}(z) = [R' + j\omega L'] \tilde{I}(z) \quad (21.8.1)$$

$$-\frac{\partial}{\partial z} \tilde{I}(z) = [G' + j\omega C'] \tilde{V}(z) \quad (21.8.2)$$

An obstacle to using these equations is that we require both equations to solve for either the potential or the current. In this section, we reduce these equations to a single equation – a *wave equation* – that is more convenient to use and provides some additional physical insight.

We begin by differentiating both sides of Equation 21.8.1 with respect to z , yielding:

$$-\frac{\partial^2}{\partial z^2} \tilde{V}(z) = [R' + j\omega L'] \frac{\partial}{\partial z} \tilde{I}(z)$$

Then using Equation 21.8.2 to eliminate $\tilde{I}(z)$, we obtain

$$-\frac{\partial^2}{\partial z^2} \tilde{V}(z) = -[R' + j\omega L'][G' + j\omega C'] \tilde{V}(z)$$

This equation is normally written as follows:

$$\boxed{\frac{\partial^2}{\partial z^2} \tilde{V}(z) - \gamma^2 \tilde{V}(z) = 0} \quad (21.8.3)$$

where we have made the substitution:

$$\gamma^2 = (R' + j\omega L')(G' + j\omega C')$$

The principal square root of γ^2 is known as the *propagation constant*:

$$\gamma \triangleq \sqrt{(R' + j\omega L')(G' + j\omega C')} \quad (21.8.4)$$

The *propagation constant* γ (units of m^{-1}) captures the effect of materials, geometry, and frequency in determining the variation in potential and current with distance on a TEM transmission line.

Following essentially the same procedure but beginning with Equation 21.8.2, we obtain

$$\boxed{\frac{\partial^2}{\partial z^2} \tilde{I}(z) - \gamma^2 \tilde{I}(z) = 0} \quad (21.8.5)$$

Equations 21.8.3 and 21.8.5 are the *wave equations* for $\tilde{V}(z)$ and $\tilde{I}(z)$, respectively.

Note that both $\tilde{V}(z)$ and $\tilde{I}(z)$ satisfy the *same* linear homogeneous differential equation. This does *not* mean that $\tilde{V}(z)$ and $\tilde{I}(z)$ are equal. Rather, it means that $\tilde{V}(z)$ and $\tilde{I}(z)$ can differ by no more than a multiplicative constant. Since $\tilde{V}(z)$ is potential and $\tilde{I}(z)$ is current, that constant must be an impedance. This impedance is known as the *characteristic impedance* and is determined in Section 3.7.

The general solutions to Equations 21.8.3 and 21.8.5 are

$$\tilde{V}(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z} \quad (21.8.6)$$

$$\tilde{I}(z) = I_0^+ e^{-\gamma z} + I_0^- e^{+\gamma z} \quad (21.8.7)$$

where V_0^+ , V_0^- , I_0^+ , and I_0^- are complex-valued constants. It is shown in Section 3.8 that Equations 21.8.6 and 21.8.7 represent sinusoidal waves propagating in the $+z$ and $-z$ directions along the length of the line. The constants may represent sources, loads, or simply discontinuities in the materials and/or geometry of the line. The values of the constants are determined by boundary conditions; i.e., constraints on $\tilde{V}(z)$ and $\tilde{I}(z)$ at some position(s) along the line.

The reader is encouraged to verify that the Equations 21.8.6 and 21.8.7 are in fact solutions to Equations 21.8.3 and 21.8.5, respectively, for any values of the constants V_0^+ , V_0^- , I_0^+ , and I_0^- .

This page titled 21.8: Wave Equation for a Transmission Line is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Steven W. Ellingson (Virginia Tech Libraries' Open Education Initiative).

- 3.6: Wave Equation for a TEM Transmission Line by Steven W. Ellingson is licensed CC BY-SA 4.0. Original source:
<https://doi.org/10.21061/electromagnetics-vol-1>.

21.9: Characteristic Impedance of a Transmission Line

Characteristic impedance is the ratio of voltage to current for a wave that is propagating in single direction on a transmission line. This is an important parameter in the analysis and design of circuits and systems using transmission lines. In this section, we formally define this parameter and derive an expression for this parameter in terms of the equivalent circuit model introduced in Section 3.4.

Consider a transmission line aligned along the z axis. Employing some results from Section 3.6, recall that the phasor form of the wave equation in this case is

$$\frac{\partial^2}{\partial z^2} \tilde{V}(z) - \gamma^2 \tilde{V}(z) = 0 \quad (21.9.1)$$

where

$$\gamma \triangleq \sqrt{(R' + j\omega L')(G' + j\omega C')} \quad (21.9.2)$$

Equation 21.9.1 relates the potential phasor $\tilde{V}(z)$ to the equivalent circuit parameters R' , G' , C' , and L' . An equation of the same form relates the current phasor $\tilde{I}(z)$ to the equivalent circuit parameters:

$$\frac{\partial^2}{\partial z^2} \tilde{I}(z) - \gamma^2 \tilde{I}(z) = 0 \quad (21.9.3)$$

Since both $\tilde{V}(z)$ and $\tilde{I}(z)$ satisfy the *same* linear homogeneous differential equation, they may differ by no more than a multiplicative constant. Since $\tilde{V}(z)$ is potential and $\tilde{I}(z)$ is current, that constant can be expressed in units of impedance. Specifically, this is the *characteristic impedance*, so-named because it depends only on the materials and cross-sectional geometry of the transmission line – i.e., things which determine γ – and not length, excitation, termination, or position along the line.

To derive the characteristic impedance, first recall that the general solutions to Equations 21.9.1 and 21.9.3 are

$$\tilde{V}(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z} \quad (21.9.4)$$

$$\tilde{I}(z) = I_0^+ e^{-\gamma z} + I_0^- e^{+\gamma z} \quad (21.9.5)$$

where V_0^+ , V_0^- , I_0^+ , and I_0^- are complex-valued constants whose values are determined by boundary conditions; i.e., constraints on $\tilde{V}(z)$ and $\tilde{I}(z)$ at some position(s) along the line. Also, we will make use of the telegrapher's equations (Section 3.5):

$$-\frac{\partial}{\partial z} \tilde{V}(z) = [R' + j\omega L'] \tilde{I}(z) \quad (21.9.6)$$

$$-\frac{\partial}{\partial z} \tilde{I}(z) = [G' + j\omega C'] \tilde{V}(z) \quad (21.9.7)$$

We begin by differentiating Equation 21.9.4 with respect to z , which yields

$$\frac{\partial}{\partial z} \tilde{V}(z) = -\gamma [V_0^+ e^{-\gamma z} - V_0^- e^{+\gamma z}]$$

Now we use this to eliminate $\partial \tilde{V}(z)/\partial z$ in Equation 21.9.6, yielding

$$\gamma [V_0^+ e^{-\gamma z} - V_0^- e^{+\gamma z}] = [R' + j\omega L'] \tilde{I}(z)$$

Solving the above equation for $\tilde{I}(z)$ yields:

$$\tilde{I}(z) = \frac{\gamma}{R' + j\omega L'} [V_0^+ e^{-\gamma z} - V_0^- e^{+\gamma z}]$$

Comparing this to Equation 21.9.5, we note

$$I_0^+ = \frac{\gamma}{R' + j\omega L'} V_0^+$$

$$I_0^- = \frac{-\gamma}{R' + j\omega L'} V_0^-$$

We now make the substitution

$$Z_0 = \frac{R' + j\omega L'}{\gamma} \quad (21.9.8)$$

and observe

$$\boxed{\frac{V_0^+}{I_0^+} = \frac{-V_0^-}{I_0^-} \triangleq Z_0}$$

As anticipated, we have found that coefficients in the equations for potentials and currents are related by an impedance, namely, Z_0 . Characteristic impedance can be written entirely in terms of the equivalent circuit parameters by substituting Equation 21.9.2 into Equation 21.9.8, yielding:

$$\boxed{Z_0 = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}}}$$

The characteristic impedance Z_0 (Ω) is the ratio of potential to current in a wave traveling in a single direction along the transmission line.

Take care to note that Z_0 is *not* the ratio of $\tilde{V}(z)$ to $\tilde{I}(z)$ in general; rather, Z_0 relates only the potential and current waves traveling in the *same* direction.

Finally, note that transmission lines are normally designed to have a characteristic impedance that is completely real-valued – that is, with no imaginary component. This is because the imaginary component of an impedance represents energy *storage* (think of capacitors and inductors), whereas the purpose of a transmission line is energy *transfer*.

This page titled [21.9: Characteristic Impedance of a Transmission Line](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.7: Characteristic Impedance](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source:
<https://doi.org/10.21061/electromagnetics-vol-1>.

21.10: Wave Propagation on a Transmission Line

In Section 3.6, it is shown that expressions for the phasor representations of the potential and current along a transmission line are

$$\tilde{V}(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z} \quad (21.10.1)$$

$$\tilde{I}(z) = I_0^+ e^{-\gamma z} + I_0^- e^{+\gamma z} \quad (21.10.2)$$

where γ is the propagation constant and it assumed that the transmission line is aligned along the z axis. In this section, we demonstrate that these expressions represent sinusoidal waves, and point out some important features. Before attempting this section, the reader should be familiar with the contents of Sections 3.4, 3.6 and 3.7. A refresher on fundamental wave concepts (Section 1.3) may also be helpful.

We first define real-valued quantities α and β to be the real and imaginary components of γ ; i.e.,

$$\alpha \triangleq \operatorname{Re}\{\gamma\}$$

$$\beta \triangleq \operatorname{Im}\{\gamma\}$$

and subsequently

$$\gamma = \alpha + j\beta$$

Then we observe

$$e^{\pm\gamma z} = e^{\pm(\alpha+j\beta)z} = e^{\pm\alpha z} e^{\pm j\beta z}$$

It may be easier to interpret this expression by reverting to the time domain:

$$\operatorname{Re}\{e^{\pm\gamma z} e^{j\omega t}\} = e^{\pm\alpha z} \cos(\omega t \pm \beta z)$$

Thus, $e^{-\gamma z}$ represents a damped sinusoidal wave traveling in the $+z$ direction, and $e^{+\gamma z}$ represents a damped sinusoidal wave traveling in the $-z$ direction.

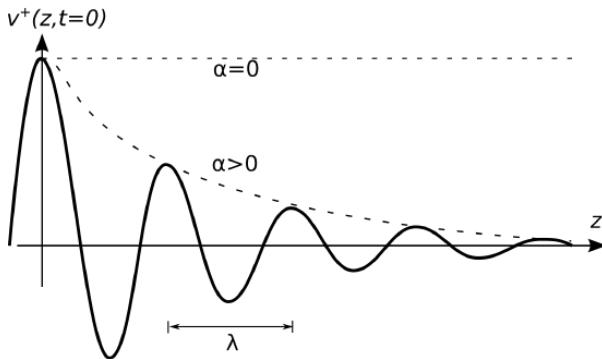


Figure 21.10.1: The potential $v^+(z, t)$ of the wave traveling in the $+z$ direction at $t = 0$ for $\psi = 0$.

Let's define $\tilde{V}^+(z)$ and $\tilde{I}^+(z)$ to be the potential and current associated with a wave propagating in the $+z$ direction. Then:

$$\tilde{V}^+(z) \triangleq V_0^+ e^{-\gamma z}$$

or equivalently in the time domain:

$$\begin{aligned} v^+(z, t) &= \operatorname{Re}\{\tilde{V}^+(z)e^{j\omega t}\} \\ &= \operatorname{Re}\{V_0^+ e^{-\gamma z} e^{j\omega t}\} \\ &= |V_0^+| e^{-\alpha z} \cos(\omega t - \beta z + \psi) \end{aligned} \quad (21.10.3)$$

where ψ is the phase of V_0^+ . Figure 21.10.1 shows $v^+(z, t)$. From fundamental wave theory we recognize

$\beta \triangleq \operatorname{Im}\{\gamma\}$ (rad/m) is the *phase propagation constant*, which is the rate at which phase changes as a function of distance.

Subsequently the wavelength in the line is

$$\lambda = \frac{2\pi}{\beta}$$

Also we recognize:

$\alpha \triangleq \operatorname{Re}\{\gamma\}$ (1/m) is the *attenuation constant*, which is the rate at which magnitude diminishes as a function of distance.

Sometimes the units of α are indicated as “Np/m” (“nepers” per meter), where the term “neper” is used to indicate the units of the otherwise unitless real-valued exponent of the constant e .

Note that $\alpha = 0$ for a wave that does not diminish in magnitude with increasing distance, in which case the transmission line is said to be *lossless*. If $\alpha > 0$ then the line is said to be *lossy* (or possibly “low loss” if the loss can be neglected), and in this case the rate at which the magnitude decreases with distance increases with α .

Next let us consider the speed of the wave. To answer this question, we need to be a bit more specific about what we mean by “speed.” At the moment, we mean phase velocity; that is, the speed at which a point of constant phase seems to move through space. In other words, what distance Δz does a point of constant phase traverse in time Δt ? To answer this question, we first note that the phase of $v^+(z, t)$ can be written generally as

$$\omega t - \beta z + \phi$$

where ϕ is some constant. Similarly, the phase at some time Δt later and some point Δz further along can be written as

$$\omega(t + \Delta t) - \beta(z + \Delta z) + \phi$$

The phase velocity v_p is $\Delta z / \Delta t$ when these two phases are equal; i.e., when

$$\omega t - \beta z + \phi = \omega(t + \Delta t) - \beta(z + \Delta z) + \phi$$

Solving for $v_p = \Delta z / \Delta t$, we obtain:

$$v_p = \frac{\omega}{\beta}$$

Having previously noted that $\beta = 2\pi/\lambda$, the above expression also yields the expected result

$$v_p = \lambda f$$

The phase velocity $v_p = \omega/\beta = \lambda f$ is the speed at which a point of constant phase travels along the line.

Returning now to consider the current associated with the wave traveling in the $+z$ direction:

$$\tilde{I}^+(z) = I_0^+ e^{-\gamma z}$$

We can rewrite this expression in terms of the characteristic impedance Z_0 , as follows:

$$\tilde{I}^+(z) = \frac{V_0^+}{Z_0} e^{-\gamma z}$$

Similarly, we find that the current $\tilde{I}^-(z)$ associated with $\tilde{V}^-(z)$ for the wave traveling in the $-z$ direction is

$$\tilde{I}^-(z) = \frac{-V_0^-}{Z_0} e^{+\gamma z}$$

The negative sign appearing in the above expression emerges as a result of the sign conventions used for potential and current in the derivation of the telegrapher's equations (Section 3.5). The physical significance of this change of sign is that wherever the potential of the wave traveling in the $-z$ direction is positive, then the current at the same point is flowing in the $-z$ direction.

It is frequently necessary to consider the possibility that waves travel in both directions simultaneously. A very important case where this arises is when there is reflection from a discontinuity of some kind; e.g., from a termination which is not perfectly impedance-matched. In this case, the total potential $\tilde{V}(z)$ and total current $\tilde{I}(z)$ can be expressed as the general solution to the wave equation; i.e., as the sum of the “incident” ($+z$ -traveling) wave and the reflected ($-z$ -traveling) waves:

$$\tilde{V}(z) = \tilde{V}^+(z) + \tilde{V}^-(z)$$

$$\tilde{I}(z) = \tilde{I}^+(z) + \tilde{I}^-(z)$$

The existence of waves propagating simultaneously in both directions gives rise to a phenomenon known as a *standing wave*. Standing waves and the calculation of the coefficients V_0^- and I_0^- due to reflection are addressed in Sections 3.13 and 3.12 respectively.

This page titled [21.10: Wave Propagation on a Transmission Line](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.8: Wave Propagation on a TEM Transmission Line](#) by Steven W. Ellingson is licensed [CC BY-SA 4.0](#). Original source:
<https://doi.org/10.21061/electromagnetics-vol-1>.

21.11: Lossless and Low-Loss Transmission Lines

Quite often the loss in a transmission line is small enough that it may be neglected. In this case, several aspects of transmission line theory may be simplified. In this section, we present these simplifications.

First, recall that “loss” refers to the reduction of magnitude as a wave propagates through space. In the lumped-element equivalent circuit model (Section 3.4), the parameters R' and G' of the represent physical mechanisms associated with loss. Specifically, R' represents the resistance of conductors, whereas G' represents the undesirable current induced between conductors through the spacing material. Also recall that the propagation constant γ is, in general, given by

$$\gamma \triangleq \sqrt{(R' + j\omega L')(G' + j\omega C')}$$

With this in mind, we now define “low loss” as meeting the conditions:

$$R' \ll \omega L' \quad (21.11.1)$$

$$G' \ll \omega C' \quad (21.11.2)$$

When these conditions are met, the propagation constant simplifies as follows:

$$\begin{aligned} \gamma &\approx \sqrt{(j\omega L')(j\omega C')} \\ &= \sqrt{-\omega^2 L' C'} \\ &= j\omega \sqrt{L' C'} \end{aligned} \quad (21.11.3)$$

and subsequently

$$\alpha \triangleq \operatorname{Re}\{\gamma\} \approx 0 \quad (\text{low-loss approx.}) \quad (21.11.4)$$

$$\beta \triangleq \operatorname{Im}\{\gamma\} \approx \omega \sqrt{L' C'} \quad (\text{low-loss approx.}) \quad (21.11.5)$$

$$v_p = \omega / \beta \approx \frac{1}{\sqrt{L' C'}} \quad (\text{low-loss approx.}) \quad (21.11.6)$$

Similarly:

$$Z_0 = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}} \approx \sqrt{\frac{L'}{C'}} \quad (\text{low-loss approx.})$$

Of course if the line is strictly lossless (i.e., $R' = G' = 0$) then these are not approximations, but rather the exact expressions.

In practice, these approximations are quite commonly used, since practical transmission lines typically meet the conditions expressed in Inequalities 21.11.1 and 21.11.2 and the resulting expressions are much simpler. We further observe that Z_0 and v_p are approximately independent of frequency when these conditions hold.

However, also note that “low loss” does not mean “no loss,” and it is common to apply these expressions even when R' and/or G' is large enough to yield significant loss. For example, a coaxial cable used to connect an antenna on a tower to a radio near the ground typically has loss that is important to consider in the analysis and design process, but nevertheless satisfies Equations 21.11.1 and 21.11.2. In this case, the low-loss expression for β is used, but α might not be approximated as zero.

This page titled [21.11: Lossless and Low-Loss Transmission Lines](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.9: Lossless and Low-Loss Transmission Lines](#) by Steven W. Ellingson is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.12: Voltage Reflection Coefficient

We now consider the scenario shown in Figure 21.12.1. Here a wave arriving from the left along a lossless transmission line having characteristic impedance Z_0 arrives at a termination located at $z = 0$. The impedance looking into the termination is Z_L , which may be real-, imaginary-, or complex-valued. The questions are: Under what circumstances is a reflection – i.e., a leftward traveling wave – expected, and what precisely is that wave?

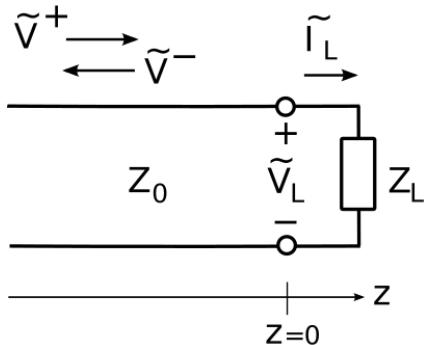


Figure 21.12.1: A wave arriving from the left incident on a termination located at $z = 0$.

The potential and current of the incident wave are related by the constant value of Z_0 . Similarly, the potential and current of the reflected wave are related by Z_0 . Therefore, it suffices to consider *either* potential or current. Choosing potential, we may express the incident wave as

$$\tilde{V}^+(z) = V_0^+ e^{-j\beta z}$$

where V_0^+ is determined by the source of the wave, and so is effectively a “given.” Any reflected wave must have the form

$$\tilde{V}^-(z) = V_0^- e^{+j\beta z}$$

Therefore, the problem is solved by determining the value of V_0^- given V_0^+ , Z_0 , and Z_L .

Considering the situation at $z = 0$, note that by definition we have

$$Z_L \triangleq \frac{\tilde{V}_L}{\tilde{I}_L} \quad (21.12.1)$$

where \tilde{V}_L and \tilde{I}_L are the potential across and current through the termination, respectively. Also, the potential and current on either side of the $z = 0$ interface must be equal. Thus,

$$\tilde{V}^+(0) + \tilde{V}^-(0) = \tilde{V}_L \quad (21.12.2)$$

$$\tilde{I}^+(0) + \tilde{I}^-(0) = \tilde{I}_L \quad (21.12.3)$$

where $\tilde{I}^+(z)$ and $\tilde{I}^-(z)$ are the currents associated with $\tilde{V}^+(z)$ and $\tilde{V}^-(z)$, respectively. Since the voltage and current are related by Z_0 , Equation 21.12.3 may be rewritten as follows:

$$\frac{\tilde{V}^+(0)}{Z_0} - \frac{\tilde{V}^-(0)}{Z_0} = \tilde{I}_L \quad (21.12.4)$$

Evaluating the left sides of Equations 21.12.2 and 21.12.4 at $z = 0$, we find:

$$\begin{aligned} V_0^+ + V_0^- &= \tilde{V}_L \\ \frac{V_0^+}{Z_0} - \frac{V_0^-}{Z_0} &= \tilde{I}_L \end{aligned} \quad (21.12.5)$$

Substituting these expressions into Equation 21.12.1 we obtain:

$$Z_L = \frac{V_0^+ + V_0^-}{V_0^+/Z_0 - V_0^-/Z_0}$$

Solving for V_0^- we obtain

$$V_0^- = \frac{Z_L - Z_0}{Z_L + Z_0} V_0^+$$

Thus, the answer to the question posed earlier is that

$$V_0^- = \Gamma V_0^+ , \text{ where}$$

$$\boxed{\Gamma \triangleq \frac{Z_L - Z_0}{Z_L + Z_0}} \quad (21.12.6)$$

The quantity Γ is known as the *voltage reflection coefficient*. Note that when $Z_L = Z_0$, $\Gamma = 0$ and therefore $V_0^- = 0$. In other words,

If the terminating impedance is equal to the characteristic impedance of the transmission line, then there is no reflection.

If, on the other hand, $Z_L \neq Z_0$, then $|\Gamma| > 0$, $V_0^- = \Gamma V_0^+$, and a leftward-traveling reflected wave exists.

Since Z_L may be real-, imaginary-, or complex-valued, Γ too may be real-, imaginary-, or complex-valued. Therefore, V_0^- may be different from V_0^+ in magnitude, sign, or phase.

Note also that Γ is *not* the ratio of I_0^- to I_0^+ . The ratio of the *current* coefficients is actually $-\Gamma$. It is quite simple to show this with a simple modification to the above procedure and is left as an exercise for the student.

Summarizing:

The voltage reflection coefficient Γ , given by Equation 21.12.6, determines the magnitude and phase of the reflected wave given the incident wave, the characteristic impedance of the transmission line, and the terminating impedance.

We now consider values Γ that arise for commonly-encountered terminations.

Matched Load. ($Z_L = Z_0$). In this case, the termination may be a device with impedance Z_0 , or the termination may be another transmission line having the same characteristic impedance. When $Z_L = Z_0$, $\Gamma = 0$ and there is no reflection.

Open Circuit. An “open circuit” is the absence of a termination. This condition implies $Z_L \rightarrow \infty$, and subsequently $\Gamma \rightarrow +1$. Since the *current* reflection coefficient is $-\Gamma$, the reflected current wave is 180° out of phase with the incident current wave, making the total current at the open circuit equal to zero, as expected.

Short Circuit. “Short circuit” means $Z_L = 0$, and subsequently $\Gamma = -1$. In this case, the phase of Γ is 180° , and therefore, the potential of the reflected wave cancels the potential of the incident wave at the open circuit, making the total potential equal to zero, as it must be. Since the *current* reflection coefficient is $-\Gamma = +1$ in this case, the reflected current wave is in phase with the incident current wave, and the magnitude of the total current at the short circuit non-zero as expected.

Purely Reactive Load. A purely reactive load, including that presented by a capacitor or inductor, has $Z_L = jX$ where X is reactance. In particular, an inductor is represented by $X > 0$ and a capacitor is represented by $X < 0$. We find

$$\Gamma = \frac{-Z_0 + jX}{+Z_0 + jX}$$

The numerator and denominator have the same magnitude, so $|\Gamma| = 1$. Let ϕ be the phase of the denominator ($+Z_0 + jX$). Then, the phase of the numerator is $\pi - \phi$. Subsequently, the phase of Γ is $(\pi - \phi) - \phi = \pi - 2\phi$. Thus, we see that the phase of Γ is no longer limited to be 0° or 180° , but can be any value in between. The phase of reflected wave is subsequently shifted by this amount.

Other Terminations. Any other termination, including series and parallel combinations of any number of devices, can be expressed as a value of Z_L which is, in general, complex-valued. The associated value of $|\Gamma|$ is limited to the range 0 to 1. To see this, note:

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{Z_L/Z_0 - 1}{Z_L/Z_0 + 1}$$

Note that the smallest possible value of $|\Gamma|$ occurs when the numerator is zero; i.e., when $Z_L = Z_0$. Therefore, the smallest value of $|\Gamma|$ is zero. The largest possible value of $|\Gamma|$ occurs when $Z_L/Z_0 \rightarrow \infty$ (i.e., an open circuit) or when $Z_L/Z_0 = 0$ (a short circuit); the result in either case is $|\Gamma| = 1$. Thus,

$$0 \leq |\Gamma| \leq 1$$

This page titled [21.12: Voltage Reflection Coefficient](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.12: Voltage Reflection Coefficient](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.13: Standing Waves

A standing wave consists of waves moving in opposite directions. These waves add to make a distinct magnitude variation as a function of distance that does not vary in time.

To see how this can happen, first consider that an incident wave $V_0^+ e^{-j\beta z}$, which is traveling in the $+z$ axis along a lossless transmission line. Associated with this wave is a reflected wave $V_0^- e^{+j\beta z} = \Gamma V_0^+ e^{+j\beta z}$, where Γ is the voltage reflection coefficient. These waves add to make the total potential

$$\begin{aligned}\tilde{V}(z) &= V_0^+ e^{-j\beta z} + \Gamma V_0^+ e^{+j\beta z} \\ &= V_0^+ (e^{-j\beta z} + \Gamma e^{+j\beta z})\end{aligned}$$

The magnitude of $\tilde{V}(z)$ is most easily found by first finding $|\tilde{V}(z)|^2$, which is:

$$\begin{aligned}\tilde{V}(z)\tilde{V}^*(z) &= |V_0^+|^2 (e^{-j\beta z} + \Gamma e^{+j\beta z})(e^{-j\beta z} + \Gamma e^{+j\beta z})^* \\ &= |V_0^+|^2 (e^{-j\beta z} + \Gamma e^{+j\beta z})(e^{+j\beta z} + \Gamma^* e^{-j\beta z}) \\ &= |V_0^+|^2 (1 + |\Gamma|^2 + \Gamma e^{+j2\beta z} + \Gamma^* e^{-j2\beta z})\end{aligned}\tag{21.13.1}$$

Let ϕ be the phase of Γ ; i.e.,

$$\Gamma = |\Gamma| e^{j\phi}$$

Then, continuing from the previous expression:

$$\begin{aligned}|V_0^+|^2 (1 + |\Gamma|^2 + |\Gamma| e^{+j(2\beta z + \phi)} + |\Gamma| e^{-j(2\beta z + \phi)}) \\ = |V_0^+|^2 (1 + |\Gamma|^2 + |\Gamma| [e^{+j(2\beta z + \phi)} + e^{-j(2\beta z + \phi)}])\end{aligned}\tag{21.13.2}$$

The quantity in square brackets can be reduced to a cosine function using the identity

$$\cos \theta = \frac{1}{2} [e^{j\theta} + e^{-j\theta}]$$

yielding:

$$|V_0^+|^2 [1 + |\Gamma|^2 + 2|\Gamma| \cos(2\beta z + \phi)]$$

Recall that this is $|\tilde{V}(z)|^2$. $|\tilde{V}(z)|$ is therefore the square root of the above expression:

$$|\tilde{V}(z)| = |V_0^+| \sqrt{1 + |\Gamma|^2 + 2|\Gamma| \cos(2\beta z + \phi)}$$

Thus, we have found that the magnitude of the resulting total potential varies sinusoidally along the line. This is referred to as a standing wave because the variation of the magnitude of the phasor resulting from the interference between the incident and reflected waves does not vary with time.

We may perform a similar analysis of the current, leading to:

$$|\tilde{I}(z)| = \frac{|V_0^+|}{Z_0} \sqrt{1 + |\Gamma|^2 - 2|\Gamma| \cos(2\beta z + \phi)}$$

Again we find the result is a standing wave.

Now let us consider the outcome for a few special cases.

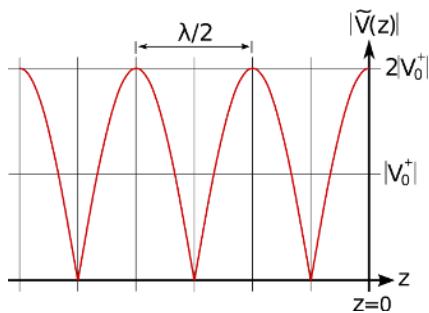
Matched load. When the impedance of the termination of the transmission line, Z_L , is equal to the characteristic impedance of the transmission line, Z_0 , $\Gamma = 0$ and there is no reflection. In this case, the above expressions reduce to $|\tilde{V}(z)| = |V_0^+|$ and $|\tilde{I}(z)| = |V_0^+|/Z_0$, as expected.

Open or Short-Circuit. In this case, $\Gamma = \pm 1$ and we find:

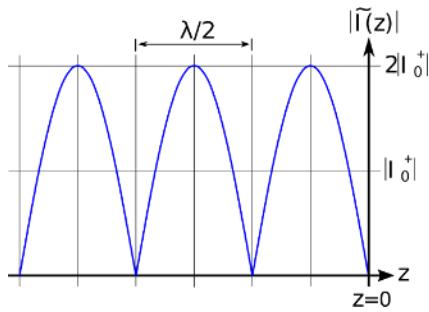
$$|\tilde{V}(z)| = |V_0^+| \sqrt{2 + 2 \cos(2\beta z + \phi)}$$

$$|\tilde{I}(z)| = \frac{|V_0^+|}{Z_0} \sqrt{2 - 2 \cos(2\beta z + \phi)}$$

where $\phi = 0$ for an open circuit and $\phi = \pi$ for a short circuit. The result for an open circuit termination is shown in Figure 21.13.1(a) (potential) and 21.13.1(b) (current). The result for a short circuit termination is identical except the roles of potential and current are reversed. In either case, note that voltage maxima correspond to current minima, and vice versa.



(a) Potential.



(b) Current.

Figure 21.13.1: Standing wave associated with an opencircuit termination at $z = 0$ (incident wave arrives from left).

Also note:

The period of the standing wave is $\lambda/2$; i.e., one-half of a wavelength.

This can be confirmed as follows. First, note that the frequency argument of the cosine function of the standing wave is $2\beta z$. This can be rewritten as $2\pi(\beta/\pi) z$, so the frequency of variation is β/π and the period of the variation is π/β . Since $\beta = 2\pi/\lambda$, we see that the period of the variation is $\lambda/2$. Furthermore, this is true regardless of the value of Γ .

Mismatched loads. A common situation is that the termination is neither perfectly-matched ($\Gamma = 0$) nor an open/short circuit ($|\Gamma| = 1$). Examples of the resulting standing waves are shown in Figure 21.13.2.

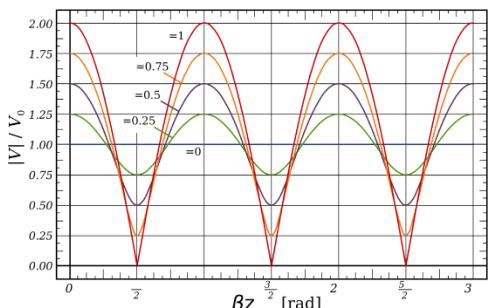


Figure 21.13.2: Standing waves associated with loads exhibiting various reflection coefficients. In this figure the incident wave arrives from the right.

This page titled [21.13: Standing Waves](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.13: Standing Waves](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.14: Standing Wave Ratio

Precise matching of transmission lines to terminations is often not practical or possible. Whenever a significant mismatch exists, a standing wave (Section 3.13) is apparent. The quality of the match is commonly expressed in terms of the *standing wave ratio* (SWR) of this standing wave.

Standing wave ratio (SWR) is defined as the ratio of the maximum magnitude of the standing wave to minimum magnitude of the standing wave.

In terms of the potential:

$$\text{SWR} \triangleq \frac{\text{maximum } |\tilde{V}|}{\text{minimum } |\tilde{V}|}$$

SWR can be calculated using a simple expression, which we shall now derive. In Section 3.13, we found that:

$$|\tilde{V}(z)| = |V_0^+| \sqrt{1 + |\Gamma|^2 + 2|\Gamma| \cos(2\beta z + \phi)}$$

The maximum value occurs when the cosine factor is equal to **+1**, yielding:

$$\max |\tilde{V}| = |V_0^+| \sqrt{1 + |\Gamma|^2 + 2|\Gamma|}$$

Note that the argument of the square root operator is equal to $(1 + |\Gamma|)^2$; therefore:

$$\max |\tilde{V}| = |V_0^+| (1 + |\Gamma|)$$

Similarly, the minimum value is achieved when the cosine factor is equal to **-1**, yielding:

$$\min |\tilde{V}| = |V_0^+| \sqrt{1 + |\Gamma|^2 - 2|\Gamma|}$$

So:

$$\min |\tilde{V}| = |V_0^+| (1 - |\Gamma|)$$

Therefore:

$$\text{SWR} = \frac{1 + |\Gamma|}{1 - |\Gamma|}$$

(21.14.1)

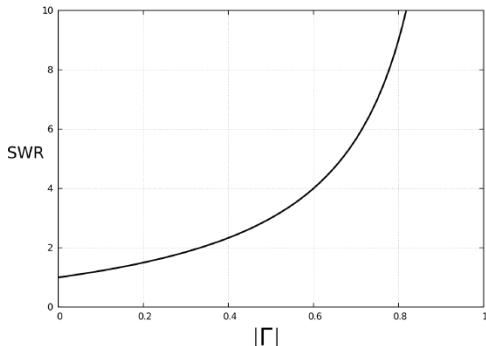


Figure 21.14.1: Relationship between SWR and $|\Gamma|$.

This relationship is shown graphically in Figure 21.14.1. Note that SWR ranges from 1 for perfectly-matched terminations ($\Gamma = 0$) to infinity for open- and short-circuit terminations ($|\Gamma| = 1$).

It is sometimes of interest to find the magnitude of the reflection coefficient given SWR. Solving Equation 21.14.1 for $|\Gamma|$ we find:

$$|\Gamma| = \frac{\text{SWR} - 1}{\text{SWR} + 1} \quad (21.14.2)$$

SWR is often referred to as the *voltage standing wave ratio* (VSWR), although repeating the analysis above for the current reveals that the current SWR is equal to potential SWR, so the term “SWR” suffices.

SWR < 2 or so is usually considered a “good match,” although some applications require SWR < 1.1 or better, and other applications are tolerant to SWR of 3 or greater.

✓ Example 21.14.1: Reflection Coefficient for Various Values of SWR

What is the reflection coefficient for the above-cited values of SWR? Using Equation 21.14.2, we find:

- SWR = 1.1 corresponds to $|\Gamma| = 0.0476$.
- SWR = 2.0 corresponds to $|\Gamma| = 1/3$.
- SWR = 3.0 corresponds to $|\Gamma| = 1/2$.

This page titled [21.14: Standing Wave Ratio](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.14: Standing Wave Ratio](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-1>.

21.15: Parallel Wire Transmission Line

A parallel wire transmission line consists of wires separated by a dielectric spacer. Figure 21.15.1 shows a common implementation, commonly known as “twin lead.” The wires in twin lead line are held in place by a mechanical spacer comprised of the same low-loss dielectric material that forms the jacket of each wire. Very little of the total energy associated with the electric and magnetic fields lies inside this material, so the jacket and spacer can usually be neglected for the purposes of analysis and electrical design.



Figure 21.15.1: Twin lead, a commonly-encountered form of parallel wire transmission line. (CC BY SA 3.0 (modified); SpinningSpark)

Parallel wire transmission line is often employed in radio applications up to about 100 MHz as an alternative to coaxial line. Parallel wire line has the advantages of lower cost and lower loss than coaxial line in this frequency range. However, parallel wire line lacks the self-shielding property of coaxial cable; i.e., the electromagnetic fields of coaxial line are isolated by the outer conductor, whereas those of parallel wire line are exposed and prone to interaction with nearby structures and devices. This prevents the use of parallel wire line in many applications.

Another discriminator between parallel wire line and coaxial line is that parallel wire line is differential.¹ The conductor geometry is symmetric and neither conductor is favored as a signal datum (“ground”). Thus, parallel wire line is commonly used in applications where the signal sources and/or loads are also differential; common examples are the dipole antenna and differential amplifiers.²

Figure 21.15.2 shows a cross-section of parallel wire line.

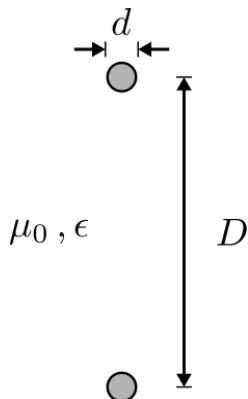


Figure 21.15.2: Parallel wire transmission line structure and design parameters. (CC BY SA 4.0; C. Wang)

Relevant parameters include the wire diameter, d ; and the center-to-center spacing, D .

The associated field structure is transverse electromagnetic (TEM) and is therefore completely described by a single cross-section along the line, as shown in Figure 21.15.3.

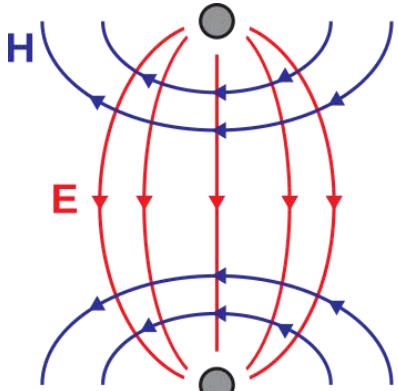


Figure 21.15.3: Structure of the electric and magnetic fields for a cross-section of parallel wire line. In this case, the wave is propagating away from the viewer. (CC BY SA 4.0; S. Lally)

Expressions for these fields exist, but are complex and not particularly useful except as a means to calculate other parameters of interest. One of these parameters is, of course, the characteristic impedance since this parameter plays an important role in the analysis and design of systems employing transmission lines. The characteristic impedance may be determined using the “lumped element” transmission line model using the following expression:

$$Z_0 = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}}$$

where R' , G' , C' , and L' are the resistance, conductance, capacitance, and inductance per unit length, respectively. This analysis is considerably simplified by neglecting loss; therefore, let us assume the “low-loss” conditions $R' \ll \omega L'$ and $G' \ll \omega C'$. Then we find:

$$Z_0 \approx \sqrt{\frac{L'}{C'}} \quad (\text{low loss}) \tag{21.15.1}$$

and the problem is reduced to determining inductance and capacitance of the transmission line. These are

$$L' = \frac{\mu_0}{\pi} \ln \left[(D/d) + \sqrt{(D/d)^2 - 1} \right]$$

$$C' = \frac{\pi \epsilon}{\ln \left[(D/d) + \sqrt{(D/d)^2 - 1} \right]}$$

Because the wire separation D is typically much greater than the wire diameter d , $D/d \gg 1$ and so $\sqrt{(D/d)^2 - 1} \approx D/d$. This leads to the simplified expressions

$$L' \approx \frac{\mu_0}{\pi} \ln (2D/d) \quad (D \gg d)$$

$$C' \approx \frac{\pi \epsilon}{\ln (2D/d)} \quad (D \gg d)$$

Now returning to Equation 21.15.1:

$$Z_0 \approx \frac{1}{\pi} \sqrt{\frac{\mu_0}{\epsilon}} \ln (2D/d)$$

Noting that $\epsilon = \epsilon_r \epsilon_0$ and $\sqrt{\mu_0/\epsilon_0} \triangleq \eta_0$, we obtain

$Z_0 \approx \frac{1}{\pi} \frac{\eta_0}{\sqrt{\epsilon_r}} \ln (2D/d)$

(21.15.2)

The characteristic impedance of parallel wire line, assuming low-loss conditions and wire spacing much greater than wire diameter, is given by Equation 21.15.2.

Observe that the characteristic impedance of parallel wire line increases with increasing D/d . Since this ratio is large, the characteristic impedance of parallel wire line tends to be large relative to common values of other kinds of TEM transmission line, such as coaxial line and microstrip line. An example follows.

✓ Example 21.15.1: 300 Ω twin-lead

A commonly-encountered form of parallel wire transmission line is 300 Ω twin-lead. Although implementations vary, the wire diameter is usually about 1 mm and the wire spacing is usually about 6 mm. The relative permittivity of the medium $\epsilon_r \approx 1$ for the purposes of calculating transmission line parameters, since the jacket and spacer have only a small effect on the fields. For these values, Equation 21.15.2 gives $Z_0 \approx 298 \Omega$, as expected.

Under the assumption that the wire jacket/spacer material has a negligible effect on the electromagnetic fields, and that the line is suspended in air so that $\epsilon_r \approx 1$, the phase velocity v_p for a parallel wire line is approximately that of any electromagnetic wave in free space; i.e., c . In practical twin-lead, the effect of a plastic jacket/spacer material is to reduce the phase velocity by a few percent up to about 20%, depending on the materials and details of construction. So in practice $v_p \approx 0.8c$ to $0.9c$ for twin-lead line.

Additional Reading:

- “Twin-lead” on Wikipedia.
- “Differential signaling” on Wikipedia.
- Sec. 8.7 (“Differential Circuits”) in S.W. Ellingson, *Radio Systems Engineering*, Cambridge Univ. Press, 2016.

-
1. The references in “Additional Reading” at the end of this section may be helpful if you are not familiar with this concept. ↪
 2. This is in contrast to “single-ended” line such as coaxial line, which has conductors of different cross-sections and the outer conductor is favored as the datum. ↪

This page titled [21.15: Parallel Wire Transmission Line](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [7.1: Parallel Wire Transmission Line](#) by Steven W. Ellingson is licensed [CC BY-SA 4.0](#). Original source:
<https://doi.org/10.21061/electromagnetics-vol-2>.

21.16: Attenuation in Coaxial Cable

In this section, we consider the issue of attenuation in coaxial transmission line. Recall that attenuation can be interpreted in the context of the “lumped element” equivalent circuit transmission line model as the contributions of the resistance per unit length R' and conductance per unit length G' . In this model, R' represents the physical resistance in the inner and outer conductors, whereas G' represents loss due to current flowing directly between the conductors through the spacer material.

The parameters used to describe the relevant features of coaxial cable are shown in Figure 21.16.1. In this figure, a and b are the radii of the inner and outer conductors, respectively. σ_{ic} and σ_{oc} are the conductivities (SI base units of S/m) of the inner and outer conductors, respectively. Conductors are assumed to be non-magnetic; i.e., having permeability μ equal to the free space value μ_0 . The spacer material is assumed to be a lossy dielectric having relative permittivity ϵ_r and conductivity σ_s .

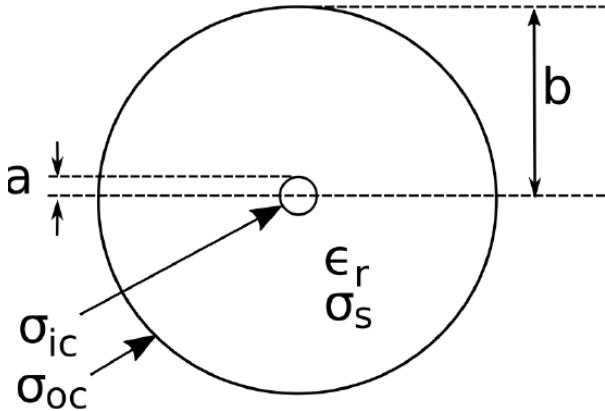


Figure 21.16.1: Parameters defining the design of a coaxial cable.

Resistance per unit length

The resistance per unit length is the sum of the resistances of the inner and outer conductor per unit length. The resistance per unit length of the inner conductor is determined by σ_{ic} and the effective cross-sectional area through which the current flows. The latter is equal to the circumference $2\pi a$ times the skin depth δ_{ic} of the inner conductor, so:

$$R'_{ic} \approx \frac{1}{(2\pi a \cdot \delta_{ic}) \sigma_{ic}} \quad \text{for } \delta_{ic} \ll a$$

This expression is only valid for $\delta_{ic} \ll a$ because otherwise the cross-sectional area through which the current flows is not well-modeled as a thin ring near the surface of the conductor. Similarly, we find the resistance per unit length of the outer conductor is

$$R'_{oc} \approx \frac{1}{(2\pi b \cdot \delta_{oc}) \sigma_{oc}} \quad \text{for } \delta_{oc} \ll t$$

where δ_{oc} is the skin depth of the outer conductor and t is the thickness of the outer conductor. Therefore, the total resistance per unit length is

$$\begin{aligned} R' &= R'_{ic} + R'_{oc} \\ &\approx \frac{1}{(2\pi a \cdot \delta_{ic}) \sigma_{ic}} + \frac{1}{(2\pi b \cdot \delta_{oc}) \sigma_{oc}} \end{aligned} \tag{21.16.1}$$

Recall that skin depth depends on conductivity. Specifically:

$$\delta_{ic} = \sqrt{\frac{2}{\omega \mu \sigma_{ic}}} \tag{21.16.2}$$

$$\delta_{oc} = \sqrt{\frac{2}{\omega \mu \sigma_{oc}}} \tag{21.16.3}$$

Expanding Equation 21.16.1 to show explicitly the dependence on conductivity, we find:

$$R' \approx \frac{1}{2\pi \sqrt{2/\omega \mu_0}} \left[\frac{1}{a \sqrt{\sigma_{ic}}} + \frac{1}{b \sqrt{\sigma_{oc}}} \right]$$

At this point it is convenient to identify two particular cases for the design of the cable. In the first case, “Case I,” we assume $\sigma_{oc} \gg \sigma_{ic}$. Since $b > a$, we have in this case

$$\begin{aligned} R' &\approx \frac{1}{2\pi\sqrt{2/\omega\mu_0}} \left[\frac{1}{a\sqrt{\sigma_{ic}}} \right] \\ &= \frac{1}{2\pi\delta_{ic}\sigma_{ic}} \frac{1}{a} \quad (\text{Case I}) \end{aligned} \quad (21.16.4)$$

In the second case, “Case II,” we assume $\sigma_{oc} = \sigma_{ic}$. In this case, we have

$$\begin{aligned} R' &\approx \frac{1}{2\pi\sqrt{2/\omega\mu_0}} \left[\frac{1}{a\sqrt{\sigma_{ic}}} + \frac{1}{b\sqrt{\sigma_{ic}}} \right] \\ &= \frac{1}{2\pi\delta_{ic}\sigma_{ic}} \left[\frac{1}{a} + \frac{1}{b} \right] \quad (\text{Case II}) \end{aligned} \quad (21.16.5)$$

A simpler way to deal with these two cases is to represent them both using the single expression

$$R' \approx \frac{1}{2\pi\delta_{ic}\sigma_{ic}} \left[\frac{1}{a} + \frac{C}{b} \right]$$

where $C = 0$ in Case I and $C = 1$ in Case II.

Conductance per unit length

The conductance per unit length of coaxial cable is simply that of the associated coaxial structure at DC; i.e.,

$$G' = \frac{2\pi\sigma_s}{\ln(b/a)}$$

Unlike resistance, the conductance is independent of frequency, at least to the extent that σ_s is independent of frequency.

Attenuation

The attenuation of voltage and current waves as they propagate along the cable is represented by the factor $e^{-\alpha z}$, where z is distance traversed along the cable. It is possible to find an expression for α in terms of the material and geometry parameters using:

$$\gamma \triangleq \sqrt{(R' + j\omega L')(G' + j\omega C')} = \alpha + j\beta \quad (21.16.6)$$

where L' and C' are the inductance per unit length and capacitance per unit length, respectively. These are given by

$$L' = \frac{\mu}{2\pi} \ln(b/a)$$

and

$$C' = \frac{2\pi\epsilon_0\epsilon_r}{\ln(b/a)}$$

In principle we could solve Equation 21.16.6 for α . However, this course of action is quite tedious, and a simpler approximate approach facilitates some additional insights. In this approach, we define parameters α_R associated with R' and α_G associated with G' such that

$$e^{-\alpha_R z} e^{-\alpha_G z} = e^{-(\alpha_R + \alpha_G)z} = e^{-\alpha z}$$

which indicates

$$\alpha = \alpha_R + \alpha_G$$

Next we postulate

$$\alpha_R \approx K_R \frac{R'}{Z_0} \quad (21.16.7)$$

where Z_0 is the characteristic impedance

$$Z_0 \approx \frac{\eta_0}{2\pi} \frac{1}{\sqrt{\epsilon_r}} \ln \frac{b}{a} \quad (\text{low loss}) \quad (21.16.8)$$

and where K_R is a unitless constant to be determined. The justification for Equation 21.16.7 is as follows: First, α_R must increase monotonically with increasing R' . Second, R' must be divided by an impedance in order to obtain the correct units of 1/m. Using similar reasoning, we postulate

$$\alpha_G \approx K_G G' Z_0 \quad (21.16.9)$$

where K_G is a unitless constant to be determined. The following example demonstrates the validity of Equations 21.16.7 and 21.16.9, and will reveal the values of K_R and K_G .

✓ Example 21.16.1: Attenuation constant for RG-59

RG-59 is a popular form of coaxial cable having the parameters $a \approx 0.292$ mm, $b \approx 1.855$ mm, $\sigma_{ic} \approx 2.28 \times 10^7$ S/m, $\sigma_s \approx 5.9 \times 10^{-5}$ S/m, and $\epsilon_r \approx 2.25$. The conductivity σ_{oc} of the outer conductor is difficult to quantify because it consists of a braid of thin metal strands. However, $\sigma_{oc} \gg \sigma_{ic}$, so we may assume Case I; i.e., $\sigma_{oc} \gg \sigma_{ic}$, and subsequently $C = 0$.

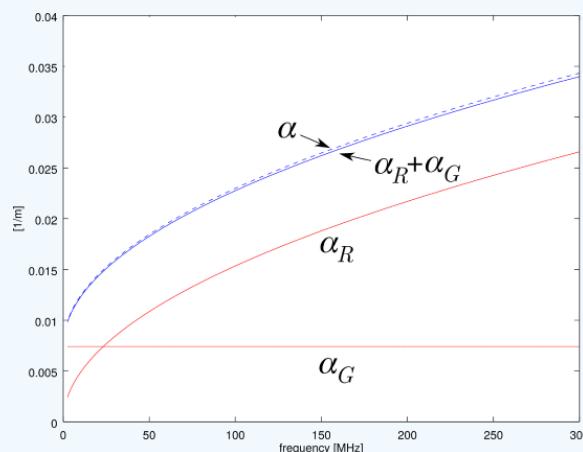


Figure 21.16.2: Comparison of $\alpha = \text{Re}\{\gamma\}$ to α_R , α_G , and $\alpha_R + \alpha_G$ for $K_R = K_G = 1/2$. The result for α has been multiplied by 1.01; otherwise the curves would be too close to tell apart.

Figure 21.16.2 shows the components α_R and α_G computed for the particular choice $K_R = K_G = 1/2$. The figure also shows $\alpha_R + \alpha_G$, along with α computed using Equation 21.16.6. We find that the agreement between these values is very good, which is compelling evidence that the ansatz is valid and $K_R = K_G = 1/2$.

Note that there is nothing to indicate that the results demonstrated in the example are not generally true. Thus, we come to the following conclusion:

The attenuation constant $\alpha \approx \alpha_R + \alpha_G$ where $\alpha_R \triangleq R'/2Z_0$ and $\alpha_G \triangleq G'Z_0/2$.

Minimizing attenuation

Let us now consider if there are design choices which minimize the attenuation of coaxial cable. Since $\alpha = \alpha_R + \alpha_G$, we may consider α_R and α_G independently. Let us first consider α_G :

$$\begin{aligned} \alpha_G &\triangleq \frac{1}{2} G' Z_0 \\ &\approx \frac{1}{2} \cdot \frac{2\pi\sigma_s}{\ln(b/a)} \cdot \frac{1}{2\pi} \frac{\eta_0}{\sqrt{\epsilon_r}} \ln(b/a) \\ &= \frac{\eta_0}{2} \frac{\sigma_s}{\sqrt{\epsilon_r}} \end{aligned} \quad (21.16.10)$$

It is clear from this result that α_G is minimized by minimizing $\sigma_s/\sqrt{\epsilon_r}$. Interestingly the physical dimensions a and b have no discernible effect on α_G . Now we consider α_R :

$$\begin{aligned}
 \alpha_R &\triangleq \frac{R'}{2Z_0} \\
 &= \frac{1}{2} \frac{(1/2\pi\delta_{ic}\sigma_{ic}) [1/a + C/b]}{(1/2\pi) (\eta_0/\sqrt{\epsilon_r}) \ln(b/a)} \\
 &= \frac{\sqrt{\epsilon_r}}{2\eta_0\delta_{ic}\sigma_{ic}} \cdot \frac{[1/a + C/b]}{\ln(b/a)}
 \end{aligned} \tag{21.16.11}$$

Now making the substitution $\delta_{ic} = \sqrt{2/\omega\mu_0\sigma_{ic}}$ in order to make the dependences on the constitutive parameters explicit, we find:

$$\alpha_R = \frac{1}{2\sqrt{2} \cdot \eta_0} \sqrt{\frac{\omega\mu_0\epsilon_r}{\sigma_{ic}}} \cdot \frac{[1/a + C/b]}{\ln(b/a)}$$

Here we see that α_R is minimized by minimizing ϵ_r/σ_{ic} . It's not surprising to see that we should maximize σ_{ic} . However, it's a little surprising that we should minimize ϵ_r . Furthermore, this is in contrast to α_G , which is minimized by *maximizing* ϵ_r . Clearly there is a tradeoff to be made here. To determine the parameters of this tradeoff, first note that the result depends on frequency: Since α_R dominates over α_G at sufficiently high frequency (as demonstrated in Figure 21.16.2), it seems we should minimize ϵ_r if the intended frequency of operation is sufficiently high; otherwise the optimum value is frequency-dependent. However, σ_s may vary as a function of ϵ_r , so a general conclusion about optimum values of σ_s and ϵ_r is not appropriate.

However, we also see that α_R – unlike α_G – depends on a and b . This implies the existence of a generally-optimum geometry. To find this geometry, we minimize α_R by taking the derivative with respect to a , setting the result equal to zero, and solving for a and/or b . Here we go:

$$\frac{\partial}{\partial a} \alpha_R = \frac{1}{2\sqrt{2} \cdot \eta_0} \sqrt{\frac{\omega\mu_0\epsilon_r}{\sigma_{ic}}} \cdot \frac{\partial}{\partial a} \frac{[1/a + C/b]}{\ln(b/a)} \tag{21.16.12}$$

This derivative is worked out in an addendum at the end of this section. Using the result from the addendum, the right side of Equation 21.16.12 can be written as follows:

$$\frac{1}{2\sqrt{2} \cdot \eta_0} \sqrt{\frac{\omega\mu_0\epsilon_r}{\sigma_{ic}}} \cdot \left[\frac{-1}{a^2 \ln(b/a)} + \frac{1/a + C/b}{a \ln^2(b/a)} \right] \tag{21.16.13}$$

In order for $\partial\alpha_R/\partial a = 0$, the factor in the square brackets above must be equal to zero. After a few steps of algebra, we find:

$$\ln(b/a) = 1 + \frac{C}{b/a}$$

In Case I ($\sigma_{oc} \gg \sigma_{ic}$), $C = 0$ so:

$$b/a = e \cong 2.72 \quad (\text{Case I})$$

In Case II ($\sigma_{oc} = \sigma_{ic}$), $C = 1$. The resulting equation can be solved by plotting the function, or by a few iterations of trial and error; either way one quickly finds

$$b/a \cong 3.59 \quad (\text{Case II})$$

Summarizing, we have found that α is minimized by choosing the ratio of the outer and inner radii to be somewhere between 2.72 and 3.59, with the precise value depending on the relative conductivity of the inner and outer conductors.

Substituting these values of b/a into Equation 21.16.8, we obtain:

$$Z_0 \approx \frac{59.9 \Omega}{\sqrt{\epsilon_r}} \text{ to } \frac{76.6 \Omega}{\sqrt{\epsilon_r}} \tag{21.16.14}$$

as the range of impedances of coaxial cable corresponding to physical designs that minimize attenuation.

Equation 21.16.14 gives the range of characteristic impedances that minimize attenuation for coaxial transmission lines. The precise value within this range depends on the ratio of the conductivity of the outer conductor to that of the inner conductor.

Since $\epsilon_r \geq 1$, the impedance that minimizes attenuation is less for dielectric-filled cables than it is for air-filled cables. For example, let us once again consider the RG-59 from Example 21.16.1. In that case, $\epsilon_r \approx 2.25$ and $C = 0$, indicating $Z_0 \approx 39.9 \Omega$ is optimum for attenuation. The actual characteristic impedance of Z_0 is about 75Ω , so clearly RG-59 is not optimized for attenuation. This is simply because other considerations apply, including power handling capability (addressed in Section 7.4) and the convenience of standard values (addressed in Section 7.5).

Addendum: Derivative of $a^2 \ln(b/a)$

Evaluation of Equation 21.16.12 requires finding the derivative of $a^2 \ln(b/a)$ with respect to a . Using the chain rule, we find:

$$\begin{aligned} \frac{\partial}{\partial a} \left[a^2 \ln \left(\frac{b}{a} \right) \right] &= \left[\frac{\partial}{\partial a} a^2 \right] \ln \left(\frac{b}{a} \right) \\ &\quad + a^2 \left[\frac{\partial}{\partial a} \ln \left(\frac{b}{a} \right) \right] \end{aligned} \quad (21.16.15)$$

Note

$$\frac{\partial}{\partial a} a^2 = 2a$$

and

$$\begin{aligned} \frac{\partial}{\partial a} \ln \left(\frac{b}{a} \right) &= \frac{\partial}{\partial a} [\ln(b) - \ln(a)] \\ &= -\frac{\partial}{\partial a} \ln(a) \\ &= -\frac{1}{a} \end{aligned} \quad (21.16.16)$$

So:

$$\begin{aligned} \frac{\partial}{\partial a} \left[a^2 \ln \left(\frac{b}{a} \right) \right] &= [2a] \ln \left(\frac{b}{a} \right) + a^2 \left[-\frac{1}{a} \right] \\ &= \boxed{2a \ln \left(\frac{b}{a} \right) - a} \end{aligned} \quad (21.16.17)$$

This result is substituted for $a^2 \ln(b/a)$ in Equation 21.16.12 to obtain Equation 21.16.13.

This page titled [21.16: Attenuation in Coaxial Cable](#) is shared under a CC BY-SA license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [7.3: Attenuation in Coaxial Cable](#) by Steven W. Ellingson is licensed CC BY-SA 4.0. Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

21.17: Power Handling Capability of Coaxial Cable

The term “power handling” refers to maximum power that can be safely transferred by a transmission line. This power is limited because when the electric field becomes too large, dielectric breakdown and arcing may occur. This may result in damage to the line and connected devices, and so must be avoided. Let E_{pk} be the maximum safe value of the electric field intensity within the line, and let P_{max} be the power that is being transferred under this condition. This section addresses the following question: How does one design a coaxial cable to maximize P_{max} for a given E_{pk} ?

We begin by finding the electric potential V within the cable. This can be done using Laplace’s equation:

$$\nabla^2 V = 0$$

Using the cylindrical (ρ, ϕ, z) coordinate system with the z axis along the inner conductor, we have $\partial V / \partial \phi = 0$ due to symmetry. Also we set $\partial V / \partial z = 0$ since the result should not depend on z . Thus, we have:

$$\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial V}{\partial \rho} \right) = 0$$

Solving for V , we have

$$V(\rho) = A \ln \rho + B$$

where A and B are arbitrary constants, presumably determined by boundary conditions. Let us assume a voltage V_0 measured from the inner conductor (serving as the “+” terminal) to the outer conductor (serving as the “−” terminal). For this choice, we have:

$$V(a) = V_0 \rightarrow A \ln a + B = V_0 \quad (21.17.1)$$

$$V(b) = 0 \rightarrow A \ln b + B = 0 \quad (21.17.2)$$

Subtracting the second equation from the first and solving for A , we find $A = -V_0 / \ln(b/a)$. Subsequently, B is found to be $V_0 \ln(b) / \ln(b/a)$, and so

$$V(\rho) = \frac{-V_0}{\ln(b/a)} \ln \rho + \frac{V_0 \ln(b)}{\ln(b/a)}$$

The electric field intensity is given by:

$$\mathbf{E} = -\nabla V$$

Again we have $\partial V / \partial \phi = \partial V / \partial z = 0$, so

$$\mathbf{E} = -\hat{\rho} \frac{\partial}{\partial \rho} V \quad (21.17.3)$$

$$= -\hat{\rho} \frac{\partial}{\partial \rho} \left[\frac{-V_0}{\ln(b/a)} \ln \rho + \frac{V_0 \ln(b)}{\ln(b/a)} \right] \quad (21.17.4)$$

$$= +\hat{\rho} \frac{V_0}{\rho \ln(b/a)} \quad (21.17.5)$$

Note that the maximum electric field intensity in the spacer occurs at $\rho = a$; i.e., at the surface of the inner conductor. Therefore:

$$E_{pk} = \frac{V_0}{a \ln(b/a)}$$

The power transferred by the line is maximized when the impedances of the source and load are matched to Z_0 . In this case, the power transferred is $V_0^2 / 2Z_0$. Recall that the characteristic impedance Z_0 is given in the “low-loss” case as

$$Z_0 \approx \frac{1}{2\pi} \frac{\eta_0}{\sqrt{\epsilon_r}} \ln \left(\frac{b}{a} \right) \quad (21.17.6)$$

Therefore, the maximum safe power is

$$P_{max} = \frac{V_0^2}{2Z_0} \quad (21.17.7)$$

$$\approx \frac{E_{pk}^2 a^2 \ln^2(b/a)}{2 \cdot (1/2\pi) (\eta_0/\sqrt{\epsilon_r}) \ln(b/a)} \quad (21.17.8)$$

$$= \frac{\pi E_{pk}^2}{\eta_0/\sqrt{\epsilon_r}} a^2 \ln(b/a) \quad (21.17.9)$$

Now let us consider if there is a value of a which maximizes P_{max} . We do this by seeing if $\partial P_{max}/\partial a = 0$ for some values of a and b . The derivative is worked out in an addendum at the end of this section. Using the result from the addendum, we find:

$$\frac{\partial}{\partial a} P_{max} = \frac{\pi E_{pk}^2}{\eta_0/\sqrt{\epsilon_r}} [2a \ln(b/a) - a] \quad (21.17.10)$$

For the above expression to be zero, it must be true that $2\ln(b/a) - 1 = 0$. Solving for b/a , we obtain:

$$\frac{b}{a} = \sqrt{e} \cong 1.65 \quad (21.17.11)$$

for optimum power handling. In other words, 1.65 is the ratio of the radii of the outer and inner conductors that maximizes the power that can be safely handled by the cable.

Equation 21.17.9 suggests that ϵ_r should be maximized in order to maximize power handling, and you wouldn't be wrong for noting that, however, there are some other factors that may indicate otherwise. For example, a material with higher ϵ_r may also have higher σ_s , which means more current flowing through the spacer and thus more ohmic heating. This problem is so severe that cables that handle high RF power often use air as the spacer, even though it has the *lowest* possible value of ϵ_r . Also worth noting is that σ_{ic} and σ_{oc} do not matter according to the analysis we've just done; however, to the extent that limited conductivity results in significant ohmic heating in the conductors – which we have also not considered – there may be something to consider. Suffice it to say, the actionable finding here concerns the ratio of the radii; the other parameters have not been suitably constrained by this analysis.

Substituting \sqrt{e} for b/a in Equation 21.17.6, we find:

$$Z_0 \approx \frac{30.0 \Omega}{\sqrt{\epsilon_r}}$$

This is the characteristic impedance of coaxial line that optimizes power handling, subject to the caveats identified above. For air-filled cables, we obtain **30 Ω**. Since $\epsilon_r \geq 1$, this optimum impedance is less for dielectric-filled cables than it is for air-filled cables.

Summarizing:

The power handling capability of coaxial transmission line is optimized when the ratio of radii of the outer to inner conductors b/a is about 1.65. For the air-filled cables typically used in high-power applications, this corresponds to a characteristic impedance of about **30 Ω**.

Addendum: Derivative of $(1/a + C/b) / \ln(b/a)$

Evaluation of Equation 21.17.9 requires finding the derivative of $(1/a + C/b) / \ln(b/a)$ with respect to a . Using the chain rule, we find:

$$\begin{aligned} \frac{\partial}{\partial a} \left[\frac{1/a + C/b}{\ln(b/a)} \right] &= \left[\frac{\partial}{\partial a} \left(\frac{1}{a} + \frac{C}{b} \right) \right] \ln^{-1}\left(\frac{b}{a}\right) \\ &\quad + \left(\frac{1}{a} + \frac{C}{b} \right) \left[\frac{\partial}{\partial a} \ln^{-1}\left(\frac{b}{a}\right) \right] \end{aligned} \quad (21.17.12)$$

Note

$$\frac{\partial}{\partial a} \left(\frac{1}{a} + \frac{C}{b} \right) = -\frac{1}{a^2}$$

To handle the quantity in the second set of square brackets, first define $v = \ln u$, where $u = b/a$. Then:

$$\begin{aligned}
 \frac{\partial}{\partial a} v^{-1} &= \left[\frac{\partial}{\partial v} v^{-1} \right] \left[\frac{\partial v}{\partial u} \right] \left[\frac{\partial u}{\partial a} \right] \\
 &= [-v^{-2}] \left[\frac{1}{u} \right] [-ba^{-2}] \\
 &= \left[-\ln^{-2} \left(\frac{b}{a} \right) \right] \left[\frac{a}{b} \right] [-ba^{-2}] \\
 &= \frac{1}{a} \ln^{-2} \left(\frac{b}{a} \right)
 \end{aligned} \tag{21.17.13}$$

So:

$$\begin{aligned}
 \frac{\partial}{\partial a} \left[\frac{1/a + C/b}{\ln(b/a)} \right] &= \left[-\frac{1}{a^2} \right] \ln^{-1} \left(\frac{b}{a} \right) \\
 &\quad + \left(\frac{1}{a} + \frac{C}{b} \right) \left[\frac{1}{a} \ln^{-2} \left(\frac{b}{a} \right) \right]
 \end{aligned} \tag{21.17.14}$$

This result is substituted in Equation 21.17.9 to obtain Equation 21.17.10.

This page titled [21.17: Power Handling Capability of Coaxial Cable](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson](#) (Virginia Tech Libraries' Open Education Initiative).

- [7.4: Power Handling Capability of Coaxial Cable](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source:
<https://doi.org/10.21061/electromagnetics-vol-2>.

21.18: Why 50 Ohms?

The quantity 50Ω appears in a broad range of applications across the field of electrical engineering. In particular, it is a very popular value for the characteristic impedance of transmission line, and is commonly specified as the port impedance for signal sources, amplifiers, filters, antennas, and other RF components. So, what's special about 50Ω ? The short answer is "nothing." In fact, other standard impedances are in common use – prominent among these is 75Ω . It is shown in this section that a broad range of impedances – on the order of 10s of ohms – emerge as useful values based on technical considerations such as minimizing attenuation, maximizing power handling, and compatibility with common types of antennas. Characteristic impedances up to 300Ω and beyond are useful in particular applications. However, it is not practical or efficient to manufacture and sell products for every possible impedance in this range. Instead, engineers have settled on 50Ω as a round number that lies near the middle of this range, and have chosen a few other values to accommodate the smaller number of applications where there may be specific compelling considerations.

So, the question becomes "what makes characteristic impedances in the range of 10s of ohms particularly useful?" One consideration is attenuation in coaxial cable. Coaxial cable is by far the most popular type of transmission line for connecting devices on separate printed circuit boards or in separate enclosures. The attenuation of coaxial cable is addressed in Section 7.3. In that section, it is shown that attenuation is minimized for characteristic impedances in the range $(60 \Omega) / \sqrt{\epsilon_r}$ to $(77 \Omega) / \sqrt{\epsilon_r}$, where ϵ_r is the relative permittivity of the spacer material. So, we find that Z_0 in the range 60Ω to 77Ω is optimum for air-filled cable, but more like 40Ω to 50Ω for cables using a plastic spacer material having typical $\epsilon_r \approx 2.25$. Thus, 50Ω is clearly a reasonable choice if a single standard value is to be established for all such cable.

Coaxial cables are often required to carry high power signals. In such applications, power handling capability is also important, and is addressed in Section 7.4. In that section, we find the power handling capability of coaxial cable is optimized when the ratio of radii of the outer to inner conductors b/a is about 1.65. For the air-filled cables typically used in high-power applications, this corresponds to a characteristic impedance of about 30Ω . This is significantly less than the 60Ω to 77Ω that minimizes attenuation in air-filled cables. So, 50Ω can be viewed as a compromise between minimizing attenuation and maximizing power handling in air-filled coaxial cables.

Although the preceding arguments justify 50Ω as a standard value, one can also see how one might make a case for 75Ω as a secondary standard value, especially for applications where attenuation is the primary consideration.

Values of 50Ω and 75Ω also offer some convenience when connecting RF devices to antennas. For example, 75Ω is very close to the impedance of the commonly-encountered half-wave dipole antenna (about $73 + j42 \Omega$), which may make impedance matching to that antenna easier. Another commonly-encountered antenna is the quarter-wave monopole, which exhibits an impedance of about $36 + j21 \Omega$, which is close to 50Ω . In fact, we see that if we desire a single characteristic impedance that is equally convenient for applications involving either type of antenna, then 50Ω is a reasonable choice.

A third commonly-encountered antenna is the *folded* half-wave dipole. This type of antenna is similar to a half-wave dipole but has better bandwidth, and is commonly used in FM and TV systems and land mobile radio (LMR) base stations. A folded half-wave dipole has an impedance of about 300Ω and is balanced (not single-ended); thus, there is a market for balanced transmission line having $Z_0 = 300 \Omega$. However, it is very easy and inexpensive to implement a balun (a device which converts the dipole output from balanced to unbalanced) while simultaneously stepping down impedance by a factor of 4; i.e., to 75Ω . Thus, we have an additional application for 75Ω coaxial line.

Finally, note that it is quite simple to implement microstrip transmission line having characteristic impedance in the range 30Ω to 75Ω . For example, 50Ω on commonly-used 1.575 mm FR4 requires a width-to-height ratio of about 2, so the trace is about 3 mm wide. This is a very manageable size and easily implemented in printed circuit board designs.

Additional Reading:

- "Dipole antenna" on Wikipedia.
- "Monopole antenna" on Wikipedia.
- "Balun" on Wikipedia.

This page titled [21.18: Why 50 Ohms?](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson](#) ([Virginia Tech Libraries' Open Education Initiative](#)) .

- [7.5: Why 50 Ohms?](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

21.19: Conclusion

21.19: Conclusion is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

22: Generation and Detection of Electromagnetic Waves

- 22.1: Introduction
- 22.2: Production of Electromagnetic Waves - The Antenna
- 22.3: Radiation from a Current Moment
- 22.4: Radiation from an Electrically-Short Dipole
- 22.5: Far-Field Radiation from a Half-Wave Dipole
- 22.6: Equivalent Circuit Model for Transmission; Radiation Efficiency
- 22.7: Equivalent Circuit Model for Reception
- 22.8: Potential Induced in a Dipole
- 22.9: Decibel Scale for Power Ratio
- 22.10: Antenna Radiation Patterns, Directivity, and Gain
- 22.11: Friis Transmission Equation

22: Generation and Detection of Electromagnetic Waves is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

22.1: Introduction

This chapter covers the theory of the generation and detection of electromagnetic waves in more detail than in Part 1. Most of the sections in the chapter are excerpted from the books by Steven W. Ellingson, Electromagnetics, vol. 1 and vol. 2. Before reading this chapter, you should be familiar with the content in Part 1 through the chapter on Electromagnetic Waves.

22.1: Introduction is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Ronald Kumon.

22.2: Production of Electromagnetic Waves - The Antenna

Learning Objectives

By the end of this section, you will be able to:

- Describe the electric and magnetic waves as they move out from a source, such as an AC generator.
- Explain the mathematical relationship between the magnetic field strength and the electrical field strength.
- Calculate the maximum strength of the magnetic field in an electromagnetic wave, given the maximum electric field strength.

We can get a good understanding of **electromagnetic waves** (EM) by considering how they are produced. Whenever a current varies, associated electric and magnetic fields vary, moving out from the source like waves. Perhaps the easiest situation to visualize is a varying current in a long straight wire, produced by an AC generator at its center, as illustrated in Figure 22.2.1.

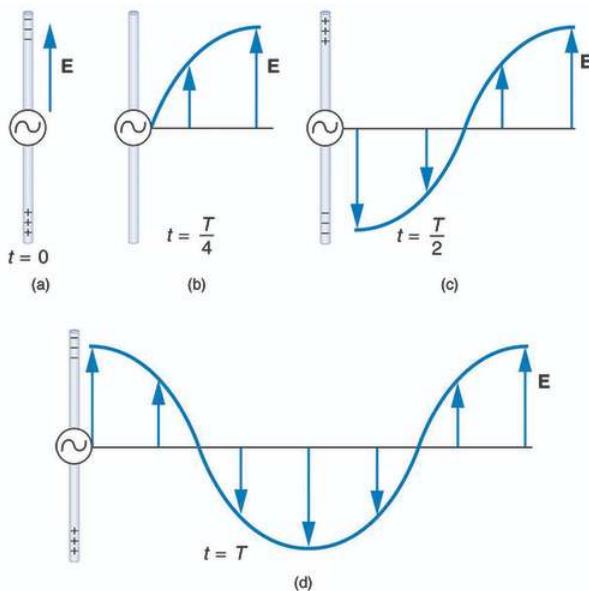


Figure 22.2.1: This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field (\mathbf{E}) propagates away from the antenna at the speed of light, forming part of an electromagnetic wave.

The **electric field (\mathbf{E})** shown surrounding the wire is produced by the charge distribution on the wire. Both the \mathbf{E} and the charge distribution vary as the current changes. The changing field propagates outward at the speed of light.

There is an associated **magnetic field (\mathbf{B})** which propagates outward as well (Figure 22.2.2). The electric and magnetic fields are closely related and propagate as an electromagnetic wave. This is what happens in broadcast antennae such as those in radio and TV stations.

Closer examination of the one complete cycle shown in Figure 22.2.1 reveals the periodic nature of the generator-driven charges oscillating up and down in the antenna and the electric field produced. At time $t = 0$, there is the maximum separation of charge, with negative charges at the top and positive charges at the bottom, producing the maximum magnitude of the electric field (or E -field) in the upward direction. One-fourth of a cycle later, there is no charge separation and the field next to the antenna is zero, while the maximum E -field has moved away at speed c .

As the process continues, the charge separation reverses and the field reaches its maximum downward value, returns to zero, and rises to its maximum upward value at the end of one complete cycle. The outgoing wave has an **amplitude** proportional to the maximum separation of charge. Its **wavelength (λ)** is proportional to the period of the oscillation and, hence, is smaller for short periods or high frequencies. (As usual, wavelength and **frequency (f)** are inversely proportional.)

Electric and Magnetic Waves: Moving Together

Following Ampere's law, current in the antenna produces a magnetic field, as shown in Figure 22.2.2. The relationship between \mathbf{E} and \mathbf{B} is shown at one instant in Figure 2a. As the current varies, the magnetic field varies in magnitude and direction.

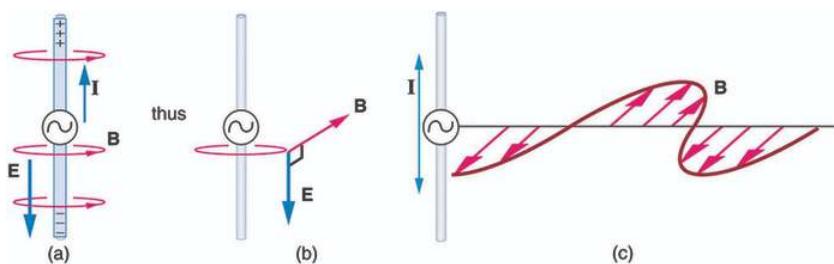


Figure 22.2.2: (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields (E and B) near the wire are perpendicular; they are shown here for one point in space. (c) The magnetic field varies with current and propagates away from the antenna at the speed of light.

The magnetic field lines also propagate away from the antenna at the speed of light, forming the other part of the electromagnetic wave, as seen in Figure 22.2.2b. The magnetic part of the wave has the same period and wavelength as the electric part, since they are both produced by the same movement and separation of charges in the antenna.

The electric and magnetic waves are shown together at one instant in time in Figure 22.2.3. The electric and magnetic fields produced by a long straight wire antenna are exactly in phase. Note that they are perpendicular to one another and to the direction of propagation, making this a **transverse wave**.

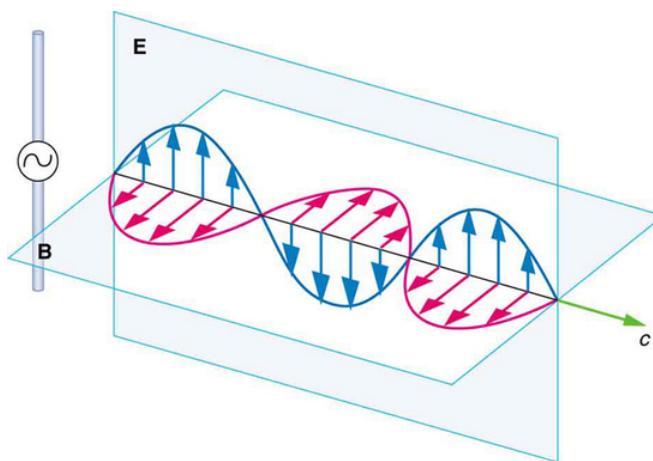


Figure 22.2.3: A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields (E and B) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one direction, but they propagate out in other directions too.

Electromagnetic waves generally propagate out from a source in all directions, sometimes forming a complex radiation pattern. A linear antenna like this one will not radiate parallel to its length, for example. The wave is shown in one direction from the antenna in Figure 22.2.3 to illustrate its basic characteristics.

Instead of the AC generator, the antenna can also be driven by an AC circuit. In fact, charges radiate whenever they are accelerated. But while a current in a circuit needs a complete path, an antenna has a varying charge distribution forming a **standing wave**, driven by the AC. The dimensions of the antenna are critical for determining the frequency of the radiated electromagnetic waves. This is a **resonant** phenomenon and when we tune radios or TV, we vary electrical properties to achieve appropriate resonant conditions in the antenna.

Receiving Electromagnetic Waves

Electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving EM signals works in reverse. And like antennas that produce EM waves, receiver

antennas are specially designed to resonate at particular frequencies.

An incoming electromagnetic wave accelerates electrons in the antenna, setting up a standing wave. If the radio or TV is switched on, electrical components pick up and amplify the signal formed by the accelerating electrons. The signal is then converted to audio and/or video format. Sometimes big receiver dishes are used to focus the signal onto an antenna.

In fact, charges radiate whenever they are accelerated. When designing circuits, we often assume that energy does not quickly escape AC circuits, and mostly this is true. A broadcast antenna is specially designed to enhance the rate of electromagnetic radiation, and shielding is necessary to keep the radiation close to zero. Some familiar phenomena are based on the production of electromagnetic waves by varying currents. Your microwave oven, for example, sends electromagnetic waves, called microwaves, from a concealed antenna that has an oscillating current imposed on it.

Relating E -Field and B -Field Strengths

There is a relationship between the E - and B - field strengths in an electromagnetic wave. This can be understood by again considering the antenna just described. The stronger the E -field created by a separation of charge, the greater the current and, hence, the greater the B -field created.

Since current is directly proportional to voltage (Ohm's law) and voltage is directly proportional to E -field strength, the two should be directly proportional. It can be shown that the magnitudes of the fields do have a constant ratio, equal to the speed of light. That is,

$$\frac{E}{B} = c \quad (22.2.1)$$

is the ratio of E -field strength to B -field strength in any electromagnetic wave. This is true at all times and at all locations in space. A simple and elegant result.

Example 22.2.1: Calculating B -Field Strength in an Electromagnetic Wave

What is the maximum strength of the B -field in an electromagnetic wave that has a maximum E -field strength of 1000V/m ?

Strategy:

To find the B -field strength, we rearrange the Equation 22.2.1 to solve for B , yielding

$$B = \frac{E}{c}. \quad (22.2.2)$$

Solution:

We are given E , and c is the speed of light. Entering these into the expression for B yields

$$B = \frac{1000\text{V/m}}{3.00 \times 10^8 \text{m/s}} = 3.33 \times 10^{-6}\text{T},$$

Where T stands for Tesla, a measure of magnetic field strength.

Discussion:

The B -field strength is less than a tenth of the Earth's admittedly weak magnetic field. This means that a relatively strong electric field of 1000 V/m is accompanied by a relatively weak magnetic field. Note that as this wave spreads out, say with distance from an antenna, its field strengths become progressively weaker.

The result of this example is consistent with the statement made in the module 24.2 that changing electric fields create relatively weak magnetic fields. They can be detected in electromagnetic waves, however, by taking advantage of the phenomenon of resonance, as Hertz did. A system with the same natural frequency as the electromagnetic wave can be made to oscillate. All radio and TV receivers use this principle to pick up and then amplify weak electromagnetic waves, while rejecting all others not at their resonant frequency.

TAKE-HOME EXPERIMENT: ANTENNAS

For your TV or radio at home, identify the antenna, and sketch its shape. If you don't have cable, you might have an outdoor or indoor TV antenna. Estimate its size. If the TV signal is between 60 and 216 MHz for basic channels, then what is the wavelength of those EM waves?

Try tuning the radio and note the small range of frequencies at which a reasonable signal for that station is received. (This is easier with digital readout.) If you have a car with a radio and extendable antenna, note the quality of reception as the length of the antenna is changed.

PHET EXPLORATIONS: RADIO WAVES AND ELECTROMAGNETIC FIELDS

Broadcast radio waves from [KPhET](#). Wiggle the transmitter electron manually or have it oscillate automatically. Display the field as a curve or vectors. The strip chart shows the electron positions at the transmitter and at the receiver.

Summary

- Electromagnetic waves are created by oscillating charges (which radiate whenever accelerated) and have the same frequency as the oscillation.
- Since the electric and magnetic fields in most electromagnetic waves are perpendicular to the direction in which the wave moves, it is ordinarily a transverse wave.
- The strengths of the electric and magnetic parts of the wave are related by

$$\frac{\mathbf{E}}{\mathbf{B}} = c,$$

which implies that the magnetic field \mathbf{B} is very weak relative to the electric field \mathbf{E} .

Glossary

electric field

a vector quantity (\mathbf{E}); the lines of electric force per unit charge, moving radially outward from a positive charge and in toward a negative charge

electric field strength

the magnitude of the electric field, denoted E -field

magnetic field

a vector quantity (\mathbf{B}); can be used to determine the magnetic force on a moving charged particle

magnetic field strength

the magnitude of the magnetic field, denoted B -field

transverse wave

a wave, such as an electromagnetic wave, which oscillates perpendicular to the axis along the line of travel

standing wave

a wave that oscillates in place, with nodes where no motion happens

wavelength

the distance from one peak to the next in a wave

amplitude

the height, or magnitude, of an electromagnetic wave

frequency

the number of complete wave cycles (up-down-up) passing a given point within one second (cycles/second)

resonant

a system that displays enhanced oscillation when subjected to a periodic disturbance of the same frequency as its natural frequency

oscillate

to fluctuate back and forth in a steady beat

This page titled [22.2: Production of Electromagnetic Waves - The Antenna](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [24.2: Production of Electromagnetic Waves](#) by [OpenStax](#) is licensed [CC BY 4.0](#). Original source:
<https://openstax.org/details/books/college-physics>.

22.3: Radiation from a Current Moment

In this section, we begin to address the following problem: Given a distribution of impressed current density $\mathbf{J}(\mathbf{r})$, what is the resulting electric field intensity $\mathbf{E}(\mathbf{r})$? One route to an answer is via Maxwell's equations. Viewing Maxwell's equations as a system of differential equations, a rigorous mathematical solution is possible given the appropriate boundary conditions. The rigorous solution following that approach is relatively complicated, and is presented beginning in Section 9.2 of this book.

If we instead limit scope to a sufficiently simple current distribution, a simple informal derivation is possible. This section presents such a derivation. The advantage of tackling a simple special case first is that it will allow us to quickly assess the nature of the solution, which will turn out to be useful once we do eventually address the more general problem. Furthermore, the results presented in this section will turn out to be sufficient to tackle many commonly-encountered applications.

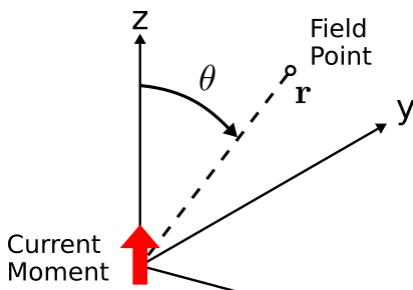


Figure 22.3.1: A $+\hat{\mathbf{z}}$ -directed current moment located at the origin. (CC BY-SA 4.0; C. Wang)

The simple current distribution considered in this section is known as a *current moment*. An example of a current moment is shown in Figure 22.3.1 and in this case is defined as follows:

$$\Delta\mathbf{J}(\mathbf{r}) = \hat{\mathbf{z}} I \Delta l \delta(\mathbf{r}) \quad (22.3.1)$$

where $I \Delta l$ is the scalar component of the current moment, having units of current times length (SI base units of A·m); and $\delta(\mathbf{r})$ is the volumetric sampling function¹ defined as follows:

$$\delta(\mathbf{r}) \triangleq 0 \text{ for } \mathbf{r} \neq 0; \text{ and} \quad (22.3.2)$$

$$\int_{\mathcal{V}} \delta(\mathbf{r}) d\mathbf{v} \triangleq 1 \quad (22.3.3)$$

where \mathcal{V} is any volume which includes the origin ($\mathbf{r} = 0$). It is evident from Equation 22.3.3 that $\delta(\mathbf{r})$ has SI base units of m^{-3} . Subsequently, $\Delta\mathbf{J}(\mathbf{r})$ has SI base units of A/m^2 , confirming that it is a volume current density. However, it is the *simplest possible* form of volume current density, since – as indicated by Equation 22.3.2 – it exists only at the origin and nowhere else.

Although some current distributions approximate the current moment, current distributions encountered in common engineering practice generally do not exist in precisely this form. Nevertheless, the current moment turns out to be generally useful as a “building block” from which practical distributions of current can be constructed, via the principle of superposition. Radiation from current distributions constructed in this manner is calculated simply by summing the radiation from each of the constituent current moments.

Now let us consider the electric field intensity $\Delta\mathbf{E}(\mathbf{r})$ that is created by this current distribution. First, if the current is steady (i.e., “DC”), this problem falls within the domain of magnetostatics; i.e., the outcome is completely described by the magnetic field, and there can be no radiation. Therefore, let us limit our attention to the “AC” case, for which radiation is possible. It will be convenient to employ phasor representation. In phasor representation, the current density is

$$\Delta\tilde{\mathbf{J}}(\mathbf{r}) = \hat{\mathbf{z}} \tilde{I} \Delta l \delta(\mathbf{r}) \quad (22.3.4)$$

where $\tilde{I} \Delta l$ is simply the scalar current moment expressed as a phasor.

Now we are ready to address the question “What is $\Delta\tilde{\mathbf{E}}(\mathbf{r})$ due to $\Delta\tilde{\mathbf{J}}(\mathbf{r})$?” Without doing any math, we know quite a bit about $\Delta\tilde{\mathbf{E}}(\mathbf{r})$. For example:

¹ Since electric fields are proportional to the currents that give rise to them, we expect $\Delta\tilde{\mathbf{E}}(\mathbf{r})$ to be proportional to $|\tilde{I}| \Delta l$.

- If we are sufficiently far from the origin, we expect $\Delta\mathbf{E}(\mathbf{r})$ to be approximately proportional to $1/r$, where $r \equiv |\mathbf{r}|$ is the distance from the source current. This is because point sources give rise to spherical waves, and the power density in a spherical wave would be proportional to $1/r^2$. Since time-average power density is proportional to $|\Delta\tilde{\mathbf{E}}(\mathbf{r})|^2$, $\Delta\tilde{\mathbf{E}}(\mathbf{r})$ must be proportional to $1/r$.
- If we are sufficiently far from the origin, and the loss due to the medium is negligible, then we expect the phase of $\Delta\tilde{\mathbf{E}}(\mathbf{r})$ to change approximately at rate β where β is the phase propagation constant $2\pi/\lambda$. Since we expect spherical phasefronts, $\Delta\tilde{\mathbf{E}}(\mathbf{r})$ should therefore contain the factor $e^{-j\beta r}$.
- Ampere's law indicates that a \hat{z} -directed current at the origin should give rise to a $\hat{\phi}$ -directed magnetic field in the $z = 0$ plane.² At the same time, Poynting's theorem requires the cross product of the electric and magnetic fields to point in the direction of power flow. In the present problem, this direction is away from the source; i.e., $+\hat{r}$. Therefore, $\Delta\tilde{\mathbf{E}}(z = 0)$ points in the $-\hat{z}$ direction. The same principle applies outside of the $z = 0$ plane, so in general we expect $\Delta\tilde{\mathbf{E}}(\mathbf{r})$ to point in the $\hat{\theta}$ direction.
- We expect $\Delta\tilde{\mathbf{E}}(\mathbf{r}) = 0$ along the z axis. Subsequently $|\Delta\tilde{\mathbf{E}}(\hat{\mathbf{r}})|$ must increase from zero at $\theta = 0$ and return to zero at $\theta = \pi$. The symmetry of the problem suggests $|\Delta\tilde{\mathbf{E}}(\hat{\mathbf{r}})|$ is maximum at $\theta = \pi/2$. This magnitude must vary in the simplest possible way, leading us to conclude that $\Delta\tilde{\mathbf{E}}(\hat{\mathbf{r}})$ is proportional to $\sin\theta$. Furthermore, the radial symmetry of the problem means that $\Delta\tilde{\mathbf{E}}(\hat{\mathbf{r}})$ should not depend at all on ϕ .

Putting these ideas together, we conclude that the radiated electric field has the following form:

$$\Delta\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} C (\tilde{I} \Delta l) (\sin\theta) \frac{e^{-j\beta r}}{r}$$

where C is a constant which accounts for all of the constants of proportionality identified in the preceding analysis. Since the units of $\Delta\tilde{\mathbf{E}}(\mathbf{r})$ are V/m, the units of C must be $\Omega\text{-m}$. We have not yet accounted for the wave impedance of the medium η , which has units of Ω , so it would be a good bet based on the units that C is proportional to η . However, here the informal analysis reaches a dead end, so we shall simply state the result from the rigorous solution: $C = j\eta\beta/4\pi$. The units are correct, and we finally obtain:

$$\Delta\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} \frac{j\eta\beta}{4\pi} (\tilde{I} \Delta l) (\sin\theta) \frac{e^{-j\beta r}}{r}$$

Additional evidence that this solution is correct comes from the fact that it satisfies the wave equation $\nabla^2 \Delta\tilde{\mathbf{E}}(\mathbf{r}) + \beta^2 \Delta\tilde{\mathbf{E}}(\mathbf{r}) = 0$.³

Note that the expression we have obtained for the radiated electric field is approximate (hence the “≈”). This is due in part to our presumption of a simple spherical wave, which may only be valid at distances far from the source. But how far? An educated guess would be distances much greater than a wavelength (i.e., $r \gg \lambda$). This will do for now; in another section, we shall show rigorously that this guess is essentially correct.

We conclude this section by noting that the current distribution analyzed in this section is sometimes referred to as a *Hertzian dipole*. A Hertzian dipole is typically defined as a straight infinitesimally-thin filament of current with length which is very small relative to a wavelength, but not precisely zero. This interpretation does not change the solution obtained in this section, thus we may view the current moment and the Hertzian dipole as effectively the same in practical engineering applications.

Additional Reading:

- “Dirac delta function” on Wikipedia.
- “Dipole antenna” (section entitled “Hertzian Dipole”) on Wikipedia.

1. Also a form of the *Dirac delta function*; see “Additional Reading” at the end of this section. ↪

2. This is sometimes described as the “right hand rule” of Ampere’s law. ↪

3. Confirming this is straightforward (simply substitute and evaluate) and is left as an exercise for the student. ↪

This page titled [22.3: Radiation from a Current Moment](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

• **9.1: Radiation from a Current Moment** by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

22.4: Radiation from an Electrically-Short Dipole

The simplest distribution of radiating current that is encountered in common practice is the **electrically-short dipole (ESD)**. This current distribution is shown in Figure 22.4.1.

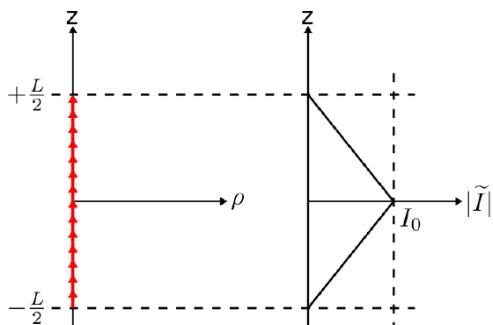


Figure 22.4.1: Current distribution of the electrically-short dipole (ESD). (CC BY-SA 4.0; C. Wang)

The two characteristics that define the ESD are (1) the current is aligned along a straight line, and (2) the length L of the line is much less than one-half of a wavelength; i.e., $L \ll \lambda/2$. The latter characteristic is what we mean by “electrically-short.”¹

The current distribution of an ESD is approximately triangular in magnitude, and approximately constant in phase. How do we know this? First, note that distributions of current cannot change in a complex or rapid way over such distances which are much less than a wavelength. If this is not immediately apparent, recall the behavior of transmission lines: The current standing wave on a transmission line exhibits a period of $\lambda/2$, regardless the source or termination. For the ESD, $L \ll \lambda/2$ and so we expect an even simpler variation. Also, we know that the current at the ends of the dipole must be zero, simply because the dipole ends there. These considerations imply that the current distribution of the ESD is well-approximated as triangular in magnitude.² Expressed mathematically:

$$\tilde{I}(z) \approx I_0 \left(1 - \frac{2}{L} |z|\right)$$

where I_0 (SI base units of A) is a complex-valued constant indicating the maximum current magnitude and phase.

There are two approaches that we might consider in order to find the electric field radiated by an ESD. The first approach is to calculate the magnetic vector potential $\tilde{\mathbf{A}}$ by integration over the current distribution, calculate $\tilde{\mathbf{H}} = (1/\mu)\nabla \times \tilde{\mathbf{A}}$, and finally calculate $\tilde{\mathbf{E}}$ from $\tilde{\mathbf{H}}$ using Ampere’s law. We shall employ a simpler approach, shown in Figure 22.4.2.

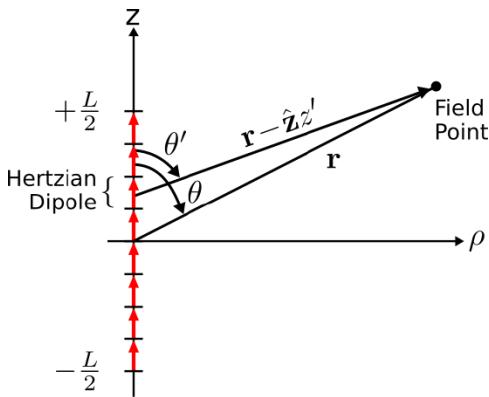


Figure 22.4.2: Current distribution of the electricallyshort dipole (ESD) approximated as a large number of Hertzian dipoles. (CC BY-SA 4.0; C. Wang)

Imagine the ESD as a collection of many shorter segments of current that radiate independently. The total field is then the sum of these short segments. Because these segments are very short relative to the length of the dipole as well as being short relative to a wavelength, we may approximate the current over each segment as approximately constant. In other words, we may interpret each of these segments as being, to a good approximation, a Hertzian dipole.

The advantage of this approach is that we already have a solution for each of the segments. In Section 9.4, it is shown that a $\hat{\mathbf{z}}$ -directed Hertzian dipole at the origin radiates the electric field

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j\eta \frac{\tilde{\mathbf{I}} \cdot \beta \Delta l}{4\pi} (\sin \theta) \frac{e^{-j\beta r}}{r}$$

where $\tilde{\mathbf{I}}$ and Δl may be interpreted as the current and length of the dipole, respectively. In this expression, η is the wave impedance of medium in which the dipole radiates (e.g., $\approx 377 \Omega$ for free space), and we have presumed lossless media such that the attenuation constant $\alpha \approx 0$ and the phase propagation constant $\beta = 2\pi/\lambda$. This expression also assumes field points far from the dipole; specifically, distances r that are much greater than λ . Repurposing this expression for the present problem, the segment at the origin radiates the electric field:

$$\tilde{\mathbf{E}}(\mathbf{r}; z' = 0) \approx \hat{\theta} j\eta \frac{I_0 \cdot \beta \Delta l}{4\pi} (\sin \theta) \frac{e^{-j\beta r}}{r}$$

where the notation $z' = 0$ indicates the Hertzian dipole is located at the origin. Letting the length Δl of this segment shrink to differential length dz' , we may describe the contribution of this segment to the field radiated by the ESD as follows:

$$d\tilde{\mathbf{E}}(\mathbf{r}; z' = 0) \approx \hat{\theta} j\eta \frac{I_0 \cdot \beta dz'}{4\pi} (\sin \theta) \frac{e^{-j\beta r}}{r}$$

Using this approach, the electric field radiated by *any* segment can be written:

$$d\tilde{\mathbf{E}}(\mathbf{r}; z') \approx \hat{\theta}' j\eta \beta \frac{\tilde{\mathbf{I}}(z')}{4\pi} (\sin \theta') \frac{e^{-j\beta |\mathbf{r} - \hat{\mathbf{z}}z'|}}{|\mathbf{r} - \hat{\mathbf{z}}z'|} dz'$$

Note that θ is replaced by θ' since the ray $\mathbf{r} - \hat{\mathbf{z}}z'$ forms a different angle (i.e., θ') with respect to $\hat{\mathbf{z}}$. Similarly, $\hat{\theta}$ is replaced by $\hat{\theta}'$, since it also varies with z' . The electric field radiated by the ESD is obtained by integration over these contributions:

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \int_{-L/2}^{+L/2} d\tilde{\mathbf{E}}(\hat{\mathbf{r}}; z')$$

yielding:

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx j \frac{\eta \beta}{4\pi} \int_{-L/2}^{+L/2} \hat{\theta}' \tilde{\mathbf{I}}(z') (\sin \theta') \frac{e^{-j\beta |\mathbf{r} - \hat{\mathbf{z}}z'|}}{|\mathbf{r} - \hat{\mathbf{z}}z'|} dz'$$

Given some of the assumptions we have already made, this expression can be further simplified. For example, note that $\theta' \approx \theta$ since $L \ll r$. For the same reason, $\hat{\theta}' \approx \hat{\theta}$. Since these variables are approximately constant over the length of the dipole, we may move them outside the integral, yielding:

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta \beta}{4\pi} (\sin \theta) \int_{-L/2}^{+L/2} \tilde{\mathbf{I}}(z') \frac{e^{-j\beta |\mathbf{r} - \hat{\mathbf{z}}z'|}}{|\mathbf{r} - \hat{\mathbf{z}}z'|} dz' \quad (22.4.1)$$

It is also possible to simplify the expression $|\mathbf{r} - \hat{\mathbf{z}}z'|$. Consider Figure 22.4.3.

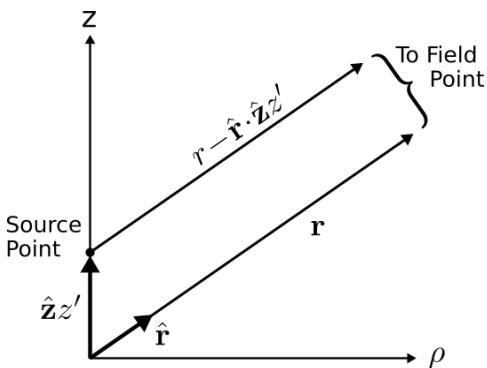


Figure 22.4.3: Parallel ray approximation for an ESD. (CC BY-SA 4.0; C. Wang)

Since we have already assumed that $r \gg L$ (i.e., the distance to field points is much greater than the length of the dipole), the vector \mathbf{r} is approximately parallel to the vector $\mathbf{r} - \hat{\mathbf{z}}z'$. Subsequently, it must be true that

$$|\mathbf{r} - \hat{\mathbf{z}}z'| \approx r - \hat{\mathbf{r}} \cdot \hat{\mathbf{z}}z' \quad (22.4.2)$$

Note that the magnitude of $r - \hat{\mathbf{r}} \cdot \hat{\mathbf{z}}z'$ must be approximately equal to r , since $r \gg L$. So, insofar as $|\mathbf{r} - \hat{\mathbf{z}}z'|$ determines the *magnitude* of $\tilde{\mathbf{E}}(\mathbf{r})$, we may use the approximation:

$$|\mathbf{r} - \hat{\mathbf{z}}z'| \approx r \quad (\text{magnitude})$$

Insofar as $|\mathbf{r} - \hat{\mathbf{z}}z'|$ determines *phase*, we have to be a bit more careful. The part of the integrand of Equation 22.4.1 that exhibits varying phase is $e^{-j\beta|\mathbf{r} - \hat{\mathbf{z}}z'|}$. Using Equation 22.4.2, we find

$$e^{-j\beta|\mathbf{r} - \hat{\mathbf{z}}z'|} \approx e^{-j\beta r} e^{+j\beta \hat{\mathbf{r}} \cdot \hat{\mathbf{z}}z'}$$

The worst case in terms of phase variation within the integral is for field points along the z axis. For these points, $\hat{\mathbf{r}} \cdot \hat{\mathbf{z}} = \pm 1$ and subsequently $|\mathbf{r} - \hat{\mathbf{z}}z'|$ varies from $z - L/2$ to $z + L/2$ where z is the location of the field point. However, since $L \ll \lambda$ (i.e., because the dipole is electrically short), this difference in lengths is much less than $\lambda/2$. Therefore, the phase $\beta\hat{\mathbf{r}} \cdot \hat{\mathbf{z}}z'$ varies by much less than π radians, and subsequently $e^{-j\beta\hat{\mathbf{r}} \cdot \hat{\mathbf{z}}z'} \approx 1$. We conclude that under these conditions,

$$e^{-j\beta|\mathbf{r} - \hat{\mathbf{z}}z'|} \approx e^{-j\beta r} \quad (\text{phase})$$

Applying these simplifications for magnitude and phase to Equation 22.4.1, we obtain:

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta\beta}{4\pi} (\sin \theta) \frac{e^{-j\beta r}}{r} \int_{-L/2}^{+L/2} \tilde{I}(z') dz'$$

The integral in this equation is very easy to evaluate; in fact, from inspection (Figure 22.4.1), we determine it is equal to $I_0 L / 2$. Finally, we obtain:

$$\boxed{\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \eta \frac{I_0 \cdot \beta L}{8\pi} (\sin \theta) \frac{e^{-j\beta r}}{r}} \quad (22.4.3)$$

Summarizing:

The electric field intensity radiated by an ESD located at the origin and aligned along the z axis is given by Equation 22.4.3. This expression is valid for $r \gg \lambda$.

It is worth noting that the variation in magnitude, phase, and polarization of the ESD with field point location is identical to that of a single Hertzian dipole having current moment $\hat{\mathbf{z}}I_0L/2$ (Section 9.4). However, the magnitude of the field radiated by the ESD is exactly one-half that of the Hertzian dipole. Why one-half? Simply because the integral over the triangular current distribution assumed for the ESD is one-half the integral over the uniform current distribution that defines the Hertzian dipole. This similarly sometimes causes confusion between Hertzian dipoles and ESDs. Remember that ESDs are physically realizable, whereas Hertzian dipoles are not.

It is common to eliminate the factor of β in the magnitude using the relationship $\beta = 2\pi/\lambda$, yielding:

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta I_0}{4} \frac{L}{\lambda} (\sin \theta) \frac{e^{-j\beta r}}{r}$$

At field points $r \gg \lambda$, the wave appears to be locally planar. Therefore, we are justified using the plane wave relationship $\tilde{\mathbf{H}} = \frac{1}{\eta} \hat{\mathbf{r}} \times \tilde{\mathbf{E}}$ to calculate $\tilde{\mathbf{H}}$. The result is:

$$\tilde{\mathbf{H}}(\mathbf{r}) \approx \hat{\phi} j \frac{I_0}{4} \frac{L}{\lambda} (\sin \theta) \frac{e^{-j\beta r}}{r} \quad (22.4.4)$$

Finally, let us consider the spatial characteristics of the radiated field. Figures 22.4.4 and 22.4.5 show the result in a plane of constant ϕ . Figures 22.4.6 and 22.4.7 show the result in the $z = 0$ plane. Note that the orientations of the electric and magnetic field vectors indicate a Poynting vector $\tilde{\mathbf{E}} \times \tilde{\mathbf{H}}$ that is always directed radially outward from the location of the dipole. This confirms that power flow is always directed radially outward from the dipole. Due to the symmetry of the problem, Figures 22.4.4 – 22.4.7 provide a complete characterization of the relative magnitudes and orientations of the radiated fields.

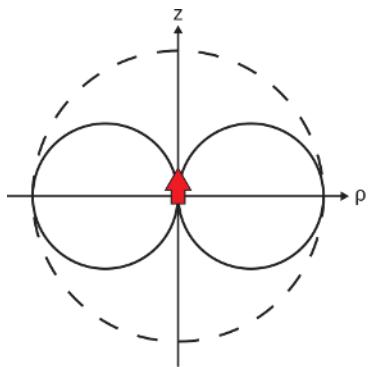
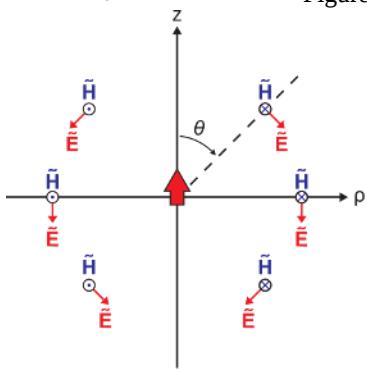
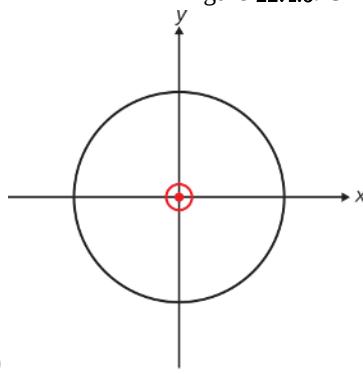


Figure 22.4.4: Magnitude of the radiated field in any plane of constant ϕ . (CC BY-SA 4.0; S.



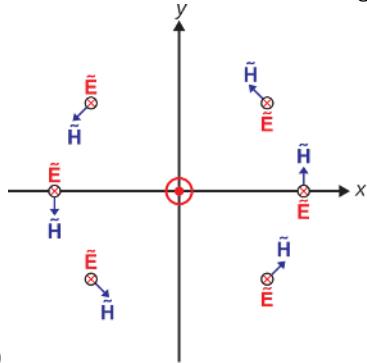
Lally)

Figure 22.4.5: Orientation of the electric and magnetic fields in any plane of constant ϕ . (



CC BY-SA 4.0; S. Lally)

Figure 22.4.6: Magnitude of the radiated field in any plane of constant



z. (CC BY-SA 4.0; S. Lally)

$z = 0$ plane. (CC BY-SA 4.0; S. Lally)

Figure 22.4.7: Orientation of the electric and magnetic fields in the

1. A potential source of confusion is that the *Hertzian dipole* is also a “dipole” which is “electrically-short.” The distinction is that the current comprising a Hertzian dipole is *constant* over its length. This condition is rarely and only approximately seen in practice, whereas the triangular magnitude distribution is a relatively good approximation to a broad class of commonly-encountered electrically-short wire antennas. Thus, the term “electrically-short dipole,” as used in this book, refers to the triangular distribution unless noted otherwise. ↪
2. A more rigorous analysis leading to the same conclusion is possible, but is beyond the scope of this book. ↪

This page titled [22.4: Radiation from an Electrically-Short Dipole](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [9.5: Radiation from an Electrically-Short Dipole](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

22.5: Far-Field Radiation from a Half-Wave Dipole

A simple and important current distribution is that of the thin half-wave dipole (HWD), shown in Figure 22.5.1.

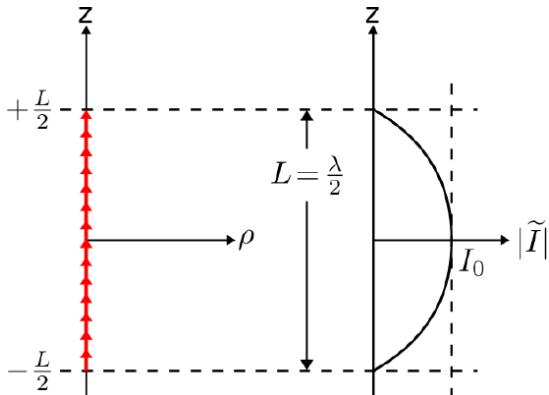


Figure 22.5.1: Current distribution of the half-wave dipole (HWD). (CC BY-SA 4.0; C. Wang)

This is the distribution expected on a thin straight wire having length $L = \lambda/2$, where λ is wavelength. This distribution is described mathematically as follows:

$$\tilde{I}(z) \approx I_0 \cos\left(\pi \frac{z}{L}\right) \quad \text{for } |z| \leq \frac{L}{2} \quad (22.5.1)$$

where I_0 (SI base units of A) is a complex-valued constant indicating the maximum magnitude of the current and its phase. Note that the current is zero at the ends of the dipole; i.e., $\tilde{I}(z) = 0$ for $|z| = L/2$. Note also that this “cosine pulse” distribution is very similar to the triangular distribution of the ESD, and is reminiscent of the sinusoidal variation of current in a standing wave.

Since $L = \lambda/2$ for the HWD, Equation 22.5.1 may equivalently be written:

$$\tilde{I}(z) \approx I_0 \cos\left(2\pi \frac{z}{\lambda}\right) \quad (22.5.2)$$

The electromagnetic field radiated by this distribution of current may be calculated using the method described in Section 9.6, in particular:

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta}{2} \frac{e^{-j\beta r}}{r} (\sin \theta) \cdot \left[\frac{1}{\lambda} \int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z' \cos \theta} dz' \right] \quad (22.5.3)$$

which is valid for field points \mathbf{r} far from the dipole; i.e., for $r \gg L$ and $r \gg \lambda$. For the HWD, the quantity in square brackets is

$$\frac{I_0}{\lambda} \int_{-\lambda/4}^{+\lambda/4} \cos\left(2\pi \frac{z'}{\lambda}\right) e^{+j\beta z' \cos \theta} dz'$$

The evaluation of this integral is straightforward, but tedious. The integral reduces to

$$\frac{I_0}{\pi} \frac{\cos[(\pi/2) \cos \theta]}{\sin^2 \theta}$$

Substitution into Equation 22.5.3 yields

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta I_0}{2\pi} \frac{\cos[(\pi/2) \cos \theta]}{\sin \theta} \frac{e^{-j\beta r}}{r}$$

The magnetic field may be determined from this result using Ampere's law. However, a simpler method is to use the fact that the electric field, magnetic field, and direction of propagation $\hat{\mathbf{r}}$ are mutually perpendicular and related by:

$$\tilde{\mathbf{H}} = \frac{1}{\eta} \hat{\mathbf{r}} \times \tilde{\mathbf{E}}$$

This relationship indicates that the magnetic field will be $+\hat{\phi}$ -directed.

The magnitude and polarization of the radiated field is similar to that of the electrically-short dipole (ESD; Section 9.5). A comparison of the magnitudes in any radial plane containing the z -axis is shown in Figure 22.5.2.

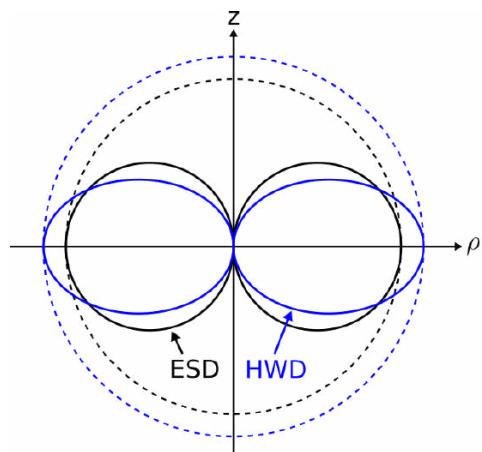


Figure 22.5.2: Comparison of the magnitude of the radiated field of the HWD to that of an electrically-short dipole also oriented along the z -axis. This result is for any radial plane that includes the z -axis. ([CC BY-SA 4.0](#); C. Wang)

For either current distribution, the maximum magnitude of the fields occurs in the $z = 0$ plane. For a given terminal current I_0 , the maximum magnitude is greater for the HWD than for the ESD. Both current distributions yield zero magnitude along the axis of the dipole. The polarization characteristics of the fields of both current distributions are identical.

Additional Reading:

- “Dipole antenna” (section entitled “Half-wave dipole”) on Wikipedia.

This page titled [22.5: Far-Field Radiation from a Half-Wave Dipole](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson](#) (Virginia Tech Libraries' Open Education Initiative).

- [9.7: Far-Field Radiation from a Half-Wave Dipole](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

22.6: Equivalent Circuit Model for Transmission; Radiation Efficiency

A radio transmitter consists of a source which generates the electrical signal intended for transmission, and an antenna which converts this signal into a propagating electromagnetic wave. Since the transmitter is an electrical system, it is useful to be able to model the antenna as an equivalent circuit. From a circuit analysis point of view, it should be possible to describe the antenna as a passive one-port circuit that presents an impedance to the source. Thus, we have the following question: What is the equivalent circuit for an antenna which is transmitting?

We begin by emphasizing that the antenna is passive. That is, the antenna does not add power. Invoking the principle of conservation of power, there are only three possible things that *can* happen to power that is delivered to the antenna by the transmitter:¹

- Power can be converted to a propagating electromagnetic wave. (The desired outcome.)
- Power can be dissipated within the antenna.
- Energy can be stored by the antenna, analogous to the storage of energy in a capacitor or inductor.

We also note that these outcomes can occur in any combination. Taking this into account, we model the antenna using the equivalent circuit shown in Figure 22.6.1.

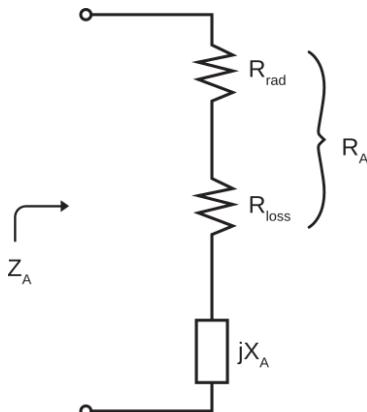


Figure 22.6.1: Equivalent circuit for an antenna which is transmitting. (CC BY-SA 4.0; S. Lally)

Since the antenna is passive, it is reasonable to describe it as an impedance Z_A which is (by definition) the ratio of voltage \tilde{V}_A to current \tilde{I}_A at the terminals; i.e.,

$$Z_A \triangleq \frac{\tilde{V}_A}{\tilde{I}_A}$$

In the phasor domain, Z_A is a complex-valued quantity and therefore has, in general, a real-valued component and an imaginary component. We may identify those components using the power conservation argument made previously: Since the real-valued component must represent power transfer and the imaginary component must represent energy storage, we infer:

$$Z_A \triangleq R_A + jX_A$$

where R_A represents power transferred to the antenna, and X_A represents energy stored by the antenna. Note that the energy stored by the antenna is being addressed in precisely the same manner that we address energy storage in a capacitor or an inductor; in all cases, as reactance. Further, we note that R_A consists of components R_{rad} and R_{loss} as follows:

$$Z_A = R_{rad} + R_{loss} + jX_A$$

where R_{rad} represents power transferred to the antenna and subsequently radiated, and R_{loss} represents power transferred to the antenna and subsequently dissipated.

To confirm that this model works as expected, consider what happens when a voltage is applied across the antenna terminals. The current \tilde{I}_A flows and the time-average power P_A transferred to the antenna is

$$P_A = \frac{1}{2} \operatorname{Re} \left\{ \tilde{V}_A \tilde{I}_A^* \right\}$$

where we have assumed peak (as opposed to root mean squared) units for voltage and current. Since $\tilde{V}_A = Z_A \tilde{I}_A$, we have:

$$P_A = \frac{1}{2} \operatorname{Re} \left\{ (R_{rad} + R_{loss} + jX_A) \tilde{I}_A \tilde{I}_A^* \right\}$$

which reduces to:

$$P_A = \frac{1}{2} |\tilde{I}_A|^2 R_{rad} + \frac{1}{2} |\tilde{I}_A|^2 R_{loss} \quad (22.6.1)$$

As expected, the power transferred to the antenna is the sum of

$$P_{rad} \triangleq \frac{1}{2} |\tilde{I}_A|^2 R_{rad} \quad (22.6.2)$$

representing power transferred to the radiating electromagnetic field, and

$$P_{loss} \triangleq \frac{1}{2} |\tilde{I}_A|^2 R_{loss} \quad (22.6.3)$$

representing power dissipated within the antenna.

The reactance X_A will play a role in determining \tilde{I}_A given \tilde{V}_A (and vice versa), but does not by itself account for disposition of power. Again, this is exactly analogous to the role played by inductors and capacitors in a circuit.

The utility of this equivalent circuit formalism is that it allows us to treat the antenna in the same manner as any other component, and thereby facilitates analysis using conventional electric circuit theory and transmission line theory. For example: Given Z_A , we know how to specify the output impedance Z_S of the transmitter so as to minimize reflection from the antenna: We would choose $Z_S = Z_A$, since in this case the voltage reflection coefficient would be

$$\Gamma = \frac{Z_A - Z_S}{Z_A + Z_S} = 0$$

Alternatively, we might specify Z_S so as to maximize power transfer to the antenna: We would choose $Z_S = Z_A^*$; i.e., conjugate matching.

In order to take full advantage of this formalism, we require values for R_{rad} , R_{loss} , and X_A . These quantities are considered below.

Radiation resistance

(R_{rad}) is referred to as *radiation resistance*. Equation 22.6.2 tells us that

$$R_{rad} = 2P_{rad} |\tilde{I}_A|^{-2} \quad (22.6.4)$$

This equation suggests the following procedure: We apply current \tilde{I}_A to the antenna terminals, and then determine the total power P_{rad} radiated from the antenna in response. For an example of this procedure, see Section 10.2 (“Total Power Radiated by an Electrically-Short Dipole”). Given \tilde{I}_A and P_{rad} , one may then use Equation 22.6.4 to determine R_{rad} .

Loss resistance

Loss resistance represents the dissipation of power within the antenna, which is usually attributable to loss intrinsic to materials comprising or surrounding the antenna. In many cases, antennas are made from good conductors – metals, in particular – so that R_{loss} is very low compared to R_{rad} . For such antennas, loss is often so low compared to R_{rad} that R_{loss} may be neglected. In the case of the electrically-short dipole, R_{loss} is typically very small but R_{rad} is also very small, so both must be considered. In many other cases, antennas contain materials with substantially greater loss than metal. For example, a microstrip patch antenna implemented on a printed circuit board typically has non-negligible R_{loss} because the dielectric material comprising the antenna exhibits significant loss.

Antenna reactance

The reactance term jX_A accounts for energy stored by the antenna. This may be due to reflections internal to the antenna, or due to energy associated with non-propagating electric and magnetic fields surrounding the antenna. The presence of significant reactance (i.e., $|X_A|$ comparable to or greater than $|R_A|$) complicates efforts to establish the desired impedance match to the source. For an example, see Section 10.4 (“Reactance of the Electrically-Short Dipole”).

Radiation efficiency

When R_{loss} is non-negligible, it is useful to characterize antennas in terms of their *radiation efficiency* e_{rad} , defined as the fraction of power which is radiated compared to the total power delivered to the antenna; i.e.,

$$e_{rad} \triangleq \frac{P_{rad}}{P_A}$$

Using Equations 22.6.1–22.6.3, we see that this efficiency can be expressed as follows:

$$e_{rad} = \frac{R_{rad}}{R_{rad} + R_{loss}} \quad (22.6.5)$$

Once again, the equivalent circuit formalism proves useful.

✓ Example 22.6.1: Impedance of an antenna

The total power radiated by an antenna is 60 mW when 20 mA (rms) is applied to the antenna terminals. The radiation efficiency of the antenna is known to be 70%. It is observed that voltage and current are in-phase at the antenna terminals. Determine (a) the radiation resistance, (b) the loss resistance, and (c) the impedance of the antenna.

Solution

From the problem statement, $P_{rad} = 60 \text{ mW}$, $|\tilde{I}_A| = 20 \text{ mA}$ (rms), and $e_{rad} = 0.7$. Also, the fact that voltage and current are in-phase at the antenna terminals indicates that $X_A = 0$. From Equation 22.6.4, the radiation resistance is

$$R_{rad} \approx \frac{2 \cdot (60 \text{ mW})}{|\sqrt{2} \cdot 20 \text{ mA}|^2} = 150 \Omega$$

Solving Equation 22.6.5 for the loss resistance, we find:

$$R_{loss} = \frac{1 - e_{rad}}{e_{rad}} R_{rad} \cong 64.3 \Omega$$

Since $Z_A = R_{rad} + R_{loss} + jX_A$, we find $Z_A \cong 214.3 + j0 \Omega$. This will be the ratio of voltage to current at the antenna terminals regardless of the source current.

-
1. Note that “delivered” power means power accepted by the antenna. We are not yet considering power reflected from the antenna due to impedance mismatch. ↵

This page titled 22.6: Equivalent Circuit Model for Transmission; Radiation Efficiency is shared under a CC BY-SA license and was authored, remixed, and/or curated by Steven W. Ellingson (Virginia Tech Libraries' Open Education Initiative).

- 10.5: Equivalent Circuit Model for Transmission; Radiation Efficiency by Steven W. Ellingson is licensed CC BY-SA 4.0. Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

22.7: Equivalent Circuit Model for Reception

In this section, we begin to address antennas as devices that convert incident electromagnetic waves into potentials and currents in a circuit. It is convenient to represent this process in the form of a Thévenin equivalent circuit. The particular circuit addressed in this section is shown in Figure 22.7.1.

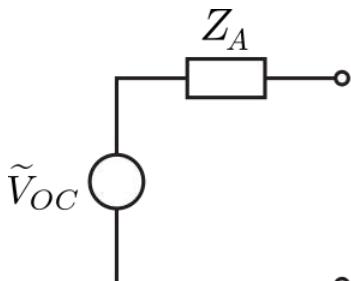


Figure 22.7.1: Thévenin equivalent circuit for an antenna in the presence of an incident electromagnetic wave. (CC BY-SA 4.0 (modified); S. Lally)

The circuit consists of a voltage source \tilde{V}_{OC} and a series impedance Z_A . The source potential \tilde{V}_{OC} is the potential at the terminals of the antenna when there is no termination; i.e., when the antenna is open-circuited. The series impedance Z_A is the output impedance of the circuit, and so determines the magnitude and phase of the current at the terminals once a load is connected. Given \tilde{V}_{OC} and the current through the equivalent circuit, it is possible to determine the power delivered to the load. Thus, this model is quite useful, but only if we are able to determine \tilde{V}_{OC} and Z_A . This section provides an informal derivation of these quantities that is sufficient to productively address the subsequent important topics of effective aperture and impedance matching of receive antennas.¹

Vector effective length

With no derivation required, we can deduce the following about \tilde{V}_{OC} :

- \tilde{V}_{OC} must depend on the incident electric field intensity $\tilde{\mathbf{E}}^i$. Presumably the relationship is linear, so \tilde{V}_{OC} is proportional to the magnitude of $\tilde{\mathbf{E}}^i$.
- Since $\tilde{\mathbf{E}}^i$ is a vector whereas \tilde{V}_{OC} is a scalar, there must be some vector \mathbf{l}_e for which

$$\tilde{V}_{OC} = \tilde{\mathbf{E}}^i \cdot \mathbf{l}_e \quad (22.7.1)$$

- Since $\tilde{\mathbf{E}}^i$ has SI base units of V/m and \tilde{V}_{OC} has SI base units of V, \mathbf{l}_e must have SI base units of m; i.e., length.
- We expect that \tilde{V}_{OC} increases as the size of the antenna increases, so the magnitude of \mathbf{l}_e likely increases with the size of the antenna.
- The direction of \mathbf{l}_e must be related to the orientation of the incident electric field relative to that of the antenna, since this is clearly important yet we have not already accounted for this.

It may seem at this point that \mathbf{l}_e is unambiguously determined, and we need merely to derive its value. However, this is not the case. There are in fact multiple unique definitions of \mathbf{l}_e that will reduce the vector $\tilde{\mathbf{E}}^i$ to the observed scalar \tilde{V}_{OC} via Equation 22.7.1. In this section, we shall employ the most commonly-used definition, in which \mathbf{l}_e is referred to as *vector effective length*. Following this definition, the scalar part l_e of $\mathbf{l}_e = \hat{\mathbf{l}} l_e$ is commonly referred to as any of the following: *effective length* (the term used in this book), *effective height*, or *antenna factor*.

In this section, we shall merely define vector effective length, and defer a formal derivation to Section 10.11. In this definition, we arbitrarily set $\hat{\mathbf{l}}$, the real-valued unit vector indicating the direction of \mathbf{l}_e , equal to the direction in which the electric field transmitted from this antenna would be polarized in the far field. For example, consider a $\hat{\mathbf{z}}$ -oriented electrically-short dipole (ESD) located at the origin. The electric field transmitted from this antenna would have only a $\hat{\theta}$ component, and no $\hat{\phi}$ component (and certainly no $\hat{\mathbf{r}}$ component). Thus, $\hat{\mathbf{l}} = \hat{\theta}$ in this case.

Applying this definition, $\tilde{\mathbf{E}}^i \cdot \hat{\mathbf{l}}$ yields the scalar component of $\tilde{\mathbf{E}}^i$ that is co-polarized with electric field radiated by the antenna when transmitting. Now l_e is uniquely defined to be the factor that converts this component into \tilde{V}_{OC} . Summarizing:

The *vector effective length* $\mathbf{l}_e = \hat{\mathbf{l}}_e$ is defined as follows: $\hat{\mathbf{l}}$ is the real-valued unit vector corresponding to the polarization of the electric field that would be transmitted from the antenna in the far field. Subsequently, the *effective length* \mathbf{l}_e is

$$\mathbf{l}_e \triangleq \frac{\tilde{V}_{OC}}{\tilde{\mathbf{E}}^i \cdot \hat{\mathbf{l}}}$$

where \tilde{V}_{OC} is the open-circuit potential induced at the antenna terminals in response to the incident electric field intensity $\tilde{\mathbf{E}}^i$.

While this definition yields an unambiguous value for \mathbf{l}_e , it is not yet clear what that value is. For most antennas, effective length is quite difficult to determine directly, and one must instead determine effective length indirectly from the transmit characteristics via reciprocity. This approach is relatively easy (although still quite a bit of effort) for thin dipoles, and is presented in Section 10.11.

To provide an example of how effective length works right away, consider the $\hat{\mathbf{z}}$ -oriented ESD described earlier in this section. Let the length of this ESD be L . Let $\tilde{\mathbf{E}}^i$ be a $\hat{\theta}$ -polarized plane wave arriving at the ESD. The ESD is open-circuited, so the potential induced in its terminals is \tilde{V}_{OC} . One observes the following:

- When $\tilde{\mathbf{E}}^i$ arrives from anywhere in the $\theta = \pi/2$ plane (i.e., broadside to the ESD), $\tilde{\mathbf{E}}^i$ points in the $-\hat{\mathbf{z}}$ direction, and we find that $\mathbf{l}_e \approx L/2$. It should not be surprising that \mathbf{l}_e is proportional to L ; this expectation was noted earlier in this section.
- When $\tilde{\mathbf{E}}^i$ arrives from the directions $\theta = 0$ or $\theta = \pi$ – i.e., along the axis of the ESD – $\tilde{\mathbf{E}}^i$ is perpendicular to the axis of the ESD. In this case, we find that \mathbf{l}_e equals zero.

Taken together, these findings suggest that \mathbf{l}_e should contain a factor of $\sin \theta$. We conclude that the vector effective length for a $\hat{\mathbf{z}}$ -directed ESD of length L is

$$\mathbf{l}_e \approx \hat{\theta} \frac{L}{2} \sin \theta \quad (\text{ESD}) \quad (22.7.2)$$

✓ Example 22.7.1: Potential induced in an ESD

A thin straight dipole of length 10 cm is located at the origin and aligned with the \mathbf{z} -axis. A plane wave is incident on the dipole from the direction ($\theta = \pi/4, \phi = \pi/2$). The frequency of the wave is 30 MHz. The magnitude of the incident electric field is 10 μ V/m (rms). What is the magnitude of the induced open-circuit potential when the electric field is (a) $\hat{\theta}$ -polarized and (b) $\hat{\phi}$ -polarized?

Solution

The wavelength in this example is $c/f \cong 10$ m, so this dipole is electrically-short. Using Equation 22.7.2:

$$\begin{aligned} \mathbf{l}_e &\approx \hat{\theta} \frac{10 \text{ cm}}{2} \sin \frac{\pi}{4} \\ &\approx \hat{\theta} (3.54 \text{ cm}) \end{aligned}$$

Thus, the effective length $\mathbf{l}_e = 3.54$ cm. When the electric field is $\hat{\theta}$ -polarized, the magnitude of the induced open-circuit voltage is

$$\begin{aligned} |\tilde{V}_{OC}| &= |\tilde{\mathbf{E}}^i \cdot \mathbf{l}_e| \\ &\approx (10 \mu\text{V}/\text{m}) \hat{\theta} \cdot \hat{\theta} (3.54 \text{ cm}) \\ &\approx \underline{354 \text{ nV rms}} \quad (\text{a}) \end{aligned}$$

When the electric field is $\hat{\phi}$ -polarized:

$$\begin{aligned} |\tilde{V}_{OC}| &\approx (10 \mu\text{V}/\text{m}) \hat{\phi} \cdot \hat{\theta} (3.54 \text{ cm}) \\ &\approx \underline{0} \quad (\text{b}) \end{aligned}$$

This is because the polarization of the incident electric field is orthogonal to that of the ESD. In fact, the answer to part (b) is zero for *any* angle of incidence (θ, ϕ) .

Output impedance

The output impedance Z_A is somewhat more difficult to determine without a formal derivation, which is presented in Section 10.12. For the purposes of this section, it suffices to jump directly to the result:

The output impedance Z_A of the equivalent circuit for an antenna in the receive case is equal to the input impedance of the same antenna in the *transmit* case.

This remarkable fact is a consequence of the reciprocity property of antenna systems, and greatly simplifies the analysis of receive antennas.

Now a demonstration of how the antenna equivalent circuit can be used to determine the power delivered by an antenna to an attached electrical circuit:

✓ Example 22.7.2: Power captured by an ESD

Continuing with part (a) of Example 22.7.1: If this antenna is terminated into a conjugate-matched load, then what is the power delivered to that load? Assume the antenna is lossless.

Solution

First, we determine the impedance Z_A of the equivalent circuit of the antenna. This is equal to the input impedance of the antenna in transmission. Let R_A and X_A be the real and imaginary parts of this impedance; i.e., $Z_A = R_A + jX_A$. Further, R_A is the sum of the radiation resistance R_{rad} and the loss resistance. The loss resistance is zero because the antenna is lossless. Since this is an ESD:

$$R_{rad} \approx 20\pi^2 \left(\frac{L}{\lambda}\right)^2$$

Therefore, $R_A = R_{rad} \approx 4.93 \text{ m}\Omega$. We do not need to calculate X_A , as will become apparent in the next step.

A conjugate-matched load has impedance Z_A^* , so the potential \tilde{V}_L across the load is

$$\tilde{V}_L = \tilde{V}_{OC} \frac{Z_A^*}{Z_A + Z_A^*} = \tilde{V}_{OC} \frac{Z_A^*}{2R_A}$$

The current \tilde{I}_L through the load is

$$\tilde{I}_L = \frac{\tilde{V}_{OC}}{Z_A + Z_A^*} = \frac{\tilde{V}_{OC}}{2R_A}$$

Taking \tilde{V}_{OC} as an RMS quantity, the power P_L delivered to the load is

$$P_L = \operatorname{Re} \{ V_L I_L^* \} = \frac{|\tilde{V}_{OC}|^2}{4R_A}$$

In part (a) of Example 22.7.1, $|\tilde{V}_{OC}|$ is found to be $\approx 354 \text{ nV rms}$, so $P_L \approx 6.33 \text{ pW}$.

1. Formal derivations of these quantities are provided in subsequent sections. The starting point is the section on reciprocity. ↩

This page titled [22.7: Equivalent Circuit Model for Reception](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by Steven W. Ellingson ([Virginia Tech Libraries' Open Education Initiative](#)) .

- [10.9: Equivalent Circuit Model for Reception](#) by Steven W. Ellingson is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

22.8: Potential Induced in a Dipole

An electromagnetic wave incident on an antenna will induce a potential at the terminals of the antenna. In this section, we shall derive this potential. To simplify the derivation, we shall consider the special case of a straight thin dipole of arbitrary length that is illuminated by a plane wave. However, certain aspects of the derivation will apply to antennas generally. In particular, the concepts of *effective length* (also known as *effective height*) and *vector effective length* emerge naturally from this derivation, so this section also serves as a stepping stone in the development of an equivalent circuit model for a receiving antenna. The derivation relies on the transmit properties of dipoles as well as the principle of reciprocity, so familiarity with those topics is recommended before reading this section.

The scenario of interest is shown in Figure 22.8.1. Here a thin \hat{z} -aligned straight dipole is located at the origin. The total length of the dipole is L . The arms of the dipole are perfectly-conducting. The terminals consist of a small gap of length Δl between the arms. The incident plane wave is described in terms of its electric field intensity $\tilde{\mathbf{E}}^i$. The question is: What is \tilde{V}_{OC} , the potential at the terminals when the terminals are open-circuited?

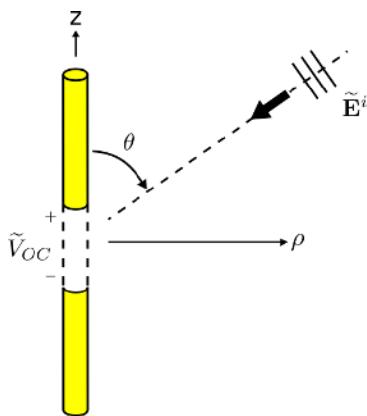


Figure 22.8.1: A potential is induced at the terminals of a thin straight dipole in response to an incident plane wave. (CC BY-SA 4.0; C. Wang)

There are multiple approaches to solve this problem. A direct attack is to invoke the principle that potential is equal to the integral of the electric field intensity over a path. In this case, the path begins at the “-” terminal and ends at the “+” terminal, crossing the gap that defines the antenna terminals. Thus:

$$\tilde{V}_{OC} = - \int_{gap} \tilde{\mathbf{E}}_{gap} \cdot d\mathbf{l} \quad (22.8.1)$$

where $\tilde{\mathbf{E}}_{gap}$ is the electric field in the gap. The problem with this approach is that the value of $\tilde{\mathbf{E}}_{gap}$ is not readily available. It is not simply $\tilde{\mathbf{E}}^i$, because the antenna structure (in particular, the electromagnetic boundary conditions) modify the electric field in the vicinity of the antenna.¹

Fortunately, we can bypass this obstacle using the principle of reciprocity. In a reciprocity-based strategy, we establish a relationship between two scenarios that take place within the same electromagnetic system. The first scenario is shown in Figure 22.8.2.

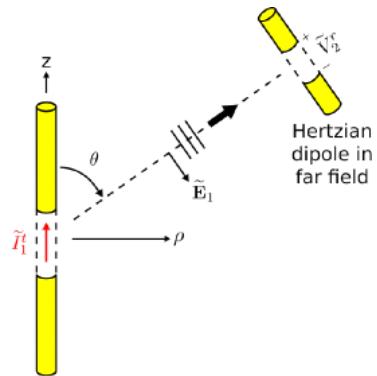


Figure 22.8.2: The dipole of interest driven by current \tilde{I}_1^t radiates electric field $\tilde{\mathbf{E}}_1$, resulting in open-circuit potential \tilde{V}_2^r at the terminals of a Hertzian dipole in the far field. (CC BY-SA 4.0; C. Wang)

In this scenario, we have two dipoles. The first dipole is precisely the dipole of interest (Figure 22.8.1), except that a current \tilde{I}_1^t is applied to the antenna terminals. This gives rise to a current distribution $\tilde{I}(z)$ (SI base units of A) along the dipole, and subsequently the dipole radiates the electric field (Section 9.6):

$$\begin{aligned}\tilde{\mathbf{E}}_1(\mathbf{r}) \approx & \hat{\theta} j \frac{\eta}{2} \frac{e^{-j\beta r}}{r} (\sin \theta) \\ & \cdot \left[\frac{1}{\lambda} \int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z' \cos \theta} dz' \right]\end{aligned}\quad (22.8.2)$$

The second antenna is a $\hat{\theta}$ -aligned Hertzian dipole in the far field, which receives $\tilde{\mathbf{E}}_1$. (For a refresher on the properties of Hertzian dipoles, see Section 9.4. A key point is that Hertzian dipoles are vanishingly small.) Specifically, we measure (conceptually, at least) the open-circuit potential \tilde{V}_2^r at the terminals of the Hertzian dipole. We select a Hertzian dipole for this purpose because – in contrast to essentially all other antennas – it is simple to determine the open circuit potential. As explained earlier:

$$\tilde{V}_2^r = - \int_{gap} \tilde{\mathbf{E}}_{gap} \cdot d\mathbf{l}$$

For the Hertzian dipole, $\tilde{\mathbf{E}}_{gap}$ is simply the incident electric field, since there is negligible structure (in particular, a negligible amount of material) present to modify the electric field. Thus, we have simply:

$$\tilde{V}_2^r = - \int_{gap} \tilde{\mathbf{E}}_1 \cdot d\mathbf{l} \quad (22.8.3)$$

Since the Hertzian dipole is very short and very far away from the transmitting dipole, $\tilde{\mathbf{E}}_1$ is essentially constant over the gap. Also recall that we required the Hertzian dipole to be aligned with $\tilde{\mathbf{E}}_1$. Choosing to integrate in a straight line across the gap, Equation 22.8.3 reduces to:

$$\tilde{V}_2^r = - \tilde{\mathbf{E}}_1(\mathbf{r}_2) \cdot \hat{\theta} \Delta l \quad (22.8.4)$$

where Δl is the length of the gap and \mathbf{r}_2 is the location of the Hertzian dipole. Substituting the expression for $\tilde{\mathbf{E}}_1$ from Equation 22.8.2, we obtain:

$$\begin{aligned}\tilde{V}_2^r \approx & - j \frac{\eta}{2} \frac{e^{-j\beta r_2}}{r_2} (\sin \theta) \\ & \cdot \left[\frac{1}{\lambda} \int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z' \cos \theta} dz' \right] \Delta l\end{aligned}\quad (22.8.5)$$

where $r_2 = |\mathbf{r}_2|$.

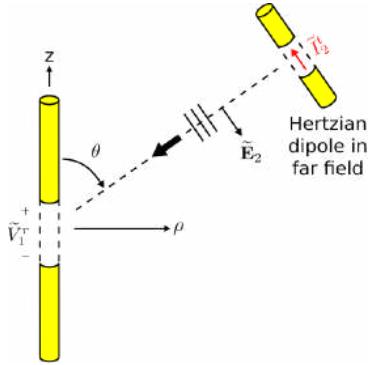


Figure 22.8.3: The Hertzian dipole driven by current \tilde{I}_2^t radiates electric field $\tilde{\mathbf{E}}_2$, resulting in open-circuit potential \tilde{V}_1^r at the terminals of the dipole of interest. (CC BY-SA 4.0; C. Wang)

The second scenario is shown in Figure 22.8.3. This scenario is identical to the first scenario, with the exception that the Hertzian dipole transmits and the dipole of interest receives. The field radiated by the Hertzian dipole in response to applied current \tilde{I}_2^t , evaluated at the origin, is (Section 9.4):

$$\tilde{\mathbf{E}}_2(\mathbf{r} = 0) \approx \hat{\theta} j \eta \frac{\tilde{I}_2^t \cdot \beta \Delta l}{4\pi} (1) \frac{e^{-j\beta r_2}}{r_2} \quad (22.8.6)$$

The “ $\sin \theta$ ” factor in the general expression is equal to 1 in this case, since, as shown in Figure 22.8.3, the origin is located broadside (i.e., at $\pi/2$ rad) relative to the axis of the Hertzian dipole. Also note that because the Hertzian dipole is presumed to be in the far field, \mathbf{E}_2 may be interpreted as a plane wave in the region of the receiving dipole of interest.

Now we ask: What is the induced potential \tilde{V}_1^r in the dipole of interest? Once again, Equation 22.8.1 is not much help, because the electric field in the gap is not known. However, we do know that \tilde{V}_1^r should be proportional to $\tilde{\mathbf{E}}_2(\mathbf{r} = 0)$, since this is presumed to be a linear system. Based on this much information alone, there must be some vector $\mathbf{l}_e = \hat{\mathbf{l}}l_e$ for which

$$\tilde{V}_1^r = \tilde{\mathbf{E}}_2(\mathbf{r} = 0) \cdot \mathbf{l}_e \quad (22.8.7)$$

This does not uniquely define either the unit vector $\hat{\mathbf{l}}$ nor the scalar part l_e , since a change in the definition of the former can be compensated by a change in the definition of the latter and vice-versa. So at this point we invoke the standard definition of \mathbf{l}_e as the *vector effective length*, introduced in Section 10.9. Thus, $\hat{\mathbf{l}}$ is defined to be the direction in which an electric field *transmitted* from the antenna would be polarized. In the present example, $\hat{\mathbf{l}} = -\hat{\theta}$, where the minus sign reflects the fact that positive terminal potential results in terminal current which flows in the $-\hat{\mathbf{z}}$ direction. Thus, Equation 22.8.7 becomes:

$$\tilde{V}_1^r = -\tilde{\mathbf{E}}_2(\mathbf{r} = 0) \cdot \hat{\theta} l_e$$

We may go a bit further and substitute the expression for $\tilde{\mathbf{E}}_2(\mathbf{r} = 0)$ from Equation 22.8.6:

$$\tilde{V}_1^r \approx -j\eta \frac{\tilde{I}_2^t \cdot \beta \Delta l}{4\pi} \frac{e^{-j\beta r_2}}{r_2} l_e \quad (22.8.8)$$

Now we invoke reciprocity. As a two-port linear time-invariant system, it must be true that:

$$\tilde{I}_1^t \tilde{V}_1^r = \tilde{I}_2^t \tilde{V}_2^r$$

Thus:

$$\tilde{V}_1^r = \frac{\tilde{I}_2^t}{\tilde{I}_1^t} \tilde{V}_2^r$$

Substituting the expression for \tilde{V}_2^r from Equation 22.8.5:

$$\begin{aligned} \tilde{V}_1^r &\approx -\frac{\tilde{I}_2^t}{\tilde{I}_1^t} \cdot j \frac{\eta}{2} \frac{e^{-j\beta r_2}}{r_2} (\sin \theta) \\ &\quad \cdot \left[\frac{1}{\lambda} \int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z' \cos \theta} dz' \right] \Delta l \end{aligned}$$

Thus, reciprocity has provided a second expression for \tilde{V}_1^r . We may solve for \mathbf{l}_e by setting this expression equal to the expression from Equation 22.8.8, yielding

$$l_e \approx \frac{2\pi}{\beta \tilde{I}_1^t} \left[\frac{1}{\lambda} \int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z' \cos \theta} dz' \right] \sin \theta$$

Noting that $\beta = 2\pi/\lambda$, this simplifies to:

$$l_e \approx \left[\frac{1}{\tilde{I}_1^t} \int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z' \cos \theta} dz' \right] \sin \theta \quad (22.8.9)$$

Thus, you can calculate \mathbf{l}_e using the following procedure:

1. Apply a current \tilde{I}_1^t to the dipole of interest.
2. Determine the resulting current distribution $\tilde{I}(z)$ along the length of the dipole. (Note that precisely this is done for the electrically-short dipole in Section 9.5 and for the half-wave dipole in Section 9.7.)
3. Integrate $\tilde{I}(z)$ over the length of the dipole as indicated in Equation 22.8.9. Then divide (“normalize”) by \tilde{I}_1^t (which is simply $\tilde{I}(0)$). Note that the result is independent of the excitation \tilde{I}_1^t , as expected since this is a linear system.
4. Multiply by $\sin \theta$.

We have now determined that the open-circuit terminal potential \tilde{V}_{OC} in response to an incident electric field $\tilde{\mathbf{E}}^i$ is

$$\tilde{V}_{OC} = \tilde{\mathbf{E}}^i \cdot \mathbf{l}_e \quad (22.8.10)$$

where $\mathbf{l}_e = \hat{\mathbf{l}} l_e$ is the vector effective length defined previously.

This result is remarkable. In plain English, we have found that:

The potential induced in a dipole is the co-polarized component of the incident electric field times a normalized integral of the *transmit* current distribution over the length of the dipole, times sine of the angle between the dipole axis and the direction of incidence.

In other words, the reciprocity property of linear systems allows this property of a receiving antenna to be determined relatively easily if the transmit characteristics of the antenna are known.

✓ Example 22.8.1: Effective length of a thin electrically-short dipole (ESD)

As explained in Section 9.5, the current distribution on a thin ESD is

$$\tilde{I}(z) \approx I_0 \left(1 - \frac{2}{L} |z| \right)$$

where L is the length of the dipole and I_0 is the terminal current. Applying Equation 22.8.9, we find:

$$l_e \approx \left[\frac{1}{I_0} \int_{-L/2}^{+L/2} I_0 \left(1 - \frac{2}{L} |z'| \right) e^{+j\beta z' \cos \theta} dz' \right] \sin \theta$$

Recall $\beta = 2\pi/\lambda$, so $\beta z' = 2\pi(z'/\lambda)$. Since this is an *electrically-short* dipole, $z' \ll \lambda$ over the entire integral, and subsequently we may assume $e^{+j\beta z' \cos \theta} \approx 1$ over the entire integral. Thus:

$$l_e \approx \left[\int_{-L/2}^{+L/2} \left(1 - \frac{2}{L} |z'| \right) dz' \right] \sin \theta$$

The integral is easily solved using standard methods, or simply recognize that the “area under the curve” in this case is simply one-half “base” (L) times “height” (1). Either way, we find

$$l_e \approx \frac{L}{2} \sin \theta$$

Example 10.9.1 (Section 10.9) demonstrates how Equation 22.8.10 with the vector effective length determined in the preceding example is used to obtain the induced potential.

1. Also, if this were true, then the antenna itself would not matter; only the relative spacing and orientation of the antenna terminals would matter! ↵
-

This page titled [22.8: Potential Induced in a Dipole](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [10.11: Potential Induced in a Dipole](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source:
<https://doi.org/10.21061/electromagnetics-vol-2>.

22.9: Decibel Scale for Power Ratio

In many disciplines within electrical engineering, it is common to evaluate the ratios of powers and power densities that differ by many orders of magnitude. These ratios could be expressed in scientific notation, but it is more common to use the logarithmic *decibel* (dB) scale in such applications.

In the conventional (linear) scale, the ratio of power P_1 to power P_0 is simply

$$G = \frac{P_1}{P_0} \quad (\text{linear units})$$

Here, " G " might be interpreted as "power gain." Note that $G < 1$ if $P_1 < P_0$ and $G > 1$ if $P_1 > P_0$. In the decibel scale, the ratio of power P_1 to power P_0 is

$$G \triangleq 10 \log_{10} \frac{P_1}{P_0} \quad (\text{dB}) \quad (22.9.1)$$

where "dB" denotes a unitless quantity which is expressed in the decibel scale. Note that $G < 0$ dB (i.e., is "negative in dB") if $P_1 < P_0 > 0$ and $G > 0$ dB if $P_1 > P_0$.

The power gain P_1/P_0 in dB is given by Equation 22.9.1.

Alternatively, one might choose to interpret a power ratio as a loss L with $L \triangleq 1/G$ in linear units, which is $L = -G$ when expressed in dB. Most often, but not always, engineers interpret a power ratio as "gain" if the output power is expected to be greater than input power (e.g., as expected for an amplifier) and as "loss" if output power is expected to be less than input power (e.g., as expected for a lossy transmission line).

Power loss L is the reciprocal of power gain G . Therefore, $L = -G$ when these quantities are expressed in dB.

✓ Example 22.9.1: Power loss from a long cable

A 2 W signal is injected into a long cable. The power arriving at the other end of the cable is 10 μW . What is the power loss in dB?

Solution

In linear units:

$$G = \frac{10 \mu\text{W}}{2 \text{ W}} = 5 \times 10^{-6} \quad (\text{linear units})$$

In dB:

Misplaced &

The decibel scale is used in precisely the same way to relate ratios of spatial power densities for waves. For example, the loss incurred when the spatial power density is reduced from S_0 (SI base units of W/m^2) to S_1 is

$$L = 10 \log_{10} \frac{S_0}{S_1} \quad (\text{dB})$$

This works because the common units of m^{-2} in the numerator and denominator cancel, leaving a power ratio.

A common point of confusion is the proper use of the decibel scale to represent voltage or current ratios. To avoid confusion, simply refer to the definition expressed in Equation 22.9.1. For example, let's say $P_1 = V_1^2/R_1$ where V_1 is potential and R_1 is the impedance across which V_1 is defined. Similarly, let us define $P_0 = V_0^2/R_0$ where V_0 is potential and R_0 is the impedance across which V_0 is defined. Applying Equation 22.9.1:

$$\begin{aligned} G &\triangleq 10 \log_{10} \frac{P_1}{P_0} \quad (\text{dB}) \\ &= 10 \log_{10} \frac{V_1^2/R_1}{V_0^2/R_0} \quad (\text{dB}) \end{aligned} \tag{22.9.2}$$

Now, if $R_1 = R_0$, then

$$\begin{aligned} G &= 10 \log_{10} \frac{V_1^2}{V_0^2} \quad (\text{dB}) \\ &= 10 \log_{10} \left(\frac{V_1}{V_0} \right)^2 \quad (\text{dB}) \\ &= 20 \log_{10} \frac{V_1}{V_0} \quad (\text{dB}) \end{aligned} \tag{22.9.3}$$

However, note that this is *not* true if $R_1 \neq R_0$.

A power ratio in dB is equal to $20 \log_{10}$ of the voltage ratio only if the associated impedances are equal.

Adding to the potential for confusion on this point is the concept of *voltage gain* G_v :

$$G_v \triangleq 20 \log_{10} \frac{V_1}{V_0} \quad (\text{dB})$$

which applies regardless of the associated impedances. Note that $G_v = G$ only if the associated impedances are equal, and that these ratios are different otherwise. Be careful!

The decibel scale simplifies common calculations. Here's an example. Let's say a signal having power P_0 is injected into a transmission line having loss L . Then the output power $P_1 = P_0/L$ in linear units. However, in dB, we find:

$$\begin{aligned} 10 \log_{10} P_1 &= 10 \log_{10} \frac{P_0}{L} \\ &= 10 \log_{10} P_0 - 10 \log_{10} L \end{aligned}$$

Division has been transformed into subtraction; i.e.,

$$P_1 = P_0 - L \quad (\text{dB}) \tag{22.9.4}$$

This form facilitates easier calculation and visualization, and so is typically preferred.

Finally, note that the units of P_1 and P_0 in Equation 22.9.4 are not dB *per se*, but rather dB with respect to the original power units. For example, if P_1 is in mW, then taking $10 \log_{10}$ of this quantity results in a quantity having units of dB relative to 1 mW. A power expressed in dB relative to 1 mW is said to have units of "dBm." For example, "0 dBm" means 0 dB relative to 1 mW, which is simply 1 mW. Similarly +10 dBm is 10 mW, -10 dBm is 0.1 mW, and so on.

This page titled [22.9: Decibel Scale for Power Ratio](#) is shared under a CC BY-SA license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [3.8: Decibel Scale for Power Ratio](#) by [Steven W. Ellingson](#) is licensed CC BY-SA 4.0. Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

22.10: Antenna Radiation Patterns, Directivity, and Gain

A transmitting antenna does not radiate power uniformly in all directions. Inevitably more power is radiated in some directions than others. *Directivity* quantifies this behavior. In this section, we introduce the concept of directivity and the related concepts of *maximum directivity* and *antenna gain*.

Consider an antenna located at the origin. The power radiated in a single direction (θ, ϕ) is formally zero. This is because a single direction corresponds to a solid angle of zero, which intercepts an area of zero at any given distance from the antenna. Since the power flowing through any surface having zero area is zero, the power flowing in a single direction is formally zero. Clearly we need a different metric of power in order to develop a sensible description of the spatial distribution of power flow.

The appropriate metric is *spatial power density*; that is, power per unit area, having SI base units of W/m^2 . Therefore, directivity is defined in terms of spatial power density in a particular direction, as opposed to power in a particular direction. Specifically, directivity in the direction (θ, ϕ) is:

$$D(\theta, \phi) \triangleq \frac{S(\mathbf{r})}{S_{ave}(\mathbf{r})} \quad (22.10.1)$$

In this expression, $S(\mathbf{r})$ is the power density at $(\mathbf{r}, \theta, \phi)$; i.e., at a distance \mathbf{r} in the direction (θ, ϕ) . $S_{ave}(\mathbf{r})$ is the *average* power density at that distance; that is, $S(\mathbf{r})$ averaged over all possible directions at distance \mathbf{r} . Since directivity is a ratio of power densities, it is unitless. Summarizing:

Directivity is ratio of power density in a specified direction to the power density averaged over all directions at the same distance from the antenna.

Despite Equation 22.10.1, directivity does not depend on the distance from the antenna. To be specific, directivity is the same at every distance \mathbf{r} . Even though the numerator and denominator of Equation 22.10.1 both vary with \mathbf{r} , one finds that the distance dependence always cancels because power density and average power density are both proportional to \mathbf{r}^{-2} . This is a key point: Directivity is a convenient way to characterize an antenna because it does not change with distance from the antenna.

In general, directivity is a function of direction. However, one is often not concerned about all directions, but rather only the directivity in the direction in which it is maximum. In fact it is quite common to use the term “directivity” informally to refer to the maximum directivity of an antenna. This is usually what is meant when the directivity is indicated to be a single number; in any event, the intended meaning of the term is usually clear from context.

✓ Example 22.10.1: Directivity of the electrically-short dipole

An electrically-short dipole (ESD) consists of a straight wire having length $L \ll \lambda/2$. What is the directivity of the ESD?

Solution

The field radiated by an ESD is derived in Section 9.5. In that section, we find that the electric field intensity in the far field of a $\hat{\mathbf{z}}$ -oriented ESD located at the origin is:

$$\tilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j\eta \frac{I_0 \cdot \beta L}{8\pi} (\sin \theta) \frac{e^{-j\beta r}}{r} \quad (22.10.2)$$

where I_0 represents the magnitude and phase of the current applied to the terminals, η is the wave impedance of the medium, and $\beta = 2\pi/\lambda$. In Section 10.2, we find that the power density of this field is:

$$S(\mathbf{r}) \approx \eta \frac{|I_0|^2 (\beta L)^2}{128\pi^2} (\sin \theta)^2 \frac{1}{r^2} \quad (22.10.3)$$

and we subsequently find that the total power radiated is:

$$P_{rad} \approx \eta \frac{|I_0|^2 (\beta L)^2}{48\pi} \quad (22.10.4)$$

The average power density S_{ave} is simply the total power divided by the area of a sphere centered on the ESD. Let us place this sphere at distance r , with $r \gg L$ and $r \gg \lambda$ as required for the validity of Equations 22.10.2 and 22.10.3. Then:

$$S_{ave} = \frac{P_{rad}}{4\pi r^2} \approx \eta \frac{|I_0|^2 (\beta L)^2}{192\pi^2 r^2}$$

Finally the directivity is determined by applying the definition:

$$\begin{aligned} D(\theta, \phi) &\triangleq \frac{S(\mathbf{r})}{S_{ave}(r)} \\ &\approx 1.5(\sin \theta)^2 \end{aligned} \quad (22.10.5)$$

The maximum directivity occurs in the $\theta = \pi/2$ plane. Therefore, the maximum directivity is , meaning the maximum power density is 1.5 times greater than the power density averaged over all directions.

Since directivity is a unitless ratio, it is common to express it in decibels. For example, the maximum directivity of the ESD in the preceding example is $10 \log_{10} 1.5 \cong 1.76$ dB. (Note “ $10 \log_{10}$ ” here since directivity is the ratio of power-like quantities.)

Gain

The gain $G(\theta, \phi)$ of an antenna is its directivity modified to account for loss within the antenna. Specifically:

$$G(\theta, \phi) \triangleq \frac{S(\mathbf{r}) \text{ for actual antenna}}{S_{ave}(r) \text{ for identical but lossless antenna}}$$

In this equation, the numerator is the actual power density radiated by the antenna, which is less than the nominal power density due to losses within the antenna. The denominator is the average power density for an antenna which is identical, but lossless. Since the actual antenna radiates less power than an identical but lossless version of the same antenna, gain in any particular direction is always less than directivity in that direction. Therefore, an equivalent definition of antenna gain is

$$G(\theta, \phi) \triangleq e_{rad} D(\theta, \phi)$$

where e_{rad} is the radiation efficiency of the antenna (Section 10.5).

Gain is directivity times radiation efficiency; that is, directivity modified to account for loss within the antenna.

The receive case

To conclude this section, we make one additional point about directivity, which applies equally to gain. The preceding discussion has presumed an antenna which is radiating; i.e., transmitting. Directivity can also be defined for the receive case, in which it quantifies the effectiveness of the antenna in converting power in an incident wave to power in a load attached to the antenna. Receive directivity is formally introduced in Section 10.13 (“Effective Aperture”). When receive directivity is defined as specified in Section 10.13, it is equal to transmit directivity as defined in this section. Thus, it is commonly said that the directivity of an antenna is the same for receive and transmit.

Additional Reading:

- “[Directivity](#)” on Wikipedia.

This page titled [22.10: Antenna Radiation Patterns, Directivity, and Gain](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [10.7: Directivity and Gain](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

22.11: Friis Transmission Equation

A common task in radio systems applications is to determine the power delivered to a receiver due to a distant transmitter. The scenario is shown in Figure 22.11.1:

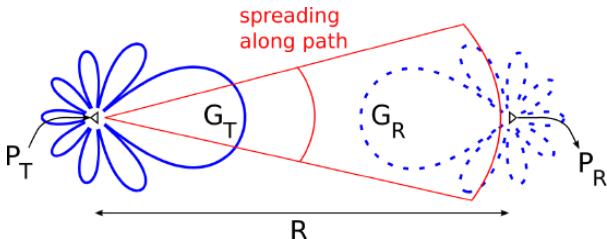


Figure 22.11.1: Copy and Paste Caption here

A transmitter delivers power P_T to an antenna which has gain G_T in the direction of the receiver. The receiver's antenna has gain G_R . As always, antenna gain is equal to directivity times radiation efficiency, so G_T and G_R account for losses internal to the antenna, but not losses due to impedance mismatch.

A simple expression for P_R can be derived as follows. First, let us assume “free space conditions”; that is, let us assume that the intervening terrain exhibits negligible absorption, reflection, or other scattering of the transmitted signal. In this case, the spatial power density at range R from the transmitter which radiates this power through a *lossless and isotropic* antenna would be:

$$\frac{P_T}{4\pi R^2}$$

that is, total transmitted power divided by the area of a sphere of radius R through which all the power must flow. The *actual* power density S^i is this amount times the gain of the transmit antenna, i.e.:

$$S^i = \frac{P_T}{4\pi R^2} G_T$$

The maximum received power is the incident co-polarized power density times the effective aperture A_e of the receive antenna:

$$\begin{aligned} P_{R,max} &= A_e S^i_{co} \\ &= A_e \frac{P_T}{4\pi R^2} G_T \end{aligned} \tag{22.11.1}$$

This assumes that the receive antenna is co-polarized with the incident electric field, and that the receiver is conjugate-matched to the antenna. The effective aperture can also be expressed in terms of the gain G_R of the receive antenna:

$$A_e = \frac{\lambda^2}{4\pi} G_R$$

Thus, Equation 22.11.1 may be written in the following form:

$$P_{R,max} = P_T G_T \left(\frac{\lambda}{4\pi R} \right)^2 G_R \tag{22.11.2}$$

This is the *Friis transmission equation*. Summarizing:

The *Friis transmission equation* (Equation 22.11.2) gives the power delivered to a conjugate-matched receiver in response to a distant transmitter, assuming co-polarized antennas and free space conditions.

The factor $(\lambda/4\pi R)^2$ appearing in the Friis transmission equation is referred to as *free space path gain*. More often this is expressed as the reciprocal quantity:

$$L_p \triangleq \left(\frac{\lambda}{4\pi R} \right)^{-2}$$

which is known as *free space path loss*. Thus, Equation 22.11.2 may be expressed as follows:

$$P_{R,\max} = P_T G_T L_p^{-1} G_R \quad (22.11.3)$$

The utility of the concept of path loss is that it may also be determined for conditions which are different from free space. The Friis transmission equation still applies; one simply uses the appropriate (and probably significantly different) value of L_p .

A common misconception is that path loss is equal to the reduction in power density due to spreading along the path between antennas, and therefore this “spreading loss” increases with frequency. In fact, the reduction in power density due to spreading between any two distances $R_1 < R_2$ is:

$$\frac{P_T / 4\pi R_1^2}{P_T / 4\pi R_2^2} = \left(\frac{R_1}{R_2} \right)^2$$

which is clearly independent of frequency. The path loss L_p , in contrast, depends only on the total distance R and does depend on frequency. The dependence on frequency reflects the dependence of the effective aperture on wavelength. Thus, path loss is not loss in the traditional sense, but rather accounts for a combination of spreading and the λ^2 dependence of effective aperture that is common to all receiving antennas.

Finally, note that Equation 22.11.3 is merely the simplest form of the Friis transmission equation. Commonly encountered alternative forms include forms in which G_T and/or G_R are instead represented by the associated effective apertures, and forms in which the effects of antenna impedance mismatch and/or cross-polarization are taken into account.

✓ Example 22.11.1: 6 GHz point-to-point link

Terrestrial telecommunications systems commonly aggregate large numbers of individual communications links into a single high-bandwidth link. This is often implemented as a radio link between dish-type antennas having gain of about 27 dBi (that's dB relative to a lossless isotropic antenna) mounted on very tall towers and operating at frequencies around 6 GHz. Assuming the minimum acceptable receive power is -120 dBm (that's -120 dB relative to 1 mW; i.e., 10^{-15} W) and the required range is 30 km, what is the minimum acceptable transmit power?

Solution

From the problem statement:

$$G_T = G_R = 10^{27/10} \cong 501$$

$$\lambda = \frac{c}{f} \cong \frac{3 \times 10^8 \text{ m/s}}{6 \times 10^9 \text{ Hz}} \cong 5.00 \text{ cm}$$

$R = 30$ km, and $P_R \geq 10^{-15}$ W. We assume that the height and high directivity of the antennas yield conditions sufficiently close to free space. We further assume conjugate-matching at the receiver, and that the antennas are co-polarized. Under these conditions, $P_R = P_{R,\max}$ and Equation 22.11.2 applies. We find:

$$\begin{aligned} P_T &\geq \frac{P_{R,\max}}{G_T (\lambda / 4\pi R)^2 G_R} \\ &\cong 2.26 \times 10^{-7} \text{ W} \\ &\cong 2.26 \times 10^{-4} \text{ mW} \\ &\cong -36.5 \text{ dBm} \end{aligned}$$

Additional Reading:

- “Free-space path loss” on Wikipedia.
- “Friis transmission equation” on Wikipedia.

This page titled [22.11: Friis Transmission Equation](#) is shared under a [CC BY-SA](#) license and was authored, remixed, and/or curated by [Steven W. Ellingson \(Virginia Tech Libraries' Open Education Initiative\)](#).

- [10.14: Friis Transmission Equation](#) by [Steven W. Ellingson](#) is licensed [CC BY-SA 4.0](#). Original source: <https://doi.org/10.21061/electromagnetics-vol-2>.

CHAPTER OVERVIEW

23: Signal Modulation

- [23.1: Introduction](#)
- [23.2: Historical Context - The Origins of Radio Communication](#)
- [23.3: Radio Signal Metrics](#)
- [23.4: Modulation Overview](#)
- [23.5: Analog Modulation](#)
- [23.6: Digital Modulation](#)
- [23.7: Frequency Shift Keying, FSK](#)
- [23.8: Carrier Recovery](#)
- [23.9: Phase Shift Keying Modulation](#)
- [23.10: Quadrature Amplitude Modulation](#)
- [23.11: Digital Modulation Summary](#)
- [23.12: References](#)
- [23.13: Exercises](#)

23: Signal Modulation is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

23.1: Introduction

Most radio communication systems superimpose slowly varying information on a sinusoidal carrier that is transmitted as a radio frequency (RF) signal. This modulated RF signal is sent through a medium, usually air, by a transmitter to a receiver. In the transmitter information is initially represented at what is called baseband. The process of transferring information from baseband to the much higher frequency carrier wave is called modulation. Most modulation schemes slowly vary the amplitude and/or phase of a sinusoidal carrier waveform. In the receiver the process is reversed using demodulation to extract the baseband information from the varying state, such as the amplitude and/or phase, of the modulated carrier.

Radio has evolved subject to constraints imposed by political, hardware, and compatibility considerations. New schemes generally must be compatible and co-exist with earlier schemes. This chapter discusses the many different modulation schemes that are used in radios. Nearly all modulation schemes are supported in modern radios such as 4G and 5G cellular radios, and many are supported in WiFi. Sometimes this is to provide support for legacy radios while in other situations they are used because simpler modulation formats tolerate higher levels of interference. Indeed the level of so-

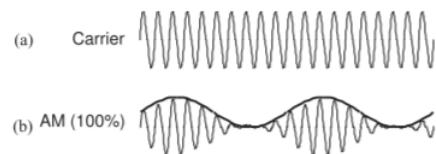


Figure 23.1.1: AM showing the relationship between the carrier and modulation envelope: (a) carrier; and (b) 100% amplitude modulated carrier.

phistication of modulation methods may need to be frequently changed to accommodate varying interference environments. Legacy analog modulation schemes and the simpler digital modulation schemes were suitable for the relatively unsophisticated hardware of years past. High-order modulation schemes enable many digital bits to be sent in each hertz of bandwidth and are only possible because of the evolution of digital signal processing and because of advances in high-density, low-power digital electronics.

Section 2.2 introduces some of the metrics that are used to compare modulation schemes and Section 2.3 introduces modulation. Section 2.4 describes analog modulation. Then Section 2.5 describes digital modulation followed by sections that deal with the specifics of various digital modulation methods: frequency shift keying (FSK) in Section 2.6; phase shift keying (PSK) in Section 2.8; and quadrature amplitude modulation (QAM) in Section 2.9. Before the discussion of PSK a concept called carrier recovery is discussed in Section 2.7 as the necessity to do this was behind the development of a variety of PSK modulation schemes. This is followed by a discussion of the metrics that can be used to quantify interference and distortion of modulated signals.

Modulation, and the hardware architectures and circuits for modulating and demodulating radio signals, are presented largely in three chapters. There is an overlap of these topics but modulation itself is largely confined to this chapter although some architecture concepts must necessarily be introduced to understand the evolution of modulation schemes. The next chapter, Chapter 3, focuses on architectures and essential circuits for modulators and demodulators.

This page titled [23.1: Introduction](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.1: Introduction](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.2: Historical Context - The Origins of Radio Communication

Communicating using EM signals has been an integral part of society since the transmission of the first telegraph signals over wires in the mid 19th century [7]. This development derived from an understanding of magnetic induction based on the experiments of Faraday in 1831 [8] in which he investigated the relationship of magnetic fields and currents. This work of Faraday is now known as Faraday's law, or Faraday's law of induction. It was one of four key laws developed between 1820 and 1835 that described the interaction of static fields and of static fields with currents. These four

Band	Frequency Range
L "long"	1 – 2 GHz
S "short"	2 – 4 GHz
C "compromise"	4 – 8 GHz
X "extended"	8 – 12 GHz
K _u "kurtz under"	12 – 18 GHz
K "kurtz" (short in German)	18 – 27 GHz
K _a "kurtz above"	27 – 40 GHz
V	40 – 75 GHz
W	75 – 110 GHz
F	90 – 140 GHz
D	110 – 170 GHz
mm	110 – 300 GHz

Table 23.2.1: IEEE radar bands [6]. The mm band designation is also used when the intent is to convey general information above 30 GHz.

Note

In Table 23.2.2 the waveguide dimensions are specified in inches (use 25.4 mm/inch to convert to mm). The number in the WR designation is the long internal dimension of the waveguide in hundredths of an inch. The EIA is the U.S.-based Electronics Industry Association. Note that the radar band (see Table 23.2.1) and waveguide band designations do not necessarily coincide.

Band	EIA Waveguide Band	Operating Frequency (GHz)	Internal Dimensions ($a \times b$, inches)
R	WR-430	1.70 – 2.60	4.300 × 2.150
D	WR-340	2.20 – 3.30	3.400 × 1.700
S	WR-284	2.60 – 3.95	2.840 × 1.340
E	WR-229	3.30 – 4.90	2.290 × 1.150
G	WR-187	3.95 – 5.85	1.872 × 0.872
F	WR-159	4.90 – 7.05	1.590 × 0.795
C	WR-137	5.85 – 8.20	1.372 × 0.622
H	WR-112	7.05 – 10.00	1.122 × 0.497
X	WR-90	8.2 – 12.4	0.900 × 0.400
Ku	WR-62	12.4 – 18.0	0.622 × 0.311
K	WR-51	15.0 – 22.0	0.510 × 0.255
K	WR-42	18.0 – 26.5	0.420 × 0.170
Ka	WR-28	26.5 – 40.0	0.280 × 0.140
Q	WR-22	33 – 50	0.224 × 0.112
U	WR-19	40 – 60	0.188 × 0.094
V	WR-15	50 – 75	0.148 × 0.074
E	WR-12	60 – 90	0.122 × 0.061
W	WR-10	75 – 110	0.100 × 0.050
F	WR-8	90 – 140	0.080 × 0.040
D	WR-6	110 – 170	0.0650 × 0.0325
G	WR-5	140 – 220	0.0510 × 0.0255

Table 23.2.2: Selected waveguide bands with operating frequencies and internal dimensions (refer to Figure 1.2.2).

laws are the Biot–Savart law (developed around 1820), Ampere's law (1826), Faraday's law (1831), and Gauss's law (1835). These are all static laws and do not describe propagating fields.

1.3.1 Electromagnetic Fields

We now know that there are two components of the EM field, the **electric field**, \mathbf{E} , with units of volts per meter (V/m), and the magnetic field, \mathbf{H} , with units of amperes per meter (A/m). \mathbf{E} and \mathbf{H} fields together describe the force between charges. There are also two flux quantities that are necessary to understand the interactions between these fields and vacuum or matter. The first is \mathbf{D} , the **electric flux** density, with units of coulombs per square meter (C/m^2), and the other is \mathbf{B} , the **magnetic flux** density, with units of teslas (T). \mathbf{B} and \mathbf{H} , and \mathbf{D} and \mathbf{E} , are related to each other by the properties of the medium, which are embodied in the quantities μ and ϵ (with the calligraphic letter, e.g. \mathcal{B} , denoting a time-domain quantity):

$$\bar{\mathcal{B}} = \mu \bar{\mathcal{H}} \quad (23.2.1)$$

$$\bar{\mathcal{D}} = \epsilon \bar{\mathcal{E}} \quad (23.2.2)$$

where the over bar denotes a vector quantity, and μ is called the **permeability** of the medium and describes the ability to store **magnetic energy** in a region. The permeability in free space (or vacuum) is denoted $\mu_0 = 4\pi \times 10^{-7} \text{ H/m}$ and the magnetic flux and magnetic field are related as

$$\overline{\mathcal{B}} = \mu_0 \overline{\mathcal{H}} \quad (23.2.3)$$

The other material quantity is the **permittivity**, ϵ , which describes the ability to store energy in a volume and in a vacuum

$$\overline{\mathcal{D}} = \epsilon_0 \overline{\mathcal{E}} \quad (23.2.4)$$

where $\epsilon_0 = 8.854 \times 10^{-12} \text{ F/m}$ is the permittivity of a vacuum. The **relative permittivity**, ϵ_r , is the ratio the permittivity of a material to that of vacuum:

$$\epsilon_r = \epsilon / \epsilon_0 \quad (23.2.5)$$

Similarly, the **relative permeability**, μ_r , refers to the ratio of permeability of a material to its value in a vacuum:

$$\mu_r = \mu / \mu_0 \quad (23.2.6)$$

1.3.2 Biot-Savart Law

The Biot–Savart law relates current to magnetic field as, see Figure 23.2.1,

$$d\overline{\mathcal{H}} = \frac{I d\ell \times \hat{\mathbf{a}}_R}{4\pi R^2} \quad (23.2.7)$$

which has the units of amperes per meter in the SI system. In Equation (23.2.7) $d\overline{\mathcal{H}}$ is the incremental static \mathcal{H} field, I is current, $d\ell$ is the vector of the length of a filament of current I , $\hat{\mathbf{a}}_R$ is the unit vector in the direction from the current filament to the magnetic field, and R is the distance between the filament and the magnetic field. The $d\overline{\mathcal{H}}$ field is directed at right angles to $\hat{\mathbf{a}}_R$ and the current filament. So Equation (23.2.7) says that a filament of current produces a magnetic field at a point. The total magnetic field from a current on a wire or surface can be found by modeling the wire or surface as a number of current filaments, and the total magnetic field at a point is obtained by integrating the contributions from each filament.

1.3.3 Faraday's Law of Induction

Faraday's law relates a time-varying magnetic field to an induced voltage drop, V , around a closed path, which is now understood to be $\oint_{\ell} \overline{\mathcal{E}} \cdot d\ell$, that is, the closed contour integral of the electric field,

$$V = \oint_{\ell} \overline{\mathcal{E}} \cdot d\ell = - \oint_s \frac{\partial \overline{\mathcal{B}}}{\partial t} \cdot ds \quad (23.2.8)$$

and this has the units of volts in the SI unit system. The operation described in Equation (23.2.8) is illustrated in Figure 23.2.2.

1.3.4 Ampere's Circuital Law

Ampere's circuital law, often called just Ampere's law, relates direct current and the static magnetic field $\overline{\mathcal{H}}$. The relationship is based on Figure 23.2.3 and Ampere's circuital law is

$$\oint_{\ell} \overline{\mathcal{H}} \cdot d\ell = I_{\text{enclosed}} \quad (23.2.9)$$

That is, the integral of the magnetic field around a loop is equal to the current enclosed by the loop. Using symmetry, the magnitude of the magnetic field at a distance r from the center of the wire shown in Figure 23.2.3 is

$$\mathcal{H} = |I| / (2\pi r) \quad (23.2.10)$$

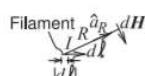


Figure 23.2.1: Diagram illustrating the Biot-Savart law. The law relates a static filament of current to the incremental \mathcal{H} field at a distance.

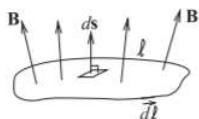


Figure 23.2.2: Diagram illustrating Faraday's law. The contour ℓ encloses the surface.

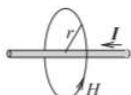


Figure 23.2.3: Diagram illustrating Ampere's law. Ampere's law relates the current, I , on a wire to the magnetic field around it, H .

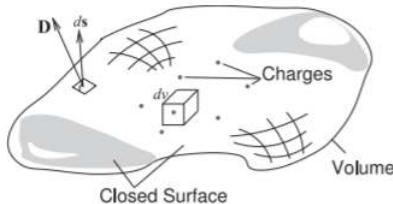


Figure 23.2.4: Diagram illustrating Gauss's law. Charges are distributed in the volume enclosed by the closed surface. An incremental area is described by the vector $d\mathbf{s}$, which is normal to the surface and whose magnitude is the area of the incremental area.

1.3.5 Gauss's Law

The final static EM law is Gauss's law, which relates the static electric flux density vector, \bar{D} , to charge. With reference to Figure 23.2.4, Gauss's law in integral form is

$$\oint_s \bar{D} \cdot d\mathbf{s} = \int_v \rho_v \cdot dv = Q_{\text{enclosed}} \quad (23.2.11)$$

This states that the integral of the electric flux vector, \bar{D} , over a closed surface is equal to the total charge enclosed by the surface, Q_{enclosed} .

1.3.6 Gauss's Law of Magnetism

Gauss's law of magnetism parallels Gauss's law which now applies to magnetic fields. In integral form the law is

$$\oint_s \bar{B} \cdot d\mathbf{s} = 0 \quad (23.2.12)$$

This states that the integral of the magnetic flux vector, \bar{D} , over a closed surface is zero reflecting the fact that magnetic charges do not exist.

1.3.7 Telegraph

With the static field laws established, the stage was set to begin the development of the transmission of EM signals over wires. While traveling by ship back to the United States from Europe in 1832, Samuel Morse learned of Faraday's experiments and conceived of an EM telegraph. He sought out partners in Leonard Gale, a professor of science at New York University, and Alfred Vail, "skilled in the mechanical arts," who constructed the telegraph models used in their experiments. In 1835 this collaboration led to an experimental version transmitting a signal over **16 km** of wire. Morse was not

Symbol	Code
1	----
2	..---
3-
4-
5
6	-....
7	--...
8	---..
9	----.
0	-----
A	. -
B	-...
C	-.-.
D	-..
E	.
F	.. .
G	--.
H
I	..
J	.---
K	-.-
L	.-..
M	--
N	- .
O	---
P	.--.
Q	--.-
R	.-.
S	...
T	-
U	..-
V	...-
W	.--

Symbol	Code
X	-..
Y	-.--
Z	--..

Table 23.2.3: International Morse code.

alone in imagining an EM telegraph, and in 1837 Charles Wheatstone opened the first commercial telegraph line between London and Camden Town, England, a distance of **2.4 km**. Subsequently, in 1844, Morse designed and developed a line to connect Washington, DC, and Baltimore, Maryland. This culminated in the first public transmission on May 24, 1844, when Morse sent a telegraph message from the Capitol in Washington to Baltimore. This event is recognized as the birth of communication over distance using wires. This rapid pace of transition from basic research into electromagnetism (Faraday's experiment) to a fielded transmission system has been repeated many times in the evolution of wired and wireless communication technology.

The early telegraph systems used EM induction and multicell batteries that were switched in and out of circuit with the long telegraph wire and so created pulses of current. We now know that these current pulses created propagating magnetic fields that were guided by the wires and were accompanied by electric fields. In 1840 Morse applied for a U.S. patent for "Improvement in the Mode of Communicating Information by Signals by the Application of Electro-Magnetism Telegraph," which described "lightning wires" and "Morse code." By 1854, **37,000 km** of telegraph wire crossed the United States, and this had a profound effect on the development of the country. Railroads made early extensive use of telegraph and a new industry was created. In the United States the telegraph industry was dominated by Western Union, which became one of the largest companies in the world. Just as with telegraph, the history of wired and wireless communication has been shaped by politics, business interests, market risk, entrepreneurship, patent ownership, and patent litigation as much as by the technology itself.

The first telegraph signals were just short bursts and slightly longer bursts of noise using Morse code in which sequences of dots, dashes, and pauses represent numbers and letters (see Table 23.2.3).¹ The speed of transmission was determined by an operator's ability to key and recognize the codes. Information transfer using EM signals in the late 19th century was therefore about **5 bits per second (bits/s)**. Morse achieved **10 words per minute**.

1.3.8 The Origins of Radio

In the 1850s Morse began to experiment with wireless transmission, but this was still based on the principle of conduction. He used a flowing river, which as is now known is a medium rich with ions, to carry the charge. On one side of the river he set up a series connection of a metal plate, a battery, a Morse key, and a second metal plate. This formed the transmitter circuit. The metal plates were inserted into the water and separated by a distance considerably greater than the width of the river. On the other side of the river, metal plates were placed directly opposite the transmitter plates and this second set of plates was connected by a wire to a galvanometer in series. This formed the receive circuit, and electric pulses established by the transmitter resulted in the charge being transferred across the river by conduction and the pulses subsequently detected by the galvanometer. This was the first wireless transmission using electromagnetism, but it was not radio.

Morse relied entirely on conduction to achieve wireless transmission and it is now known that we need alternating electric and magnetic fields to propagate information over distance without charge carriers. The next steps in the progress to radio were experiments in induction. These culminated in an experiment by Loomis who in 1866 sent the first aerial wireless signals using kites flown by copper wires [9]. The transmitter kite had a Morse key at the ground end and an electric potential would have been developed between the ground and the kite itself. Closing the key resulted in current flow along the wire and this created a magnetic field that spread out and induced a current in the receive kite and this was detected by a galvanometer. However, not much of an electric field is produced and an EM wave is not transmitted. As such, the range of this system is very limited. Practical wireless communication requires an EM wave at a high-enough frequency that it can be efficiently generated by short wires.

1.3.9 Maxwell's Equations

The essential next step in the invention of radio was the development of Maxwell's equations in 1861. Before Maxwell's equations were postulated, several static EM laws were known. These are the Biot–Savart law, Ampere's circuital law, Gauss's law, and Faraday's law. Taken together they cannot describe the propagation of EM signals, but they can be derived from Maxwell's equations. Maxwell's equations cannot be derived from the static electric and magnetic field laws. Maxwell's equations embody additional insight relating spatial derivatives to time derivatives, which leads to a description of propagating fields. Maxwell's equations are

$$\nabla \times \bar{\mathcal{E}} = -\frac{\partial \bar{\mathcal{B}}}{\partial t} - \bar{\mathcal{M}} \quad (23.2.13)$$

$$\nabla \cdot \bar{\mathcal{D}} = \rho v \quad (23.2.14)$$

$$\nabla \times \bar{\mathcal{H}} = \frac{\partial \bar{\mathcal{D}}}{\partial t} + \bar{\mathcal{J}} \quad (23.2.15)$$

$$\nabla \cdot \bar{\mathcal{B}} = \rho_m v \quad (23.2.16)$$

Several of the quantities in Maxwell's equation have already been introduced, but now the electric and magnetic fields are in vector form. The other quantities in Equations (23.2.13)–(23.2.16) are

- $\bar{\mathcal{J}}$, the **electric current density**, with units of amperes per square meter (A/m^2);
- ρv , the **electric charge density**, with units of coulombs per cubic meter (C/m^3);
- $\rho_m v$, the magnetic charge density, with units of webers per cubic meter (Wb/m^3); and
- $\bar{\mathcal{M}}$, the magnetic current density, with units of volts per square meter (V/m^2).

Magnetic charges do not exist, but their introduction through the **magnetic charge density**, $\rho_m v$, and the **magnetic current density**, $\bar{\mathcal{M}}$, introduce an aesthetically appealing symmetry to Maxwell's equations. Maxwell's equations are differential equations, and as with most differential equations, their solution is obtained with particular boundary conditions, which in radio engineering are imposed by conductors. Electric conductors (i.e., electric walls) support electric charges and hence electric current. By analogy, magnetic walls support magnetic charges and magnetic currents. Magnetic walls also provide boundary conditions to be used in the solution of Maxwell's equations. The notion of magnetic walls is important in RF and microwave engineering, as they are approximated by the boundary between two dielectrics of different permittivity. The greater the difference in permittivity, the more closely the boundary approximates a magnetic wall.

Maxwell's equations are fundamental properties and there is no underlying theory, so they must be accepted "as is," but they have been verified in countless experiments. Maxwell's equations have three types of derivatives. First, there is the time derivative, $\partial/\partial t$. Then there are two spatial derivatives, $\nabla \times$, called **curl**, capturing the way a field circulates spatially (or the amount that it curls up on itself), and $\nabla \cdot$, called the **div operator**, describing the spreading-out of a field. In rectangular coordinates, curl, $\nabla \times$, describes how much a field circles around the x , y , and z axes. That is, the curl describes how a field circulates on itself. So Equation (23.2.13) relates the amount an electric field circulates on itself to changes of the \mathbf{B} field in time. So a spatial derivative of electric fields is related to a time derivative of the magnetic field. Also in Equation (23.2.15) the spatial derivative of the magnetic field is related to the time derivative of the electric field. These are the key elements that result in self-sustaining propagation.

Div, $\nabla \cdot$, describes how a field spreads out from a point. So the presence of net electric charge (say, on a conductor) will result in the electric field spreading out from a point (see Equation (23.2.14)). In contrast, the magnetic field (Equation (23.2.16)) can never diverge from a point, which is a result of magnetic charges not existing (except when the magnetic wall approximation is used).

How fast a field varies with time, $\partial \bar{\mathcal{B}} / \partial t$ and $\partial \bar{\mathcal{D}} / \partial t$, depends on frequency. The more interesting property is how fast a field can change spatially, $\nabla \times \bar{\mathcal{E}}$ and $\nabla \times \bar{\mathcal{H}}$ —this depends on wavelength relative to geometry. So if the cross-sectional dimensions of a transmission line are less than a wavelength ($\lambda/2$ or $\lambda/4$ in different circumstances), then it will be impossible for the fields to curl up on themselves and so there will be only one solution (with no or minimal spatial variation of the \mathbf{E} and \mathbf{H} fields) or, in some cases, no solution to Maxwell's equations.

1.3.10 Transmission of Radio Signals

Now the discussion returns to the technological development of radio. About the same time as Loomis's induction experiments in 1864, James Maxwell [10] laid the foundations of modern EM theory in 1861 [11]. Maxwell theorized that electric and magnetic fields are different manifestations of the same phenomenon. The revolutionary conclusion was that if they are time varying, then they would travel through space as a wave. This insight was accepted almost immediately by many people and initiated a large number of endeavors. The period of 1875 to 1900 was a time of tremendous innovation in wireless communication.

On November 22, 1875, Edison observed EM sparks. Previously sparks were considered to be an induction phenomenon, but Edison thought that he was producing a new kind of force, which he called the etheric force. He believed that this would enable communication without wires. To put this in context, the telegraph was invented in the 1830s and the telephone was invented in 1876.

The next stage leading to radio was orchestrated by D. E. Hughes beginning in 1879. Hughes experimented with a spark gap and reasoned that in the gap there was a rapidly alternating current and not a constant current as others of his time believed. The electric oscillator was born. The spark gap transmitter was augmented with a clockwork mechanism to interrupt the transmitter circuit and produce pulsed radio signals. He used a telephone as a receiver and walked around London and detected the transmitted signals over distance. Hughes noted that he had good reception at **180 feet**. Hughes publicly demonstrated his "radio" in 1870 to the Royal Society, but the eminent scientists of the society determined that the effect was simply due to induction. This discouraged Hughes from continuing. However, Hughes has a legitimate claim to having invented radio, mobile digital radio at that, and probably was transmitting pulses on a **100 kHz** carrier. In Hugeness's radio the RF carrier was produced by the spark gap oscillator and the information was coded as pulses. It was a small leap to a Morse key-based system.

The invention of practical radio can be attributed to many people, beginning with Heinrich Hertz, who in the period from 1885 to 1889 successfully verified the essential prediction of Maxwell's equations that EM energy could propagate through the atmosphere. Hertz was much more thorough than Hughes and his results were widely accepted. In 1891 Tesla developed what is now called the Tesla coil, which is a transformer with a primary and a secondary coil, one inside the other. When one of the coils was excited by an alternating signal, a large voltage was produced across the terminals of the other coil. Tesla pursued the application of his coils to radio and realized that the coils could be tuned so that the resulting resonance greatly amplified a radio signal.

The next milestone was the establishment of the first practical radio system by Marconi, with experiments beginning in 1894. Oscillations were produced in a spark gap, which were amplified by a Tesla coil. The work culminated in the transmission of telegraph signals across the Atlantic (from Ireland to Canada) by Marconi in 1901. In 1904, crystal radio kits to detect wireless telegraph signals could be readily purchased.

Spark gap transmitters could only send pulses of noise and not voice. One generator that could be amplitude modulated was an alternator. At the end

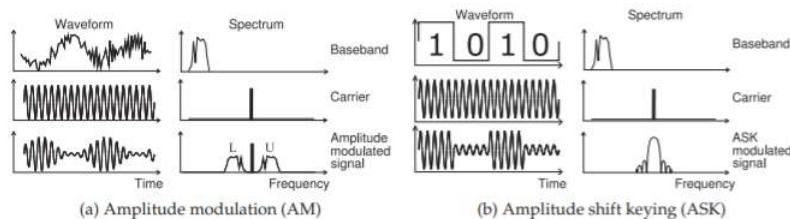


Figure 23.2.5: Waveform and spectra of simple modulation schemes. The modulating signal, at the top in (a) and (b), is also called the baseband signal.

of the 19th century, readily available alternators produced a **60 Hz** signal. Reginald Resplendent attempted to make a higher-frequency alternator and the best he achieved operated at **1 kHz**. Resplendent realized that Maxwell's equations indicated that radiation increased dramatically with frequency and so he needed a much-higher-frequency signal source. Under contract, General Electric developed a **2 kW, 100 kHz** alternator designed by Ernst Alexanderson. With this alternator, the first radio communication of voice occurred on December 23, 1900, in a transmission by Fessenden from an island in the Potomac River, near Washington,

DC. Then on December 24, 1906, Fessenden transmitted voice from Massachusetts to ships hundreds of miles away in the Atlantic Ocean. This milestone is regarded as the beginning of the radio era.

Marconi subsequently purchased 50 and 200 kW Alexanderson alternators for his trans-Atlantic transmissions. Marconi was a great integrator of ideas, with particular achievements being the design of transmitting and receiving antennas that could be tuned to a particular frequency and the development of a coherer to improve detection of a signal.

1.3.11 Early Radio

Radio works by superimposing relatively slowly varying information, at what is called the **baseband** frequency, on a carrier sinusoid by varying the amplitude and/or phase of the sinusoid. Early radio systems were based on modulating an oscillating carrier either by pulsing the carrier (using for example Morse code)—this modulation scheme is called **amplitude shift keying (ASK)**—or by varying the amplitude of the carrier, i.e. **amplitude modulation (AM)**, in the case of analog, usually voice, transmission. The waveforms and spectra of these modulation schemes are shown in Figure 23.2.5. The information is contained in the baseband signal, which is also called the modulating signal. The spectrum of the baseband signal extends to DC or perhaps down to where it rolls off at a low frequency. The carrier is a single sinewave and contains no information. The amplitude of the carrier is varied by the baseband signal to produce the modulated signal. In general, there are many cycles of the carrier relative to variations of the baseband signal so that the bandwidth of the modulated signal is relatively small compared to the frequency of the carrier.

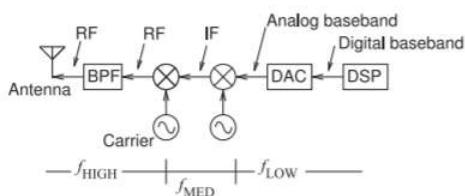


Figure 23.2.6: A simple transmitter with low, f_{LOW} , medium, f_{MED} , and high frequency, f_{HIGH} , sections. The mixers can be idealized as multipliers, shown as circles with crosses, that boost the frequency of the input baseband or IF signal by the frequency of the carrier.

AM and ASK radios are narrowband communication systems (they use a small portion of the EM spectrum), so to avoid interference with other radios it is necessary to search for an open part of the spectrum to place the carrier signal. In the decade of the 1900s there was little organization and a listener needed to search to find the desired transmission. The technology of the day necessitated this anyway, as the carrier would drift around by 10% or so since it was then not possible to build a stable oscillator. It was not until the *Titanic* sinking in 1912 that regulation was imposed on the wireless industry. Investigations of the *Titanic* sinking concluded that most of the lives lost would have been saved if a nearby ship had been monitoring its radio channels and if the frequency of the emergency channel was fixed. However, a second ship, but not close enough, did respond to *Titanic*'s “SOS” signal. A result of the investigations was the Service Regulations of the 1912 London International Radiotelegraph Convention.

These early regulations were fairly liberal and radio stations were allowed to use radio wavelengths of their own choosing, but restricted to four broad bands: a single band at 1500 kHz for amateurs; 187.5 to 500 kHz, appropriated primarily for government use; below 187.5 kHz for commercial use, and 500 kHz to 1500 kHz, also a commercial band. Subsequent years saw more stringent assignment of narrow spectral bands and the assignment of channels. The standards and regulatory environment for radio were set — there would be assigned frequency bands for particular purposes. Very quickly strong government and commercial interests struggled for exclusive use of particular bands and thus the EM spectrum developed considerable value. Entities “owned” portions of the spectrum either through a license or through government allocation.

While most of the spectrum is allocated, there are several open bands where licenses are not required. The **instrumentation, scientific, and medical (ISM)** bands at 2.4 and 5.8 GHz are examples. Since these bands are loosely regulated, radios must cope with potentially high levels of interference.

Footnotes

[1] Morse code uses sequences of dots, dashes, and spaces. The duration of a dash (or “dah”) is three times longer than that of a dot (or “dit”). Between letters there is a small gap. For example, the Morse code for PI is “. - . . .”. Between words there is a slightly longer pause and between sentences an even longer pause. Table 23.2.3 lists the international Morse code adopted in 1848. The

original Morse code developed in the 1830s is now known as “American Morse code” or “railroad code.” The “modern international Morse code” extends the international Morse code with sequences for non-English letters and special symbols.

This page titled [23.2: Historical Context - The Origins of Radio Communication](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [1.3: Communication Over Distance](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.3: Radio Signal Metrics

Radio signals are engineered to trade-off efficient use of the EM spectrum with the complexity and performance of the required RF hardware. Ultimately the goal is to efficiently use spectrum through maximal packing of information, e.g. digital bits, in a given bandwidth while, for mobile radios especially, using as little prime power as possible. The choice of the type of modulation to use is at the core of the communication system design tradeoff.

There are two families of modulation methods: analog and digital modulation. In analog modulation the RF signal has a continuous range of values; in digital modulation, the output has a number of discrete states at particular times called clock ticks, say every microsecond. There are just a few modulation schemes, all of which are digital, that achieve the optimum trade-offs of spectral efficiency and ease of use with acceptable hardware complexity. If hardware complexity is not a concern, which modulation scheme is used depends on noise and interference as well as the power required to transmit a signal, and the power required to process a received signal.

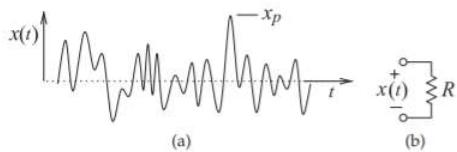


Figure 23.3.1: Definition of crest factor: (a) arbitrary waveform; and (b) voltage across a resistor.

This section introduces several metrics that characterize the variability of the amplitude of a modulated signal, and this variability has a direct impact on how analog hardware performs are designed and how efficiently hardware can be used.

2.2.1 Crest Factor and Peak-to-Average Power Ratio

Introduction

In radio engineering crest factor (CF) is a metric that describes how the voltage of a modulated carrier signal varies with time, and peak-to-average power ratio (PAPR) describes how the instantaneous power of a carrier signal varies with time. Be aware that there is one metric, **peak-to-average ratio (PAR)**, that is defined differently in the power, communications theory, and microwave communities. In some communities CF is also called the **peak-to-average ratio (PAR)**. This can lead to problems. Consider, for example, the community that works on smart power metering which combines power measurement, communications theory, and microwave design. The solution to this inevitable confusion is to skip the use of PAR and use unambiguous metrics.

Note

In standards PAR is defined as the ratio of the instantaneous peak value of a signal parameter to its time-averaged value. PAR is used with many signal parameters, e.g. voltage, current, power, and frequency [1].

Crest Factor

CF is the ratio of the maximum signal, such as a voltage, to its root-meansquare (rms) value. Referring to the arbitrary waveform shown in Figure 23.3.1(a), x_p is the absolute peak value of the waveform $x(t)$, if x_{rms} is its rms value, then the crest factor is [2]

$$\text{CF} = x_p / x_{\text{rms}} \quad (23.3.1)$$

More formally,

$$\text{CF} = \frac{\|x\|_\infty}{\|x\|_2} \quad (23.3.2)$$

where $\|x\|_\infty$ is the infinity norm, and here is the maximum value of $x(t)$, $\|x\|_\infty = \max[x(t)] = x_p$, and $\|x\|_2$ is just the rms value of $x(t)$:

$$x_{\text{rms}} = \|x\|_2 = \lim_{T \rightarrow \infty} \sqrt{\frac{1}{T} \int_0^T x(t) \cdot dt} \quad (23.3.3)$$

Note that CF is a voltage (or current) ratio rather than a power ratio. The CFs of several waveforms are given in Table 23.3.1.

Peak-to-Average Power Ratio (PAPR)

The peak-to-average power ratio (PAPR) is analogous to CF but for power. If $x(t)$ is the voltage across a resistor, as shown in Figure 23.3.1(b), then the

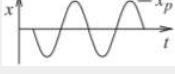
Waveform	$x(t)$	Max. value	rms (x_{rms})	CF	PAPR
DC		x_{dc}	x_{dc}	1	0 dB
Sinewave		x_p	$\frac{x_p}{\sqrt{2}}$	1.414	3.01 dB
Full-wave rectified sinewave		x_p	$\frac{x_p}{\sqrt{2}} = 0.717x_p$	1.414	3.01 dB
Half-wave rectified sinewave		x_p	$\frac{x_p}{2}$	2	6.02 dB
Triangle wave		x_p	$\frac{x_p}{\sqrt{3}} = 0.577x_p$	1.732	4.77 dB
Square wave		x_p	x_p	1	0 dB

Table 23.3.1

instantaneous peak power in the resistor is

$$P_p = |x_p|^2/R \quad (23.3.4)$$

where again x_p is the peak absolute value of the waveform. P_p is the power of the peak of a waveform treating it as though it was a DC signal. This is appropriate for a slowly varying signal such as a power frequency signal as it is this instantaneous power that determines thermal disruption of a power system. It is not the appropriate power to use with radio signals and a more suitable microwave signal metric is described in Section 2.2.2. The average power dissipated in the resistor is

$$P_{\text{avg}} = |x_{\text{rms}}|^2/R \quad (23.3.5)$$

Then

$$\text{PAPR} = \frac{P_p}{P_{\text{avg}}} = \text{CF}^2 = (x_p/x_{\text{rms}})^2 \quad (23.3.6)$$

In decibels,

$$\begin{aligned} \text{PAPR}_{\text{dB}} &= 10 \log(\text{PAPR}) \\ &= 20 \log(\text{CF}) = 20 \log(x_p/x_{\text{rms}}) \end{aligned} \quad (23.3.7)$$

The definition of PAPR above can be used with any waveform and can be used in all branches of electrical engineering. The PAPRs of several waveforms are given in Table 23.3.1.

Example 23.3.1: Crest Factor and PAPR of an Offset Sinusoid

What is the crest factor (CF) and peak-to-average power ratio (PAPR) of the signal $x(t) = 0.1 + 0.5 \sin(\omega t)$?

Solution

The signal is a sinusoid offset by a DC term. The peak value of $x(t)$ is $x_p = 0.6$, and the rms value of the signal will be the square root of the rms values squared of the individual DC and sinusoidal components. This applies to any composite signal provided that the components are uncorrelated. So $x_{\text{rms}} = \sqrt{0.12 + (0.5/\sqrt{2})^2} = 0.3674$. The general solution for a signal $x(t) = a + b \sin(\omega t)$ is, using Equation (23.3.3),

$$\begin{aligned} x_{\text{rms}} &= \sqrt{\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x(t)]^2 dt} = \sqrt{\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [a + b \sin(\omega t)]^2 dt} \\ &= \sqrt{\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [a^2 + ab \sin(\omega t) + b^2 \sin^2(\omega t)] dt} \\ &= \sqrt{\lim_{T \rightarrow \infty} \frac{1}{T} \left\{ \int_0^T a^2 dt + \int_0^T ab \sin(\omega t) dt + \int_0^T b^2 \frac{1}{2} [1 + \cos(2\omega t)] dt \right\}} \\ &= \sqrt{\lim_{T \rightarrow \infty} \frac{1}{T} \left\{ a^2 T dt + 0 + \frac{1}{2} b^2 T \right\}} \end{aligned} \quad (23.3.8)$$

since the integral of \sin and \cos over a period is zero. Thus

$$x_{\text{rms}} = \sqrt{a^2 + b^2/2} = \sqrt{0.1^2 + \frac{1}{2} 0.5^2} = 0.3674 \quad (23.3.9)$$

the crest factor is

$$\text{CF} = \frac{x_p}{x_{\text{rms}}} = \frac{0.6}{0.3674} = 1.6311 \quad (23.3.10)$$

and PAPR is

$$\text{PAPR} = 20 \log(1.6311) = 4.260 \text{ dB} \quad (23.3.11)$$

There is a quicker way of calculating PAPR by dealing with the powers directly. The peak power of the waveform is $P_p = x_p^2/R = 0.6^2/R = 0.36/R$, where x is being treated as a voltage across a resistor R . The two parts of $x(t)$, i.e. the DC component and the sinewave, are uncorrelated, so the average power of the combined signal is the sum of the powers of the uncorrelated components, so

$$P_{\text{avg}} = \frac{1}{R} \left[0.1^2 + \frac{1}{2} 0.5^2 \right] \frac{1}{R} = \frac{0.1350}{R} \quad (23.3.12)$$

Thus, in decibels,

$$\begin{aligned} \text{PAPR} \\ |\text{dB} = 10 \log \left(\frac{P_p}{P_{\text{avg}}} \right) = \frac{x_p^2}{x_{\text{rms}}^2} = 10 \log \left(\frac{0.36}{0.135} \right) = 10 \log(2.667) = 4.260 \text{ dB} \end{aligned} \quad (23.3.13)$$

2.2.2 Peak-to-Mean Envelope Power Ratio

Another metric for characterizing signals is the peak-to-mean envelope power ratio (PMEPR) and this is particularly useful for modulated signals. The amount of information sent by a communication signal is proportional to its average power, however, RF hardware must be designed with enough margin to be able to handle peaks in the signal without producing appreciable distortion. The waveform of a narrowband modulated signal appears as a carrier that slowly changes in amplitude and phase. One sinewave of this modulated signal is called a **pseudo-carrier** and the power of one cycle of the pseudo-carrier when the amplitude of the modulated signal is at its maximum (i.e. at the peak of the envelope) is called the **peak envelope power (PEP)** [1] ($\text{PEP} = P_{\text{PEP}}$). The ratio of PEP to the average signal power (the power averaged over all time) is called the PMEPR.

Then if the average power of the modulated signal is P_{avg}

$$\text{PMEPR} = \frac{\text{PEP}}{P_{\text{avg}}} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} \quad (23.3.14)$$

PMEPR is a good indicator of how sensitive a modulation format is to distortion introduced by the nonlinearity of RF hardware [3].

It is complex to determine the PMEPR for a general modulated signal. Below the mathematics is presented for an AM signal with a sinusoidal modulating signal. Determining the PMEPR otherwise requires numerical integration following the procedure outlined below.

PMEPR of an AM Signal

A good estimate of the PMEPR of an AM signal can be obtained by considering a sinusoidal modulating signal (rather than an actual baseband signal). Let $y(t) = \cos(2\pi f_m t)$ be a cosinusoidal modulating signal with frequency f_m . Then, for AM, the modulated carrier signal is

$$x(t) = A_c [1 + m \cos(2\pi f_m t)] \cos(2\pi f_c t) \quad (23.3.15)$$

where m is the modulation index (e.g. 100% AM has $m = 1$). Thus if the power of just one quasi-period of $x(t)$, i.e. one cycle of the pseudo carrier, is considered then $x(t)$ has a power that varies with time.

Consider a voltage $v(t)$ across a resistor of conductance G . The power of the signal is determined by integrating over all time, which is work, and dividing by the time period. This yields the average power:

$$P_{\text{avg}} = \lim_{\tau \rightarrow \infty} \int_{-\tau}^{\tau} \frac{1}{2\tau} G v^2(t) dt \quad (23.3.16)$$

Now, if $v(t)$ is a sinusoidal, $v(t) = A \cos \omega t$, then

$$\begin{aligned} P_{\text{avg}} &= \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} A_c^2 G \cos^2(\omega t) dt \\ &= \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} A_c^2 G \frac{1}{2} [1 + \cos(2\omega t)] dt \\ &= \frac{1}{2} A_c^2 G \left\{ \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 1 dt + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega t) dt \right\} = \frac{1}{2} A_c^2 G \end{aligned} \quad (23.3.17)$$

In the above equation, a useful equivalence has been employed by observing that the infinite integral of a cosinusoid can be simplified to just integrating over one period, $T = 2\pi/\omega$:

$$\lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos^n(\omega t) dt = \frac{1}{T} \int_{-T/2}^{T/2} \cos^n(\omega t) dt \quad (23.3.18)$$

where n is a positive integer. In power calculations there are a number of other useful simplifying techniques based on trigonometric identities. Some of the ones that will be used here are the following:

$$\begin{aligned} \cos A \cos B &= \frac{1}{2} [\cos(A - B) + \cos(A + B)] \\ \cos^2 A &= \frac{1}{2} [1 + \cos(2A)] \end{aligned} \quad (23.3.19)$$

$$\lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos \omega t dt = \frac{1}{T} \int_{-T/2}^{T/2} \cos(\omega t) dt = 0 \quad (23.3.20)$$

$$\frac{1}{T} \int_{-T/2}^{T/2} \cos^2(\omega t) dt = \frac{1}{T} \int_{-T/2}^{T/2} \frac{1}{2} [\cos(2\omega t) + \cos(0)] dt \quad (23.3.21)$$

$$= \frac{1}{2T} \left[\int_{-T/2}^{T/2} \cos(2\omega t) dt + \int_{-T/2}^{T/2} 1 dt \right] \quad (23.3.22)$$

$$= \frac{1}{2T} (0 + T) = \frac{1}{2} \quad (23.3.23)$$

More trigonometric identities are given in Appendix 1.A.2 of [4]. Also, when cosinusoids $\cos \omega_A t$ and $\cos \omega_B t$, having different frequencies ($\omega_A \neq \omega_B$), are multiplied together, for large τ ,

$$\int_{-\tau}^{\tau} \cos \omega_A t \cos \omega_B t dt = \int_{-\tau}^{\tau} \frac{1}{2} [\cos(\omega_A + \omega_B)t + \cos(\omega_A - \omega_B)t] dt = 0$$

and if $\omega_A \neq \omega_B \neq 0$

$$\int_{-\infty}^{\infty} \cos \omega_A t \cos^n \omega_B t dt = 0 \quad (23.3.24)$$

Now the discussion returns to characterizing an AM signal by considering the long-term average power and the maximum short-term power of the signal. The pseudo-carrier at its peak amplitude is, from Equation (23.3.15),

$$x_p(t) = A_c[1 + m] \cos(2\pi f_c t) \quad (23.3.25)$$

Then the power (P_{PEP}) of the peak pseudo carrier is obtained by integrating over one period of the pseudo carrier:

$$\begin{aligned} P_{\text{PEP}} &= \frac{1}{T} \int_{-T/2}^{T/2} Gx^2(t) dt = \frac{1}{T} \int_{-T/2}^{T/2} A_c^2 G(1 + m)^2 \cos^2(\omega_c t) dt \\ &= A_c^2 G(1 + m)^2 \frac{1}{T} \int_{-T/2}^{T/2} \cos^2(\omega_c t) dt = \frac{1}{2} A_c^2 G(1 + m)^2 \end{aligned} \quad (23.3.26)$$

The **average power** (P_{avg}) of the modulated signal is obtained by integrating over all time, so

$$\begin{aligned} P_{\text{avg}} &= \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} Gx^2(t) dt \\ &= A_c^2 G \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \{[1 + m \cos(\omega_m t)] \cos(\omega_c t)\}^2 dt \\ &= A_c^2 G \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \{[1 + 2m \cos(\omega_m t) + m^2 \cos^2(\omega_m t)] \cos^2(\omega_c t)\} dt \\ &= A_c^2 G \left[\lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos^2(\omega_c t) dt + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 2m \cos(\omega_m t) \cos^2(\omega_c t) dt \right. \\ &\quad \left. + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} m^2 \cos^2(\omega_m t) \cos^2(\omega_c t) dt \right] \\ &= A_c^2 G \left\{ \frac{1}{2} + 0 + m^2 \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \frac{1}{4} [1 + \cos(2\omega_m t)] [1 + \cos(2\omega_c t)] dt \right\} \\ &= A_c^2 G \left\{ \frac{1}{2} + \frac{m^2}{4} \left[\lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 1 dt + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_m t) dt \right. \right. \\ &\quad \left. \left. + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_c t) dt + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_m t) \cos(2\omega_c t) dt \right] \right\} \\ &= A_c^2 G \left[\frac{1}{2} + m^2 \left(\frac{1}{4} + 0 + 0 + 0 \right) \right] = \frac{1}{2} A_c^2 G(1 + m^2/2) \end{aligned} \quad (23.3.27)$$

Thus the rms voltage, x_{rms} , can be determined as $P_{\text{avg}} = x_{\text{rms}}^2 G$. So the PMEPR of an AM signal (i.e., PMEPR_{AM}) is

$$\text{PMEPR}_{\text{AM}} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} = \frac{\frac{1}{2} A_c^2 G(1 + m)^2}{\frac{1}{2} A_c^2 G(1 + m^2/2)} = \frac{(1 + m)^2}{1 + m^2/2}$$

For 100% AM described by $m = 1$, the PMEPR is

$$\text{PMEPR}_{100\% \text{AM}} = \frac{(1 + 1)^2}{1 + 1^2/2} = \frac{4}{1.5} = 2.667 = 4.26 \text{ dB} \quad (23.3.28)$$

In expressing the PMEPR in decibels, the formula $\text{PMEPR}_{\text{dB}} = 10 \log(\text{PMEPR})$ is used as PMEPR is a power ratio. As an example, for 50% AM, described by $m = 0.5$, the PMEPR is

$$\text{PMEPR}_{50\% \text{AM}} = \frac{(1 + 0.5)^2}{1 + 0.5^2/2} = \frac{2.25}{1.125} = 2 = 3 \text{ dB} \quad (23.3.29)$$

2.2.3 Two-Tone Signal

In assessing, either through laboratory measurements or simulations, it is common and often necessary to use very simple representations of a baseband signal or even of a modulated signal. This greatly simplifies matters and there is a justified expectation that the performance with the test signal is a good indication of performance with an actual baseband or modulated signal. With simulation at the circuit level it is usually impossible to consider real baseband signals as simulation may not even be possible or simulation may take unacceptable times. Instead it is common to use single-tone, i.e. single sinewave, or two-tone signals. A two-tone signal is a signal that is the sum of two cosinusoids:

$$y(t) = X_A \cos(\omega_A t) + X_B \cos(\omega_B t) \quad (23.3.30)$$

Generally the frequencies of the two tones are close ($|\omega_A - \omega_B| \ll \omega_A$), with the concept being that both tones fit within the passband of a transmitter's or receiver's bandpass filters. A two-tone signal is not a form of modulation, but is commonly used to characterize the nonlinear performance of RF systems and has an envelope that is similar to that of many modulated signals. The composite signal, $y(t)$, looks like a pseudo-carrier with a slowly varying amplitude, not unlike an AM signal. The tones are uncorrelated so that the average power of the composite signal, $y(t)$, is the sum of the powers of each of the individual tones. The peak power of the composite signal is that of the peak pseudo-carrier, so $y(t)$ has a peak amplitude of $\sqrt{X_A + X_B}$. The peak pseudo carrier is the single RF sinusoid where the sinusoids of each sinusoid align as much as possible. Similar concepts apply to three-tone and n -tone signals.

Example 23.3.2: PMEPR of a Two-Tone Signal

What is the PMEPR of a two-tone signal with the tones having equal amplitude?

Solution

Let the amplitudes of the two tones be X_A and X_B . Now $X_A = X_B = X$, and so the peak pseudo-carrier has amplitude $2X$, and the power of the peak RF carrier is proportional to $\frac{1}{2}(2X)^2 = 2X^2$. The average power is proportional to $\frac{1}{2}(X_A^2 + X_B^2) = \frac{1}{2}(X^2 + X^2) = X^2$, as each tone is independent of the other and so the powers can be added.

$$\text{PMEPR} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} = \frac{2X^2}{X^2} = 2 = 3 \text{ dB} \quad (23.3.31)$$

Example 23.3.3: PMEPR of Uncorrelated Signals

Consider the combination of two uncorrelated analog signals, e.g. a two-tone signal. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = 0.1 \sin(10^9 t)$ and $y(t) = 0.05 \sin(1.01 \cdot 10^9 t)$. What is the PMEPR of this combined signal?

Solution

These two signals are uncorrelated and this is key in determining the average power, P_{avg} , as the sum of the powers of each individual signal (k is a proportionality constant):

$$P_{\text{avg}} = \int_{-\infty}^{\infty} x^2(t) \cdot dt + \int_{-\infty}^{\infty} y^2(t) \cdot dt = \frac{k}{2}(0.1)^2 + \frac{k}{2}(0.5)^2 = \frac{k}{2}[0.01 + 0.0025] = 0.00625k$$

The two carriers are close in frequency so that the sum signal $z(t) = x(t) + y(t)$ looks like a slowly varying signal with a radian frequency near 10^9 rads per second. The peak amplitude of one pseudo-cycle of $z(t)$ is $0.1 + 0.05 = 0.15$. Thus the power of the largest cycle is

$$P_{\text{PEP}} = \frac{1}{2}k(0.15)^2 = 0.01125k$$

and so

$$\text{PMEPR} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} = \frac{0.01125}{0.00625} = 1.8 = 2.55 \text{ dB} \quad (23.3.32)$$

Summary

The PMEPR is an important attribute of a modulation format and impacts the types of circuit designs that can be used. It is much more challenging to develop power-efficient hardware introducing only low levels of distortion when the PMEPR is high.

It is tempting to consider if the lengthy integrations can be circumvented. Powers can be added if the signal components (the tones making up the signal) are uncorrelated. If they are correlated, then the complete integrations are required. Consider two uncorrelated sinusoids of (average) powers P_1 and P_2 , respectively, then the average power of the composite signal is $P_{\text{avg}} = P_1 + P_2$. However, in determining the peak sinusoidal power, the RF cycle where the two largest pseudo-carrier sinusoids align is considered, and here the voltages add to produce a single cycle of a sinewave with a higher amplitude. So peak power applies to just one RF pseudo-cycle. Generally the voltage amplitude of the two sinewaves would be added and then the power calculated. If the uncorrelated carriers are modulated and the modulating signals (the baseband signals) are uncorrelated, then the average power

can be determined in the same way, but the peak power calculation is much more complicated. The integrations are the only calculations that can always be relied on and can be used with all modulated signals.

Note

Signals $x(t)$ and $y(t)$ are **uncorrelated** if the integral over all time and time offsets of their product is zero: $C = \int_{-\infty}^{+\infty} x(t)y(t + \tau)dt = 0$ for all τ .

The preferred usage of PAR, PAPR, or PMEPR in RF and microwave engineering is currently in a transition phase. The most common usage of PAR and PAPR in electrical engineering refers to the peak of a signal as being the instantaneous peak value, and in the case of PAPR, the instantaneous power of the signal is calculated as if the peak is a DC value. In the past, many RF and microwave publications have taken the peak as the peak power of a sinusoid having an amplitude equal to the peak voltage of the signal and used that to calculate PAR. This usage is inconsistent with the predominant usage in electrical engineering and is a particular problem when using wireless technology in other disciplines. PMEPR is the preferred usage for what RF and microwave engineers intend to refer to when using the term PAR. A reader of RF literature encountering PAR needs to determine how the term is being used. There is no confusion if PMEPR is used.

Example 23.3.4: PAPR and PMEPR of an AM Signal

What is the PAPR and PMEPR of a 100% AM signal?

Solution

The signal is $x(t) = A_c[1 + \cos 2\pi f_m t] \cos 2\pi f_c t$ and the PMEPR of this signal, from Equation (23.3.28), is 4.26 dB. Now PAPR uses the absolute maximum value of the signal rather than the maximum short-term power of the envelope. The peak value of $x(t)$ is $2A_c$ so the peak power (if the signal is a voltage across a conductance G) is

$$P_{\text{peak, PAPR}} = (2A_c)^2 G \quad (23.3.33)$$

P_{avg} is the same for PAPR and PMEPR for the AM signal, see Equation (23.3.27), so that

$$\text{PAPR} = \frac{P_{\text{peak, PAPR}}}{P_{\text{avg}}} = \frac{(2A_c)^2 G}{\frac{1}{2}A_c^2(1 + \frac{1}{2})} = \frac{4}{3/4} = \frac{16}{3} = 5.333 = 7.27 \text{ dB} \quad (23.3.34)$$

So PAPR is 3 dB higher than PMEPR for a 100% modulated AM signal, see Equation (23.3.28). This is not always the case for other modulation schemes.

This page titled [23.3: Radio Signal Metrics](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.2: Radio Signal Metrics](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.4: Modulation Overview

There are two families of modulation methods with analog modulation used in early radios including 1G cellular radio, and digital modulation used in modern radios starting with 2G cellular radio. While 1G cellular radio transmitted voice signals using analog modulation, 1G also used a simple type of digital modulation for signaling. With the exception of **ultra-wideband (UWB)** pulse radio [5], all modern radio modulation schemes slowly vary the amplitude, phase, or frequency of a sinusoidal signal called the carrier. This results in a narrow bandwidth modulated signal perhaps with fractional bandwidth typically in the range of **0.002%** to **2%**. The early spark-gap wireless telegraph systems were ultra-wideband but they were soon discontinued because they interfered with conventional radios which were soon developed and assigned specific parts, i.e. bands, of the spectrum. The initial pulse radio concept of the 1990s occupied most of the spectrum between **3.1** and **10.6 GHz** but was never deployed mainly because capacity was relatively poor. The term ultra-wideband wireless is now widely taken to mean a wireless device such as a radar or radio with a bandwidth which is at least the lesser of **500 MHz** or **20%** of the carrier frequency [6]. So even the UWB millimeter-wave radios exploiting the high bandwidth available at millimeter wave frequencies still employ a relatively slowly varying modulation of a carrier.

This page titled [23.4: Modulation Overview](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.3: Modulation Overview](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.5: Analog Modulation

The waveforms and spectra of the signals with common analog modulation methods are shown in Figure 23.5.1. The modulating signal is generally referred to as the baseband signal and it contains all of the information to be transmitted and interpreted at the receiver. The waveforms in Figure 23.5.1 are stylized. They are presented this way so that the effects of modulation can be more easily seen. The baseband signal (Figure 23.5.1(a)) is shown as having a period that is not much greater than the period of the carrier (Figure 23.5.1(b)). In reality there would be hundreds or thousands of RF cycles for each cycle of the baseband signal so that the highest frequency component of the baseband signal is a tiny fraction of the carrier frequency. In this situation the spectra shown on the right in Figure 23.5.1(c–e) would be too narrow to enable any detail to be seen.

2.4.1 Amplitude Modulation

Amplitude Modulation (AM) is the simplest analog modulation method to implement. Here a signal is used to slowly vary the amplitude of the carrier according to the level of the modulating signal. With AM (Figure 23.5.1(c)) the amplitude of the carrier is modulated, and this results in a broadening of the spectrum of the carrier, as shown in Figure 23.5.1(c)(ii). This spectrum contains the original carrier component and upper and lower sidebands, designated as U and L, respectively. In AM, the two sidebands contain identical information, so all the information contained in the baseband signal is conveyed if just one sideband is transmitted.

The basic AM signal $x(t)$ has the form

$$x(t) = A_c [1 + my(t)] \cos(2\pi f_c t) \quad (23.5.1)$$

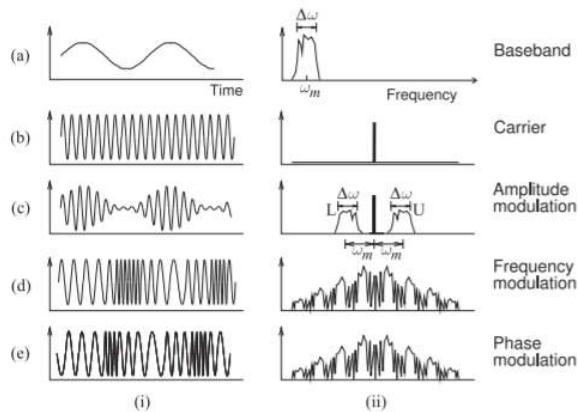


Figure 23.5.1: Basic analog modulation showing the (i) waveform and (ii) spectrum for (a) baseband signal; (b) carrier; (c) carrier modulated using amplitude modulation; (d) carrier modulated using frequency modulation; and (e) carrier modulated using phase modulation.

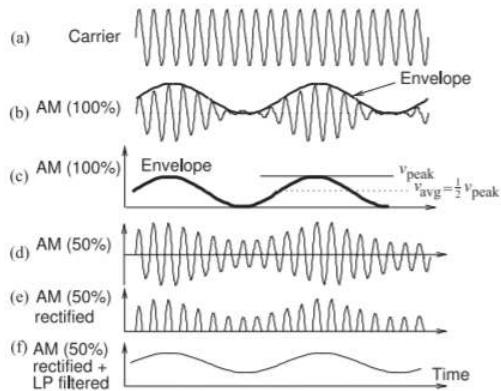


Figure 23.5.2: AM showing the relationship between the carrier and modulation envelope: (a) carrier; (b) 100% amplitude modulated carrier; (c) modulating or baseband signal; (d) 50% amplitude-modulated carrier; (e) rectified 50% AM modulated signal;

and (f) rectified and lowpass (LP) filtered 50% modulated signal. The envelope contains only amplitude information and for AM the envelope is the same as the baseband signal.

where m is the modulation index, $y(t)$ is the baseband information-bearing signal that has frequency components that are much lower than the carrier frequency f_c , and the maximum value of $|y(t)|$ is one. Provided that $y(t)$ varies slowly relative to the carrier, $x(t)$ looks like a carrier whose amplitude varies slowly. To get an idea of how slowly the amplitude varies in an actual system, consider an AM radio that broadcasts at **1 MHz** (which is in the middle of the AM broadcast band). The highest frequency component of the modulating signal corresponding to voice is about **4 kHz**. Thus the amplitude of the carrier takes **250** carrier cycles to go through a complete amplitude variation. At all times a cycle of the modulated carrier, the pseudo-carrier, appears to be periodic, but in fact it is not quite.

The concept of the envelope of a modulated RF signal is introduced in Figure 23.5.2. The envelope is an important concept and is directly related to the distortion introduced by analog hardware and to the DC power requirements which determines the battery life for mobile radios. Figure 23.5.2(a) is the carrier and the amplitude-modulated carrier is shown in Figure 23.5.2(b). The outline of the modulated carrier is called the envelope, and for AM this is identical to the modulating, i.e. baseband, signal. The envelope is shown again in Figure 23.5.2(c). At the peak of the envelope, the RF signal has maximum short-term power (considering the power of a single RF cycle). With **100% AM**, $m = 1$ in Equation (23.5.1), there is no short-term RF power when the envelope is at its minimum. The modulated signal with **50% modulation**, $m = 0.5$, is shown in Figure 23.5.2(d) and at all times there is an appreciable RF signal power.

Very simple analog hardware is required to demodulate the basic amplitude modulated signal, that is an AM signal with a carrier and both sidebands. The receiver requires bandpass filtering to select the channel from the incoming radio signal then rectifying the output of the bandpass filter. The waveform after rectification of a 50% AM signal is shown in Figure 23.5.2(e) and contains frequency components at baseband and sidebands around harmonics of the carrier, and the harmonics of the carrier itself. Lowpass filtering of the rectified waveform extracts the original baseband signal and completes demodulation, see Figure 23.5.2(f). The only electronics required is a

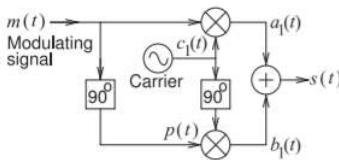


Figure 23.5.3: Hartley modulator implementing singlesideband suppressed-carrier (SSB-SC) modulation. The “90°” blocks shift the phase of the signal by +90°. The mixer indicated by the circle with a cross is an ideal multiplier, e.g. $a_1(t) = m(t) \cdot c_1(t)$.

single diode. The disadvantage is that more spectrum is used than required and the largest signal is the carrier that conveys no information but causes interference in other radios. Without the carrier and both sidebands being transmitted it is necessary to use DSP to demodulate the signal.

It is not possible to represent an actual baseband signal in a simple way and undertake the analytic derivations that illustrate the characteristics of modulation. Instead it is usual to use either a one-tone or two-tone signal, derive results, and then extrapolate the results for a finite bandwidth baseband signal. For a single-tone baseband signal $y(t) = \cos(\omega_m t + \phi)$, then the basic AM modulated signal, from Equation (23.5.1), is

$$\begin{aligned} x(t) &= [1 + m \cos(\omega_m t + \phi)] \cos(\omega_c t) \\ &= \cos(\omega_c t) + \frac{1}{2} m [\cos((\omega_c - \omega_m)t - \phi) + \cos((\omega_c + \omega_m)t + \phi)] \end{aligned} \quad (23.5.2)$$

which has three (radian) frequency components, one at the carrier frequency ω_c , one just below the carrier at $\omega_c - \omega_m$, and one just above at $\omega_c + \omega_m$ (since $\omega_m \ll \omega_c$). The extension to a finite bandwidth baseband signal, see Figure 23.5.1(a)(ii), is to imagine that ω_m ranges from a lower value $\omega_m - \frac{1}{2}\Delta\omega$ to a higher value $\omega_m + \frac{1}{2}\Delta\omega$. The discrete tones in the modulated signal below and above the carrier then become finite bandwidth sidebands with a lower sideband L centered at $\omega_c - \omega_m$ and an upper sideband U centered at $(\omega_c + \omega_m)$ each having the same bandwidth, $\Delta\omega$, as the baseband signal, see Figure 23.5.1(c)(ii).

The AM modulator described so far produces a modulated signal with a carrier and two sidebands. This modulation is called double-sideband (DSB) modulation. There is identical information in each of the sidebands and so only one of the sidebands needs to be transmitted. The carrier contains no information so if only one sideband was transmitted then the received **single-sideband (SSB) suppressed-carrier SC** (together **SSB-SC**) signal has all of the information needed to recover the original baseband signal. However the simple demodulation process using rectification as described earlier in this section no longer works. The receiver needs to use DSP but the spectrum is used efficiently.

One circuit that implements SSB-SC AM is the **Hartley modulator** shown in Figure 23.5.3. As will be seen, this basic architecture is significant and used in all modern radios. In modern radios the Hartley modulator, or a variant, takes a modulated signal which is centered at an intermediate frequency and shifts it up in frequency so that it is centered at another frequency a little below or a little above the carrier of the Hartley modulator.

In a Hartley modulator both the modulating signal $m(t)$ and the carrier are multiplied together in a mixer and then also 90° phase-shifted versions are mixed before being added together. The signal flow is as follows beginning with $m(t) = \cos(\omega_m t + \phi)$, $p(t) = \cos(\omega_m t + \phi - \pi/2) = \sin(\omega_m t + \phi)$ and carrier signal $c_1(t) = \cos(\omega_c t)$:

$$\begin{aligned} a_1(t) &= \cos(\omega_m t + \phi) \cos(\omega_c t) = \frac{1}{2} [\cos((\omega_c - \omega_m)t - \phi) + \cos((\omega_c + \omega_m)t + \phi)] \\ b_1(t) &= \sin(\omega_m t + \phi) \sin(\omega_c t) = \frac{1}{2} [\cos((\omega_c - \omega_m)t - \phi) - \cos((\omega_c + \omega_m)t + \phi)] \\ s(t) &= a_1(t) + b_1(t) = \cos((\omega_c - \omega_m)t - \phi) \end{aligned} \quad (23.5.3)$$

and so the lower sideband (LSB) is selected. An interesting observation is that the phase, ϕ , of the baseband signal is also translated up in frequency. A feature that is not exploited in AM but is in digital modulation.

2.4.2 Phase Modulation

In phase modulation (PM) the phase of the carrier depends on the instantaneous level of the baseband signal. The phase-modulated carrier is shown in Figure 23.5.1(e)(i) and it looks like the frequency of the modulated carrier is changing. What is actually happening is that when the phase is changing most quickly the apparent frequency of the RF waveform changes. Here, as the baseband signal is decreasing, the phase shift reduces and the effect is to increase the apparent frequency of the RF signal. As the baseband signal increases, the effect is to reduce the apparent frequency of the modulated RF signal. The result is that with PM is that the bandwidth of the time-varying signal is spread out, as seen in Figure 23.5.4. PM can be implemented using a phase-locked loop (PLL) but further details will be skipped here.

Consider a phase-modulated signal $s(t) = \cos(\omega_c t + \phi(t))$ where $\phi(t)$ is the baseband signal containing the information to be transmitted. The spectrum of $s(t)$ can be determined by simplifying $\phi(t)$ as a sinusoid with frequency $f_m = 2\pi\omega_m$ so that $\phi(t) = \beta \cos(\omega_m t)$ where β is the phase modulation index. (The maximum possible phase change is $\pm\pi$ and then $\beta = \pi$.) The phase-modulated signal becomes

$$\begin{aligned} s(t) &= \cos(\omega_c t + \beta \cos(\omega_m t)) \\ &= \cos(\omega_c t) \cos(\cos(\beta \omega_m t)) - \sin(\omega_c t) \sin(\cos(\beta \omega_m t)) \end{aligned} \quad (23.5.4)$$

which has the Bessel function-based expansion

$$\begin{aligned} s(t) &= J_0(\beta) \cos(\omega_c t) \\ &+ J_1(\beta) \cos(\omega_c + \omega_m)t + \pi/2) + J_1(\beta) \cos(\omega_c - \omega_m)t + \pi/2) \\ &+ J_2(\beta) \cos(\omega_c + 2\omega_m)t + \pi) + J_2(\beta) \cos(\omega_c - 2\omega_m)t + \pi) \\ &+ J_3(\beta) \cos(\omega_c + 3\omega_m)t + 3\pi/2) + J_3(\beta) \cos(\omega_c - 3\omega_m)t + 3\pi/2) + \dots \end{aligned} \quad (23.5.5)$$

where J_n is the Bessel function of the first kind of order n . The spectrum of this signal is shown in Figure 23.5.4(a) which consists of discrete tones grouped as lower- and upper-sideband sets centered on the carrier at f_c . The discrete tones in the sidebands are separated from each other and from f_c by f_m . The sidebands have lower amplitude further away from the carrier.

If the modulating signal has a finite bandwidth, approximated by f_m varying from a minimum value, $(f_m - \Delta f)$ up to the maximum frequency ($f_m + \Delta f$), then the spectrum of the modulated signal becomes that shown in Figure 23.5.4(b), with the centers of adjacent sidebands separated by f_m and the first sidebands separated from the carrier by f_m as well. This is DSB

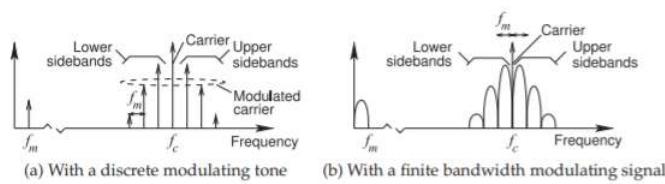


Figure 23.5.4: Spectrum of a phase-modulated carrier which includes the carrier at f_c and upper and lower sidebands with the spectrum of the discrete modulating signal at f_m .

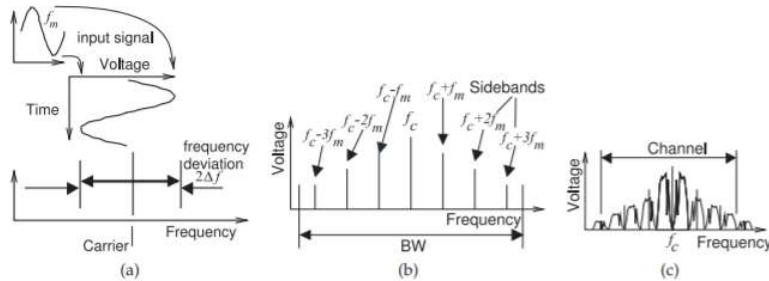


Figure 23.5.5: Frequency modulation: (a) sinusoidal baseband signal shown varying the frequency of the carrier and so FM modulating the carrier; (b) the spectrum of the resulting FM-modulated waveform; and (c) spectrum of the modulated carrier when it is modulated by a broadband baseband signal such as voice.

modulation and there is a carrier (so it is not suppressed). The sidebands do not carry identical information and several, perhaps three below and three above the carrier, are required to enable demodulation of a PM signal. Thus a rather large bandwidth is required to transmit the modulated signal.

2.4.3 Frequency Modulation

The other analog modulation schemes commonly used is frequency modulation (**FM**), see Figure 23.5.1(d). The signals produced by FM and PM appear to be similar; the difference is in how the signals are generated. In FM, the amplitude of the baseband signal determines the frequency of the modulated carrier. Consider the FM waveform in Figure 23.5.1(d)(i). When the baseband signal is at its peak value the modulated carrier is at its minimum frequency, and when the signal is at its lowest value the modulated carrier is at its maximum frequency. (Depending on the hardware implementation it could be the other way around.) The result is that the bandwidth of the time-varying signal is spread out, as seen in Figure 23.5.5.

One way of implementing the FM modulator is to use a voltage-controlled oscillator (VCO) with the baseband signal controlling the frequency of an oscillator. An FM receiver must compress, in frequency, the transmitted signal to re-create the original narrower bandwidth baseband signal. FM demodulation can be thought of as providing signal enhancement or equivalently noise suppression in a process that can be called analog processing gain. Only the components of the original FM signals are coherently collapsed to a narrower bandwidth baseband signal while noise, being uncorrelated, is still spread out (although rearranged). Thus the ratio of the signal to noise powers increases, as after demodulation only the power of the noise in the smaller bandwidth of the baseband signal is important. Thus compared to AM, FM significantly increases the tolerance to noise that may be added to the signal during transmission. PM has the same property, although the details of modulation and demodulation are different. For both FM and PM signals the peak amplitude of the RF **phasor** is equal to the average amplitude, and so the PMEPR is 1 or 0 dB.

Consider an FM signal $s(t) = \cos([\omega_c + x(t)]t)$ where $x(t)$ is the baseband signal containing the information to be transmitted. The spectrum of $s(t)$ can be determined by simplifying $x(t)$ as a sinusoid with frequency $f_m = 2\pi\omega_m$ so that $x(t) = \beta \cos(\omega_m t)$ where β is the frequency modulation index. The FM signal becomes

$$\begin{aligned} s(t) &= \cos([\omega_c + \beta \cos(\omega_m t)]t) \\ &= \cos(\omega_c t) \cos(\cos(\omega_m t)\beta t) - \sin(\omega_c t) \sin(\cos(\omega_m t)\beta t) \end{aligned} \quad (23.5.6)$$

which has the Bessel function-based expansion

$$\begin{aligned}
 s(t) = & J_0(\beta t) \cos(\omega_c t) \\
 & - J_1(\beta) \sin(\omega_c + \omega_m)t + \pi/2 - J_1(\beta t) \sin(\omega_c - \omega_m)t + \pi/2 \\
 & - J_2(\beta t) \cos(\omega_c - 2\omega_m)t + \pi + K_2(\beta t) \cos(\omega_c - 2\omega_m)t + \pi \\
 & + J_3(\beta t) \sin(\omega_c + 3\omega_m)t + 3\pi/2 + J_3(\beta t) \sin(\omega_c - 3\omega_m)t + 3\pi/2 + \dots
 \end{aligned} \tag{23.5.7}$$

where J_n is the Bessel function of the first kind of order n . The spectrum of this signal is shown in Figure 23.5.5(b) which consists of discrete tones grouped as lower- and upper-sideband sets centered on the carrier at f_c . The discrete tones in the sidebands are separated from each other and from f_c by f_m . The sidebands have lower amplitude further away from the carrier.

If the modulating signal has a finite bandwidth, approximated by $f_m = \omega_m/(2\pi)$ varying from a minimum value ($f_m - \Delta f$) up to the maximum frequency ($f_m + \Delta f$), then the spectrum of the modulated signal becomes that shown in Figure 23.5.5(c) with the centers of adjacent sidebands separated by f_m and the first sidebands from the carrier by f_m as well. This is DSB modulation and there is a carrier (so it is not suppressed but is smaller than with AM). The sidebands do not carry identical information and several, perhaps three on either side of the carrier, are required to enable demodulation of an FM signal. Thus a rather large bandwidth is required to transmit the modulated signal as it is not sufficient to transmit just one sideband to enable demodulation.

Carson's Rule

Frequency- and phase-modulated signals have a very wide spectrum and the bandwidth required to reliably transmit a PM or FM signal is subjective. The best accepted criterion for determining the bandwidth requirement is called Carson's bandwidth rule or just Carson's rule [7, 8].

An FM signal is shown in Figure 23.5.5. In particular, Figure 23.5.5(a) shows the FM function. The level (typically voltage) of the baseband signal determines the frequency deviation of the carrier from its unmodulated value. The frequency shift when the modulating signal is a DC value x_m at its maximum amplitude is called the peak frequency deviation, Δf . So, if the modulating signal changes very slowly, the bandwidth of the modulated signal is $2\Delta f$.

A rapidly varying sinusoidal modulating signal produces a modulated signal with many discrete sidebands as seen in Figure 23.5.5(b). If the modulating baseband signal is broadband, then the sidebands have finite bandwidth as seen in Figure 23.5.5(c) and many are required to recover the original baseband signal. These sidebands continue indefinitely in frequency but rapidly reduce in power away from the frequency of the unmodulated carrier. Carson's rule provides an estimate of the bandwidth that contains 98% of the energy. If the maximum frequency of the modulating signal is f_m , and the maximum value of the modulating waveform is x_m (which would produce a frequency deviation of Δf if it is DC), then Carson's rule is that the

$$\text{bandwidth required} = 2 \times (f_m + \Delta f) \tag{23.5.8}$$

Narrowband and Wideband FM

The most common type of FM signal, as used in FM broadcast radio, is called wideband FM, as the maximum frequency deviation is much greater than the highest frequency of the modulating or baseband signal, that is, $\Delta f \gg f_m$. In narrowband FM, Δf is close to f_m . Narrowband FM uses less bandwidth but requires a more sophisticated demodulation technique.

Example 23.5.1: PAPR and PMEPR of FM Signals

Consider FM signals close in frequency but whose spectra do not overlap.

- What are PAPR and PMEPR of just one FM signal?
- What are PAPR and PMEPR of a signal comprised of two uncorrelated narrowband FM signals each having a small fractional bandwidth and having the same average power.

Solution

- An FM signal has a constant envelope just like a single sinusoid, and so $\text{PAPR} = 1.414 = 3.01 \text{ dB}$ and $\text{PMEPR} = 1 = 0 \text{ dB}$.
- Since the modulation is relatively slow, each of the FM signals will look like single tone signals and the combined signal will look like a two-tone signal. However this is not enough to solve the problem. A thought experiment is required to determine the largest pseudo-carrier when the FM signals combine. If the amplitude of each tone is X , then the amplitude when the FM signal waveforms align is $2X$. (This is the same as the peak of a two-tone signal but arrived at differently.) Then

$$P_{\text{avg}} = \text{sum of the powers of each FM signal} = 2k \frac{1}{2} X^2$$

where k is a proportionality constant. For PAPR,

$$P_P = k(2X)^2 \quad \text{and} \quad \text{PMEPR} = \frac{P_P}{P_{\text{avg}}} = \frac{k(2X)^2}{2k \frac{1}{2} X^2} = 4 = 6.0 \text{ dB} \quad (23.5.9)$$

For PMEPR, $P_{\text{PEP}} = \text{power of the pseudo-carrier} = k \frac{1}{2} (2X)^2$
and

$$\text{PMEPR} = \frac{P''_P}{P_{\text{avg}}} = \frac{k \frac{1}{2} (2X)^2}{2k \frac{1}{2} X^2} = 2 = 3.0 \text{ dB} \quad (23.5.10)$$

2.4.4 Analog Modulation Summary

Analog modulation was used in the first radios and in 1G cellular radios. Radio transmission using analog modulation, i.e. analog radio, has almost ceased as it does not use spectrum efficiently. Digital modulation along with error correction, can pack much more information in a limited bandwidth. A final comparison of the analog modulation techniques is given in Figure 2.5.1 emphasizing the PMEPR of AM and FM. The PMEPR of PM is the same as for FM.

One particular event in the development of radio is illustrative of the relationship of technology and business interests. Frequency modulation was invented by Edwin H. Armstrong and patented in 1933 [9, 10]. FM is virtually static free and clearly superior to AM radio. However, it was not immediately adopted largely because AM radio was established in the 1930s, and the adoption of FM would have resulted in the scrapping of a large installed infrastructure (seen as a commercial catastrophe) and so the introduction of FM was delayed by decades. The best technology does not always win immediately! Commercial interests and the large investment in an alternative technology have a great deal to do with the success of a technology [11].

With FM and PM there are two sets of sidebands with one set above the carrier frequency and the other set below. The carrier itself is low-level but is not completely suppressed. Now SSB modulation refers to producing just one of the sideband sets. There is such as thing as SSB FM with just a few sidebands below (or above) the carrier but it is more like a combination of FM and AM [12], and was never deployed.

This page titled [23.5: Analog Modulation](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.4: Analog Modulation](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.6: Digital Modulation

Digital radio transmits bits by creating discrete states, usually discrete amplitudes and phases of a carrier. The process of creating these discrete states from a digital bitstream is called digital modulation. A state is established at a particular time called a clock tick. What that means is that the

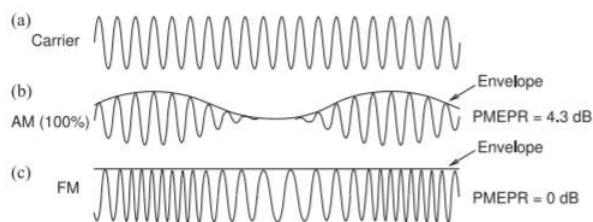


Figure 23.6.1: Comparison of 100% AM and FM highlighting the envelopes of both: (a) carrier; (b) AM signal; and (c) FM signal with constant envelope.

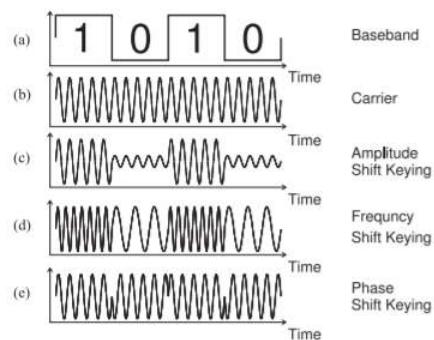


Figure 23.6.2: Modes of digital modulation: (a) modulating bitstream; (b) carrier; (c) carrier modulated using amplitude shift keying (ASK); (d) carrier modulated using frequency shift keying (FSK); and (e) carrier modulated using binary phase shift keying (BPSK).

information in the signal is the state of the waveform, such as the amplitude and the phase of a phasor, at every clock tick such as every microsecond. The time it takes to go from one state to another (a clock tick interval) defines the bandwidth of the modulated signal. For example, if a clock tick is at every microsecond the bandwidth of the modulated signal is about one megahertz as it takes about one microsecond to go from one state to another. The inverse relationship of the interval between clock ticks to bandwidth is only approximate (as will be seen when software-defined radio is considered in a future chapter).

One important digital modulation method does not fit with the description above. This is Frequency Shift Keying (FSK) modulation where the carrier is set to a particular frequency at each clock tick.

The basic digital modulation formats are shown in Figure 23.6.2. The fundamental characteristic of digital modulation is that there are discrete states, each of which is also known as a symbol, with a symbol defining the value of one or more bits. For example, the states are different frequencies in FSK and different phases in phase shift keying (PSK). With the modulated waveforms shown in Figure 23.6.2 there are only two states, which is the same as saying that there are two symbols, each symbol having one bit of information (either 0 or 1). With multiple states groups of bits can be represented.

In this section many methods of digital modulation are described. The first few methods are binary modulation methods with just two symbols with one symbol indicating that a single bit is '0' and the other symbol indicating that it is a '1'. Then four-state modulation is introduced with four symbols with each symbol indicating the values of two bits. Higher-order modulation schemes can send more than two bits per symbol and thus more bits per second (bits/s) per hertz of bandwidth. There is a limit to the number of symbols as the "distance" between symbols becomes smaller and the effect of noise, interference, and circuit distortion can cause a symbol to be misinterpreted as another. A modulation method that sends more bits per symbol is said to have higher modulation efficiency. This and other metrics that enable modulation methods to be compared are defined in the next subsection.

2.5.1 Modulation Efficiency

With digital modulation, the information being sent is in the form of bits and it is possible to send more than one bit per second in one hertz of bandwidth. This is because in digital modulation there can be several bits per symbol, however the bandwidth of the modulated signal is determined by the rate of change from one state to another, whereas the number of bits per transition depends on the number of states. It is important for the transition to be no faster than required so as to minimize bandwidth.

The ratio of the **bit rate** in bits per second (**bits/s**) to the bandwidth (BW) in hertz is called the **modulation efficiency**, η_c , and has the units of bits per second per hertz (**bits/s/Hz**). The modulation efficiency is also called the channel efficiency, hence the subscript *c* on η_c . The bits here are the gross bits which includes the information bits and bits added for error correction and others added to aid in identifying the signal, and so η_c is a measure of the performance of the modulation method itself. Thus

$$\text{modulation efficiency} = \eta_c = \frac{\text{gross bit rate}}{\text{bandwidth}} \quad (23.6.1)$$

The additional bits added to a bit stream are called coding bits and the process of adding the coding bits is called coding. If coding is used, then the information rate is lower than the gross bit rate transmitted. Thus gross bit rate refers to the bits actually transmitted and information rate (or **information bit rate**) refers to the bit rate of information transmission. The **link spectrum efficiency** is the information bit rate divided by the bandwidth. Often the term “link” is dropped and just **spectrum efficiency** is used (with units of **bits/s/Hz**). Thus

$$\text{link spectrum efficiency} = \frac{\text{information bit rate}}{\text{bandwidth}} \leq \eta_c \quad (23.6.2)$$

Example 23.6.1: Modulation Efficiency

A radio transmits a bit stream of **2 Mbits/s** using a bandwidth of **1 MHz**.

- What is the modulation efficiency?
- If **25%** of the bits are used for error correction, what is the modulation efficiency?
- With error correction coding, what is the information rate?
- With error correction coding, what is the link spectrum efficiency?

Solution

- a. The gross bit rate is **2 Mbits/s** and the bandwidth is **1 MHz**. So

$$\eta_c = \text{modulation efficiency} = \frac{\text{gross bit rate}}{\text{bandwidth}} = \frac{2 \text{ Mbits/s}}{1 \text{ MHz}} = 2 \text{ bits/s/Hz}$$

- b. The modulation efficiency is unaffected by error correction coding. So the modulation efficiency is unchanged:

$$\eta_c = \text{modulation efficiency} = \frac{\text{gross bit rate}}{\text{bandwidth}} = \frac{2 \text{ Mbits/s}}{1 \text{ MHz}} = 2 \text{ bits/s/Hz}$$

- c. With **25%** of the bits in the gross bit stream being coding bits, the information rate is **75%** of **2 Mbits/s** or **1.5 Mbits/s**.

$$\text{link spectrum efficiency} = \frac{\text{information bit rate}}{\text{bandwidth}} = \frac{1.5 \text{ Mbits/s}}{1 \text{ MHz}} = 1.5 \text{ bits/s/Hz}$$

This page titled [23.6: Digital Modulation](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.5: Digital Modulation](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.7: Frequency Shift Keying, FSK

Frequency shift keying (FSK) is one of the simplest forms of digital modulation, with the frequency of the transmitted signal at a clock tick indicating a symbol, usually representing either one or two bits. Binary FSK (BFSK) is illustrated in Figure 2.5.2(d). It can be implemented by applying a discrete signal to the input of a voltage-controlled oscillator and so was ideally suited to early digital radio as simple high-performance FM modulators were available. Four-state FSK modulation is used in the GSM 2G cellular standard, a legacy standard still widely supported by modern cellular radios and sometimes the only modulation supported by the infrastructure (i.e. basestations) in some regions where it is not economically viable to retrofit old installations.

2.6.1 Essentials of FSK Modulation

The schematic of a binary FSK modulation system is shown in Figure 23.7.1. Here, a binary bitstream is lowpass filtered and used to drive an FSK modulator, one implementation of which shifts the frequency of an oscillator according to the voltage of the baseband signal. This function can be achieved using a VCO or a PLL circuit, and an FM demodulator can be used to receive the signal. Another characteristic feature of FSK is that the amplitude of the modulated signal is constant, so efficient saturating (and hence nonlinear) amplifiers can be used without the concern of frequency distortion. Not surprisingly, FSK was the first form of digital modulation used in mobile digital radio. A particular implementation of FSK is **Minimum Shift Keying (MSK)**, which uses a baseband lowpass filter so that the transitions from one state to another are smooth in time and limit the bandwidth of the modulated signal.

The **constellation diagram** is often thought of as being like a phasor diagram and this analogy works most of the time but it does not work for FSK modulation. A phasor diagram describes a phasor that is fixed in frequency. If the phasor is very slowly phase and/or amplitude modulated, then this approximation is good. FSK modulation cannot be represented on a phasor diagram, as the information is in the frequency at the clock ticks and not the than the phase and/or amplitude of a phasor. The symbols of two-and four-state FSK modulation are shown in Figure 23.7.2 which are called constellation diagrams.

As an example consider an FSK-modulated signal with a bandwidth of **200 kHz** and a carrier at **1 GHz** (this approximately corresponds to the 2G GSM cellular system). This is a **0.02%** bandwidth, so the phasor changes very slowly. Going from one FSK state to another takes about **1230** to **3692**

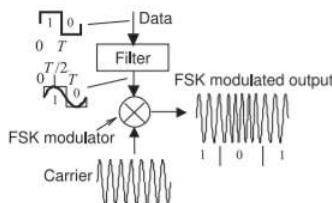


Figure 23.7.1: The frequency shift keying (FSK) modulation system. In the GSM four-state cellular system-adjacent constellation points differ in frequency by **33.25 kHz**.

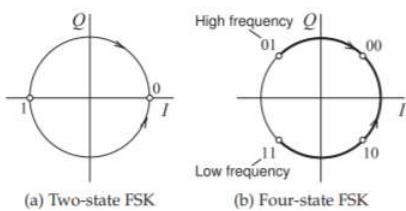


Figure 23.7.2: Constellation diagrams of FSK modulation. In two-state FSK a symbol indicates whether a bit is a '0' or a '1'. In four-state FSK there are four symbols and each symbol has a different frequency and indicates the values of two bits.

RF cycles depending on the frequency difference of the transition from one symbol to the next. With a **1 GHz** carrier the frequencies of the four symbols are (**1 GHz - 33.25 kHz**), (**1 GHz - 16.62 kHz**), (**1 GHz + 16.62 kHz**), and (**1 GHz + 33.25 kHz**). This may seem like a very small frequency difference but hardware in the basestation and in the handset can easily achieve a frequency resolution of a few hertz at **1 GHz**. In trying to represent FSK modulation on a pseudo-phasor diagram, the frequency is

approximated as being fixed and the maximum real frequency shift is arbitrarily taken as being a significant shift of the pseudo-phasor.

In FSK, the states are on a circle in the constellation diagram (see Figure 23.7.2), with two-state FSK shown in Figure 23.7.2(a) and four-state FSK shown in Figure 23.7.2(b). Note that the constellation diagram indicates that the amplitude of the phasor is constant, as FSK modulation is a form of FM modulation. Consider four-state FSK more closely. There are four frequency states ranging from the low-frequency state to the high-frequency state as shown in Figure 23.7.2(b). In four-state FSK modulation a transition from the low-frequency state to the high-frequency state takes three times longer than a transition between adjacent states. While the ‘01’ and ‘11’ states appear to be adjacent, in reality the frequency transition must traverse through the other frequency states. Filtering of the baseband modulating signal is required to minimize the bandwidth of the modulated four-state FSK signal. This reduces modulation efficiency to less than the theoretical maximum of **2 bits/s/Hz**.

In summary, there are slight inconsistencies and arbitrariness in using a phasor diagram for FSK, but FSK does have a defined constellation diagram that is closely related, but not identical, to a phasor diagram. Another difference is that a phasor diagram depends on the amplitude of the RF signal, while a constellation diagram is continuously being re-normalized to the average RF power level to maintain a constant size. With FSK modulation almost the entire modulation and demodulation paths can be implemented using analog circuitry and so was ideally suited to early cellular radios.

2.6.2 Gaussian Minimum Shift Keying

Gaussian minimum shift keying (GMSK) is the modulation scheme used in the GSM cellular wireless system and is a variant of **MSK** with waveform shaping coming from a Gaussian lowpass filter. It is a particular implementation of FSK modulation.

The modulation efficiency of GMSK as implemented in the GSM system (it depends slightly on the Gaussian filter parameters) is **1.35 bits/s/Hz**. Unfiltered MSK has a constant RF envelope. However filtering is required to limit the RF bandwidth and this results in amplitude variations of about **30%**. This is still very little so one of the fundamental advantages of this modulation scheme is that nonlinear, power-efficient amplification can be used. GMSK is essentially a digital implementation of FM with discrete changes in the frequency of modulation with the input bitstream filtered so that the change in frequency from one state to the next is smooth. It is only at the clock ticks that the modulated signal must have the specified discrete frequency. The phase of the modulating signal is always continuous and there is no information in the phase of the modulated signal.

The ideal transitions in FSK follow a circle from one state to another as shown in Figure 23.7.2 so that the PMEPR of ideal FSK is **0 dB**. With GMSK the transitions do not follow a circle because of the filtering and the transitions also overshoot. As such the amplitude of a GMSK modulated signal varies and the PMEPR of GMSK is **3.01 dB**. This is the PMEPR for a single modulated carrier, combining multiple modulated carriers as done in a basestation increases the PMEPR. Statistically the envelopes are less likely to all align if there are multiple carriers. For example, with multi-carrier **GMSK, PMEPR = 3.01 dB, 6.02 dB, 9.01 dB, 11.40 dB, 14.26 dB, and 17.39 dB** for **1, 2, 4, 8, 16, and 32** carriers respectively. (These values were calculated numerically by simulating a multi-carrier system.)

GMSK and other FSK methods have the advantage that implementation of the baseband and RF hardware is relatively simple. A GMSK transmitter can use conventional frequency modulation. On receive, an FM discriminator, i.e. an FM receiver with sampling, can be used avoiding more complex **I** and **Q** demodulation.

2.6.3 Doppler Effect

Frequency is a physical parameter that can be established and measured with great accuracy, down to a few hertz at **1 GHz** in a mobile handset for example. Thus if a receiver is stationary the frequency states at the clock ticks of an FSK modulated carrier can be measured with great accuracy. When a receiver and transmitter are moving relative to each other there will be a Doppler shift of the carrier frequency. If the relative velocity of the receiver and transmitter is v_s , the Doppler shift will be

$$\Delta f = fv_s/c \quad (23.7.1)$$

where f is the frequency of the radio transmission and c is the speed of light. For a receiver moving at **100 km/hr** receiving a **1 GHz** signal from a fixed transmitter, the Doppler frequency shift is $\Delta f = 92.6 \text{ Hz}$ which is much less than the **33 kHz** frequency spacing of adjacent states in the FSK example above. Thus the Doppler shift is not of concern. This effective fixing of the constellation points is one of the advantages of GSM.

2.6.4 Summary

GSM was not the only 2G system. The 2G NADC (for North American Digital Cellular) system modulated the phase of a carrier using phase shift keying. The NADC cellular system had higher modulation efficiency than GSM yet MSK became the dominant 2G system and is still supported as a legacy modulation system in modern cellular radio. The main reason for this is that GSM was more closely aligned with the business interests of the telephone operators of the day.

This page titled [23.7: Frequency Shift Keying, FSK](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.6: Frequency Shift Keying, FSK](#) by Michael Steer is licensed [CC BY-NC 4.0](#).

23.8: Carrier Recovery

Carrier recovery refers to establishing a local carrier reference signal which accurately reproduces the frequency and, with some modulation methods, the phase of the carrier of the modulated signal. All digital modulation methods require carrier recovery to establish a reference to determine the state of the carrier at the clock ticks. In addition digital modulation methods require that the timing of the clock ticks be established. Since radios using digital modulation all send packets of data, i.e. sequences of symbols, having a known sequence at the beginning of packet transmission enables the timing to be determined.

With FSK modulation the frequency at the clock ticks must be determined. This is relatively simple because the frequency at the clock ticks can be accurately measured as a local clock can be established within a few hertz because of the availability of accurate crystal references. The frequency of the received signal can still be shifted by the Doppler effect of the transmitter or receiver is moving but this is quite small compared to the frequency differences between the received states. With FSK it is not necessary to determine the phase of the carrier.

All digital modulation other than FSK modulates a carrier by shifting the carrier's phase and/or amplitude to a number of discrete states. Recovering the state of this modulated carrier requires that the phase of the carrier be recovered from the receive signal and to do this there must be a constant phase local version of the carrier. The circuits that implement the local version of the carrier are called carrier recovery circuits. These circuits modify a very stable internal oscillator in the receiver that after an initial setting of an approximate frequency, has a frequency and phase that can only change slowly. However, there must be a received signal at all times, because if the received signal falls below the noise level the carrier recovery circuit will try to track the noise. This requirement has led to a number of different modulation schemes that avoid the amplitude of the modulated signal from ever being small during a transition. This is important in 2G and 3G cellular radio but 4G and 5G cellular systems use pilot tones to achieve carrier recovery.

In early digital radios carrier recovery was implemented in analog circuitry and more modern radios implement carrier recovery by splitting the function between an analog oscillator signal that can be assigned to a large number of discrete states (providing coarse carrier recovery) and DSP of the (coarsely recovered) baseband signal to precisely recover the carrier signal. Thus in modern digital radios the carrier recovery circuit is implemented partially as an analog circuit and partially as a digital circuit.

This page titled [23.8: Carrier Recovery](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.12: Carrier Recovery](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.9: Phase Shift Keying Modulation

There are many variations on phase shift keying (**PSK**) modulation with the methods differing by their spectral efficiencies, PMEPR, and suitability for carrier recovery. Compared to FSK more sophisticated digital signal processing is required to demodulate a PSK-modulated signal.

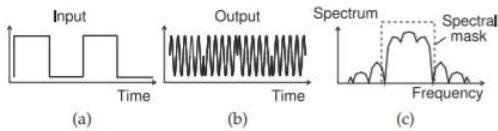


Figure 23.9.1: Binary PSK modulation: (a) modulating bitstream; (b) the modulated waveform; and (c) its spectrum after smoothing the transitions from one phase state to another.

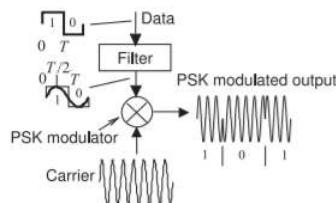


Figure 23.9.2: A binary phase shift keying (PSK) modulation system.

2.8.1 Essentials of PSK

PSK is an efficient digital modulation scheme and can be simply implemented and demodulated using a phase-locked loop. The simplest scheme is binary PSK (BPSK) with two phase states. The waveform and spectrum of BPSK are shown in Figure 23.9.2. The incoming baseband bitstream shown in Figure 23.9.1(a) modulates the phase of a carrier producing the modulated signal shown in Figure 23.9.1(b). The spectrum of the modulated signal is shown in Figure 23.9.1(c). What is very interesting about this spectrum is that it approximately fills a square. So PSK modulation results in an efficient use of the spectrum. This can be contrasted with the spectrum of an FM signal shown in Figure 2.4.5(c), which does not fill the channel uniformly. A binary PSK modulation system is shown in Figure 23.9.2 where the binary input data causes 180° phase changes of the carrier. The abrupt changes in phase shown in the output modulated waveform result in more bandwidth than is necessary. However a practical PSK modulator first lowpass filters the binary data before the carrier is modulated. This filtering eliminates the abrupt changes in the phase of the modulated signal and so reduces the required bandwidth. It is the spectrum of this signal that is shown in Figure 23.9.1(c).

There are many variants of increasing complexity, called orders, of PSK, with the fundamental characteristics being the number of phase states (e.g. with 2^n phase states, n bits of information can be transmitted) and how the phasor of the RF signal transitions from one phase state to another. PSK schemes are designed to shape the spectrum of the modulated signal to fit as much energy as possible within a spectral mask. This results in a modulated carrier whose amplitude varies (and thus has a time-varying envelope). Such schemes require highly linear amplifiers to preserve the amplitude variations of the modulated RF signal.

There are PSK methods that manage the phase transitions to achieve a constant envelope modulated RF signal but these have lower spectral efficiency. Military radios sometimes use this type of modulation scheme as it is much harder to detect and intercept communications if the amplitude of the modulated carrier is constant.

The communication limit of one symbol per hertz of bandwidth,

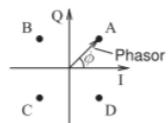


Figure 23.9.3: Phasor diagram of QPSK modulation. Here there are four discrete phase states of the phasor indicated by the points A, B, C, and D. The PSK modulator moves the phasor from one phase state to another. The task at the receiver is determining the phase of the phasor.

the **symbol rate**, comes from the **Nyquist signaling theorem**.¹ Nyquist determined that the number of independent pulses that could be put through a telegraph channel per unit of time is limited to twice the bandwidth of the channel. With a modulated RF carrier, this translates to the modulated carrier moving from one state to another in a unit of time equal to one over the bandwidth. The transition identifies a symbol, and hence one symbol can be sent per hertz of bandwidth. More accurately it could be said that the transition is a symbol rather than the end of the transition being a symbol. In PSK modulation the states are the phases of a phasor since the amplitude of the modulated signal is (ideally) constant.

The phase-shifted (i.e. phase-modulated) carrier of a PSK signal can be represented on a phasor diagram. Figure 23.9.3 is a phasor diagram with four phase states—A, B, C, D—and the phasor moves from one state to another under the control of the modulation circuit. What is shown here is 4-state PSK or quadra-phase shift keying (QPSK) and very often but less accurately called **quadrature phase shift keying**. The states, or symbols, are identified by their angle or equivalently by their rectangular coordinates, called I, for in-phase, and Q, for quadrature phase.

PSK Modulation

In PSK modulation the phase of a carrier signal is set to one of a number of discrete values at the clock ticks. For example, in **QPSK** there are four discrete settings of the phase of the carrier, e.g. 45° , 135° , -135° , and -45° . Converting this to radians the discrete baseband signal is $\phi(t) = \pi/4$, $3\pi/4$, $5\pi/4$, and $7\pi/4$, at the clock ticks. Thus if the bandwidth of the baseband signal is **1 MHz** what is shown as $\phi(t)$ are the intended phases of the carrier every microsecond. Wave-shaping or filtering is used to provide a smooth variation of $\phi(t)$ between the clock ticks and so the bandwidth of the modulated signal is constrained. High-order PSK modulation has more discrete states, e.g. 8-PSK has eight discrete phase states.

There are several ways to implement PSK modulation and one uses the quadrature modulator shown in Figure 23.9.4. The discrete baseband signal $\phi(t)$ could be internal to a DSP which is then interpolated in time and output by the DSP's DAC as two smooth signals $i(t) = \cos(\phi(t))$ and $q(t) = \sin(\phi(t))$. On a phasor diagram $i(t)$ and $q(t)$ at the clock ticks addresses one of QPSK's four states of the carrier's phasor, see Figure 23.9.3.

For PSK modulation the constellation diagram is very similar to a phasor diagram that is continuously being re-normalized to the average power of

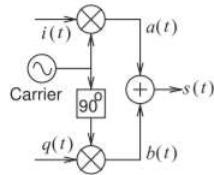


Figure 23.9.4: Quadrature modulator block diagram. In PSK modulation $i(t)$ and $q(t)$ have the same amplitude and indicate a phase ϕ of the modulated carrier so that $i(t) = \cos[\phi(t)]$ and $q(t) = \sin[\phi(t)]$. The particular example shows two possible values of I_k and Q_k and this indicates QPSK modulation.

the modulated signal. This a subtle but important distinction, for example, a PSK baseband signal has a constellation diagram even though the baseband signal does not have a phasor representation. PSK modulation using the block diagram shown in Figure 23.9.4 has a carrier that is directly input to the top multiplier and a 90° phase-shifted version input to the bottom multiplier. Let the carrier be $\cos(\omega_c t)$ and so the version of the carrier input to the bottom multiplier is $\cos(\omega_c t - \pi/2) = -\sin(\omega_c t)$. So, with $q(t)$ being a 90° phase-shifted version of $i(t)$, (using the identities in Section 1.A.2 of [4]).

$$a(t) = \cos(\phi(t)) \cos(\omega_c t) = \frac{1}{2} [\cos(\omega_c t - \phi(t)) + \cos(\omega_c t + \phi(t))] \quad (23.9.1)$$

$$b(t) = \sin(\phi(t)) [-\sin(\omega_c t)] = -\frac{1}{2} [\cos(\omega_c t - \phi(t)) - \cos(\omega_c t + \phi(t))] \quad (23.9.2)$$

$$s(t) = a(t) + b(t) = \cos(\omega_c t + \phi(t)) \quad (23.9.3)$$

Thus $s(t)$ is the single-sideband modulated carrier carrying information in the phase of the modulated carrier. The modulating signal $\phi(t)$ is driven by a digital code that is designed so that $\phi(t)$ changes at a minimum rate (it never has the same value for more than a few clock ticks). Thus there are no low frequency components of $\phi(t)$ and thus there is no modulated signal at or very close

to the carrier. Thus the carrier is suppressed but there is a sideband above and below the carrier frequency. This is SSB-SC modulation.

2.8.2 Binary Phase Shift Keying

PSK uses prescribed phase shifts to define symbols, each of which can represent one, two, or more bits. **Binary Phase Shift Keying** (BPSK), illustrated in Figures 23.9.1 and 23.9.2, has two phase states and conveys one bit per symbol and is a relatively spectrally inefficient scheme, with a maximum (i.e. ideal) modulation efficiency of **1 bits/s/Hz**. The reason why the practical modulation efficiency is less than this number is because the transition from one phase state to the other must be constrained to avoid the modulated signal becoming very small, and also because there are no ideal lowpass filters to filter the input binary data stream. Although it has low modulation efficiency, it is ideally suited to low-power applications. BPSK is commonly used in **Bluetooth**.

The operation of BPSK modulation can be described using the constellation diagram shown in Figure 23.9.5(a). The BPSK constellation diagram indicates that there are two states. These states can be interpreted as the rms values of $i(t)$ and $q(t)$ at the sampling times corresponding to the bit rate. The distance of a constellation point from the origin corresponds to (normalized) rms power of the pseudo-sinusoid of the modulated carrier at the sampling instant. (Normalization is with respect to the average power.) The curves in Figure 23.9.5(b) indicate three transitions. The states are at the ends of the transitions. If a 1, in Figure 23.9.5(b), is assigned to the positive I value and 0

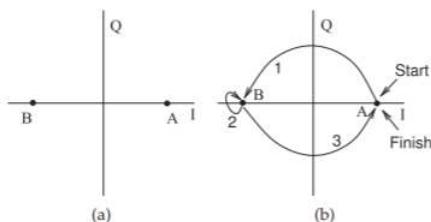


Figure 23.9.5: BPSK modulation with constellation points A and B: (a) constellation diagram; and (b) constellation diagram with possible transitions from one phase state to the other, or possibly no change in the phase state. In practical systems the transition should not go through the origin, as then the RF signal would drop below the noise level.

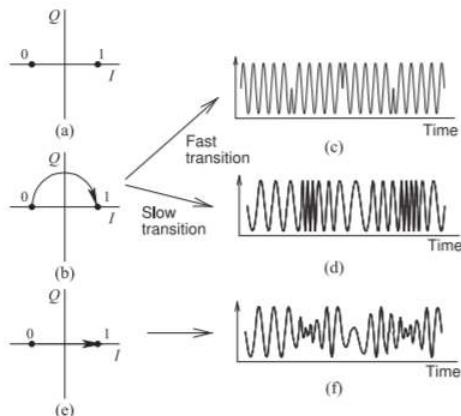


Figure 23.9.6: BPSK modulation: (a) constellation diagram; (b) constellation diagram with a constant amplitude transition; (c) time-domain waveform if the transition is fast; (d) time-domain waveform if the transition is slow; (e) constellation diagram with transition through the origin; and (f) time-domain waveform if the transition goes through the origin and is slow.

to a negative I value, then the bit sequence represented in Figure 23.9.5(b) is “1001.”

Figure 23.9.6(a) is the constellation diagram of BPSK, with two symbols denoted as 0 and 1, and the trajectory of the transition from one constellation point to the other depending on the hardware used to implement the BPSK modulator. Figure 23.9.6(b) shows the transition from the ‘0’ state to the ‘1’ state (and back) while maintaining a constant amplitude. If this transition is very fast, then the waveform produced is as shown in Figure 23.9.6(c), where there are abrupt phase transitions and these have high

spectral content. It is better to slow down the transitions, as then the waveform (shown in Figure 23.9.6(d)), has smooth transitions and the bandwidth of the modulated carrier is minimal. The preferred smooth transition is obtained by lowpass filtering the baseband signal. That is, the abrupt transitions in the modulated RF signal result in the modulated signal having a broad bandwidth. The graceful transition of BPSK modulation limits the bandwidth of the modulated carrier.

A simple implementation of BPSK modulation would result in direct transition from one state to the others causing the phasor to traverse the origin and the amplitude of the RF signal to become very small and less than the noise level (see Figure 23.9.6(e)). The resulting modulated RF waveform is

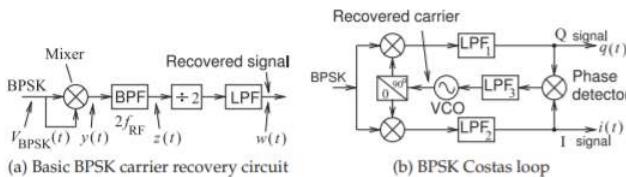


Figure 23.9.7: Block diagram of carrier recovery circuits for BPSK signals.

shown in Figure 23.9.6(f). This is a problem because the receiver would not be able to track the RF signal.

Carrier Recovery

In a PSK demodulator, a local copy of the carrier must be produced to act as a reference in determining the phase of the modulated signal. The technique that produces the local copy of the unmodulated carrier is called carrier recovery. The circuit that directly implements carrier recovery of a BPSK signal is shown in Figure 23.9.7(a). At the clock ticks the waveform of a BPSK modulated signal is

$$v_{\text{BPSK}}(t) = A(t) \cos(\omega_{\text{RF}}t + n\pi) \quad (23.9.4)$$

where the carrier frequency $f_{\text{RF}} = \omega_{\text{RF}}/(2\pi)$ and n can have a value of 0 or 1. Squaring this produces a signal

$$\begin{aligned} y(t) &= v_{\text{BPSK}}^2(t) = A^2(t) \cos^2(\omega_{\text{RF}}t + n\pi) = \frac{1}{2} A^2(t) [1 + \cos(2\omega_{\text{RF}}t + n2\pi)] \\ &= \frac{1}{2} A^2(t) [1 + \cos(2\omega_{\text{RF}}t)] \end{aligned} \quad (23.9.5)$$

This is a signal at twice the carrier frequency with no carrier modulation since $n2\pi$ and 0 radians are indistinguishable. The squaring operation is performed by mixing $v_{\text{BPSK}}(t)$ with itself. Bandpass filtering $y(t)$ produces a signal $z(t)$ at the second harmonic of the carrier. The divide-by-2 block is implemented using a phase-locked loop (PLL). The result is the recovered carrier, $w(t)$, that is used as the timing reference for sampling the demodulated I and Q components at precise times.

A better carrier recovery circuit than that in Figure 23.9.7(a) and described above is the Costas loop [14] shown in Figure 23.9.7(b). The BPSK Costas loop implements carrier recovery and I/Q demodulation simultaneously. In Figure 23.9.7(b) $i(t)$ and $q(t)$ are mixed to produce a signal applied at the input of the third lowpass filter, LPF_3 . The main function of this filter is to remove noise and to average the signal coming out of the phase detector. The output of LPF_3 drives a VCO in which the oscillation frequency is controlled by the applied voltage. The quadrature phase shifter then mixes the recovered carrier and a 90° shifted version with the BPSK signal.

It is critical that the signal-to-noise ratio (SNR), the ratio of the signal power to the noise power, of the BPSK signal be sufficiently large at all times or else the Costas loop will produce a noisy recovered carrier signal. If the modulated carrier becomes very small, for example when the trajectory on the constellation diagram goes through the origin (where the level of the carrier carrier falls below the noise level), the carrier will not be accurately recovered.

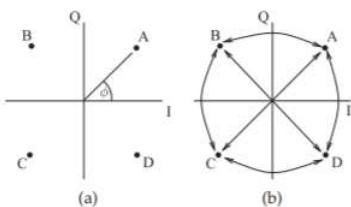


Figure 23.9.8: QPSK modulation: (a) constellation diagram; and (b) constellation diagram with possible transitions. Each constellation point indicates the phase, ϕ , of the modulated carrier, i.e. $\cos(\omega_c t + \phi)$ where ω_c is the radian frequency of the carrier.

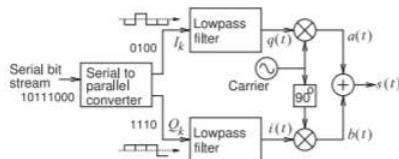


Figure 23.9.9: QPSK modulator block diagram. I_k and Q_k are similar to a stream of one-bit binary signals but are analog with a either a positive value or a negative voltage so that after lowpass filtering $i(t)$ and $q(t)$ each have either a positive or a negative voltage at each clock tick.

2.8.3 Quadra-Phase Shift Keying, QPSK

In QPSK wireless systems, modulation efficiency is obtained by sending more than one bit of information per hertz of bandwidth (i.e., more than one bit per symbol). In QPSK information is encoded in four phase states and two bits are required to identify a symbol (i.e., to identify a phase state). The constellation diagram of QPSK is shown in Figure 23.9.8(a) where the modulated RF carrier has four phase states identified as A, B, C, and D. So a QPSK modulator shifts the phase of the carrier to one of these phase states, and a QPSK demodulator must determine the phase of the received RF signal. The received RF signal is sampled with precise timing as determined by the recovered carrier signal. Thus two bits of information are transmitted per change of phase states. Each change of phase state requires at least 1 Hz of bandwidth with the minimum bandwidth obtained when the transition from one state to another is no faster than that required to reach the new phase state before the sampling instant. QPSK modulation is also referred to as quarternary PSK.

QPSK can be implemented using the modulator shown in Figure 23.9.9. In Figure 23.9.9, the input bitstream is first converted into two parallel bitstreams each containing half the number of bits of the original bit stream. Thus a two-bit sequence in the serial bitstream becomes one I_K bit and one Q_K bit. The (I_K, Q_K) pair constitutes the K th symbol. The bitstreams are converted into waveforms $i(t)$ and $q(t)$ by the wave-shaping circuit.

The constellation diagram of QPSK is the result of plotting I and Q on a rectangular graph as shown in Figure 23.9.8(a). All possible phase transitions are shown in Figure 23.9.8(b). In the absence of wave-shaping circuits, $i(t)$ and $q(t)$ have very sharp transitions, and the paths shown in Figure 23.9.5(b) occur almost instantaneously. This leads to large spectral spreads in the modulated waveform, $s(t)$. So to limit the spectrum of the RF signal $s(t)$, the shape of $i(t)$ and $q(t)$ is controlled; the waveform is shaped, usually by lowpass filtering. So a pulse-shaping circuit changes baseband binary information into a more smoothly varying signal. Each transition or path in Figure 23.9.5 represents the transfer of a symbol, with the best efficiency that can be obtained in wireless communication being one symbol per hertz of bandwidth. Each symbol contains two bits so the maximum modulation efficiency of QPSK modulation is 2 bits/s/Hz of bandwidth. What is actually achieved depends on the wave-shaping circuits and on the criteria used to establish the bandwidth of $s(t)$.

Carrier Recovery

Carrier recovery of a QPSK signal is similar to that for a BPSK signal. At the clock ticks an RF QPSK modulated signal

$$v_{\text{QPSK}}(t) = A(t) \cos(\omega_{\text{RF}} t + n\pi/2); \quad n = 0, 1, 2, 3, \quad (23.9.6)$$

where the carrier frequency $f_{\text{RF}} = \omega_{\text{RF}}/(2\pi)$. The fourth power of this produces

$$\begin{aligned} v_{\text{QPSK}}^4(t) &= A^4(t) \cos^4(\omega_{\text{RF}} t + n\pi/2) \\ &= \frac{1}{8} A^4(t) [3 + 4 \cos(2\omega_{\text{RF}} t + n\pi) + \cos(4\omega_{\text{RF}} t + n2\pi)] \end{aligned} \quad (23.9.7)$$

Following bandpass filtering at $4f_{\text{RF}}$ and then dividing the frequency by 4, the carrier is recovered. Circuits implementing this are similar to those for recovering the carrier of BPSK signals. This concept can be extended to carrier recovery for any M -PSK-modulated signal.

Example 23.9.1: QPSK Modulation and Constellation

The bit sequence **110101001100** is to be transmitted using QPSK modulation. Show the transitions on a constellation diagram.

Solution

The bit sequence **110101001100** must be converted to a two-bit-wide parallel stream of symbols resulting in the sequence of symbols **110101001100**. The symbol **11** transitions to the symbol **01** and then to the symbol **01** and so on. The states (or symbols) and the transitions from one symbol to the next required to send the bitstream **110101001100** are shown in Figure 23.9.10. QPSK modulation results in the phasor of the carrier transitioning through the origin so that the average power is lower and the PMEPR is high. A more significant problem is that the phasor will fall below the noise floor, making carrier recovery almost impossible.

2.8.4 $\pi/4$ Quadrature Phase Shift Keying

A major objective in digital modulation is to ensure that the RF trajectory from one phase state to another does not go through the origin. The transition is slow, so that if the trajectory goes through the origin, the amplitude of the carrier will be below the noise floor for a considerable time and it will not be possible to recover the carrier reference. This is why the QPSK scheme is not used directly in 2G and 3G cellular radio. (The 4G and 5G cellular radio systems do use QPSK among other modulation schemes and use pilot tones to recover the carrier.) One of the solutions developed to address this problem is the $\pi/4$ quadrature phase shift keying ($\pi/4$ -QPSK) modulation scheme. In this scheme the constellation at each symbol is rotated $\pi/4$ radians from the previous symbol, as shown in Figure 23.9.11. (In an alternative implementation of $\pi/4$ -QPSK modulation the constellation diagram could

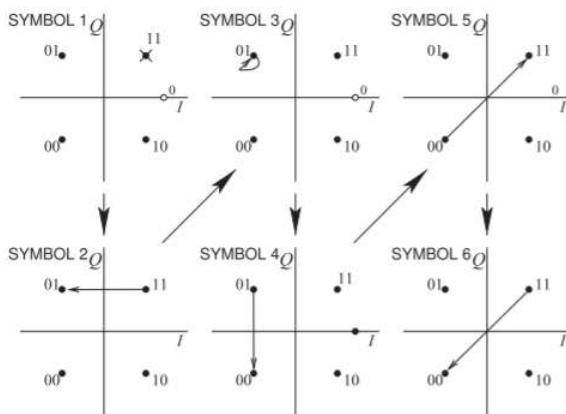


Figure 23.9.10: Constellation diagram and transitions for the bit sequence **110101001100** sent as the set of symbols **110101001100** using QPSK. Note that symbols **2** and **3** are identical, so there is no transition. The **SYMBOL** numbers indicated reference the symbol at the end of the transition (end of the arrow). The assignment of bits to symbols (e.g., assigning the bits ‘**11**’ to the symbol in the first quadrant) is arbitrary in general but the assignment of symbols is defined in a particular standard.

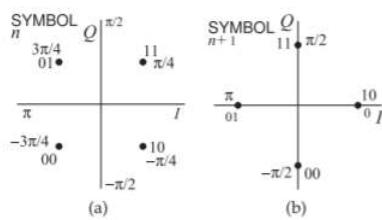


Figure 23.9.11: Constellation diagram of $\pi/4$ -QPSK modulation: (a) initial constellation diagram at one symbol; and (b) the constellation diagram at the time of the next symbol.

rotate by $\pi/4$ continuously rather than switching between conditions as described here.)

One of the unique characteristics of $\pi/4$ -QPSK modulation is that there is always a change, even if a symbol is repeated. This helps with recovering the carrier frequency. If the binary bitstream itself (with sharp transitions in time) is the modulation signal, then the

transition from one symbol to the next occurs instantaneously and hence the modulated signal has a broad spectrum around the carrier frequency. The transition, however, is slower if the bitstream is filtered, and so the bandwidth of the modulated signal will be less. Ideally the transmission of one symbol per hertz would be obtained. However, in $\pi/4$ -QPSK modulation the change from one symbol to the next has a variable distance (and so a transition takes different times) so that the ideal modulation efficiency of one symbol per second per hertz (or 2

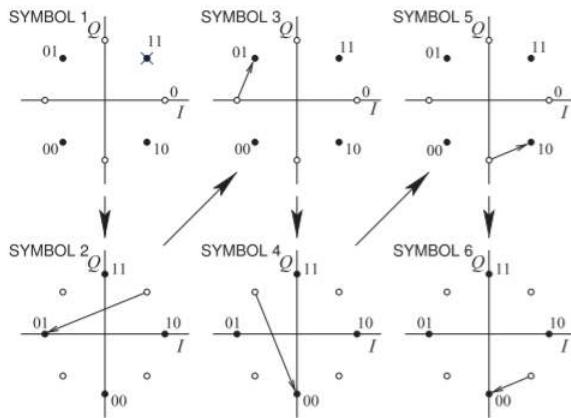


Figure 23.9.12: Constellation diagram states and transitions for the bit sequence **110101001000** sent as the set of symbols **110101001000** using $\pi/4$ -QPSK modulation.

bits/s/Hz) is not obtained. In practice, with realistic filters and allowing for the longer transitions, $\pi/4$ -QPSK modulation achieves **1.62 bits/s/Hz**.

Example 23.9.2: $\pi/4$ -QPSK Modulation and Constellation

The bit sequence **110101001000** is transmitted using $\pi/4$ -QPSK modulation. Show the transitions on a constellation diagram.

Solution

The bit sequence **110101001000** must be converted to a two-bit-wide parallel stream of symbols, resulting in the sequence of symbols **110101001000**. The symbol **11** transitions to the symbol **01** and then to the symbol **01** and so on. The constellation diagram of $\pi/4$ -QPSK modulation really consists of two QPSK constellation diagrams that are shifted by $\pi/4$ radians, as shown in Figure 23.9.11. At one symbol (or time) the constellation diagram is that shown in Figure 23.9.11(a) and at the next symbol it is that shown in Figure 23.9.11(b). The next symbol uses the constellation diagram of Figure 23.9.11(a) and the process repeats. The states (or symbols) and the transitions from one symbol to the next that are required to send the bitstream **110101001000** are shown in Figure 23.9.12.

2.8.5 Differential Quadra Phase Shift Keying, DQPSK

Multiple transmission paths, or **multipaths**, due to reflections result in constructive and destructive interference and can result in rapid additional phase rotations. Thus relying on the phase of a phasor at the symbol sample time, at the clock ticks, to determine the symbol transmitted is prone to error. When an error results at one symbol, this error accumulates when subsequent symbols are extracted. The solution is to use encoding, and one of the simplest encoding schemes is differential phase encoding. In this scheme the information of the modulated signal is contained in changes in phase rather than in the absolute phase. That is, the transition defines the symbol rather than the end point of the transition.

The $\pi/4$ -DQPSK modulation scheme is a differentially encoded form of

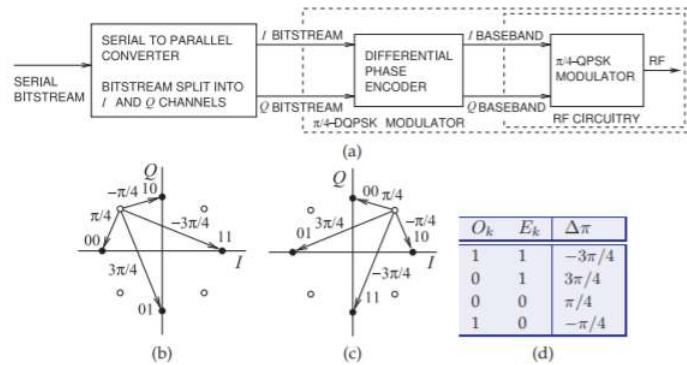


Figure 23.9.13: A $\pi/4$ -DQPSK modulator: (a) differential phase encoder with a $\pi/4$ -QPSK modulator; (b) constellation diagram of $\pi/4$ -DQPSK; (c) a second constellation diagram; and (d) phase changes in a $\pi/4$ -DQPSK modulation scheme. Note that the information is in the phase change rather than the phase state.

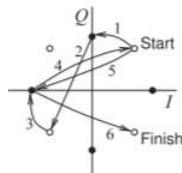


Figure 23.9.14: Constellation diagram of $\pi/4$ - DQPSK modulation showing six symbol intervals coding the bit sequence 000110110101.

$\pi/4$ -QPSK. The $\pi/4$ -DQPSK scheme incorporates the $\pi/4$ -QPSK modulator and an encoding scheme, as shown in Figure 23.9.13(a). The scheme is defined with respect to its constellation diagram, shown in Figure 23.9.14(b) and repeated in Figure 23.9.13(c) for clarity. The D indicates **differential coding**, while the $\pi/4$ denotes the rotation of the constellation by $\pi/4$ radians from one interval to the next. This can be explained by considering Figure 23.9.13(a). A four-bit stream is divided into two quadrature **nibbles** of two bits each. These nibbles independently control the **I** and **Q** encoding, respectively, so that the allowable transitions rotate according to the last transition. The information or data is in the phase transitions rather than the constellation points themselves. The relationship between the symbol value and the transition is given in Figure 23.9.13(d). For example, the transitions shown in Figure 23.9.14 for six successive time intervals describes the input bit sequence 000110110101. Its waveform and spectrum are shown in Figure 23.9.15. More detail of the spectrum is shown in Figure 23.9.16. In practice with realistic filters and allowing for the longer transitions, $\pi/4$ -DQPSK modulation achieves a modulation efficiency of 1.62 bits/s/Hz, the same as $\pi/4$ -QPSK, but of course with greater resilience to changes in the transmission path.

In a differential scheme, the data transmitted are determined by

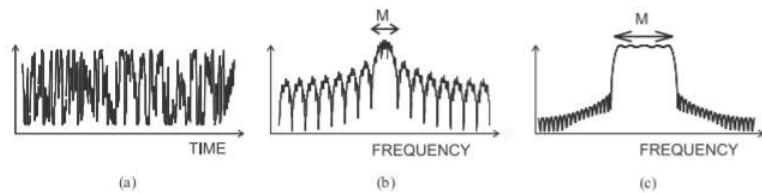


Figure 23.9.15: Details of digital modulation obtained using differential phase shift keying ($\pi/4$ - DQPSK): (a) modulating waveform; (b) spectrum of the modulated carrier, with **M** denoting the main channel; and (c) details of the spectrum of the modulated carrier focusing on the main channel.

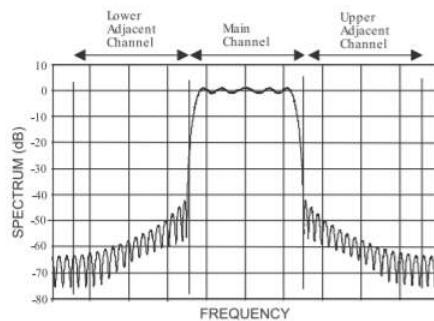


Figure 23.9.16: Detailed spectrum of a $\pi/4$ -DQPSK signal showing the main channel and lower and upper adjacent channels.

comparing a symbol with the previously received symbol, so the data are determined from the change in phase of the carrier rather than the actual phase of the carrier. This process of inferring the data actually sent from the received symbols is called decoding. When $\pi/4$ -DQPSK encoding was introduced in the early 1990s the DSP available for a mobile handset had only just reached sufficient complexity. Today, encoding is used with all digital radio systems and is more sophisticated than just the differential scheme of DQPSK. There are new ways to handle carrier phase ambiguity. The sophistication of modern coding schemes is beyond the scope of the hardware-centric theme of this book.

2.8.6 Offset Quadra Phase Shift Keying, OQPSK

The **offset quadra phase shift keying (OQPSK)** modulation scheme avoids transitions passing through the origin on the constellation diagram (see Figure 23.9.18(a)). As in all QPSK schemes, there are two bits per symbol, but now one bit is used to immediately modulate the RF signal, whereas the other bit is delayed by half a symbol period, as shown in Figure 23.9.17. The maximum phase change for a bit transition is 90° , and as Q_K is delayed, a total phase change of approximately 180° is possible during one symbol. The

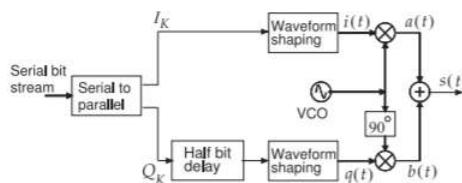


Figure 23.9.17: Block diagram of an OQPSK modulator.

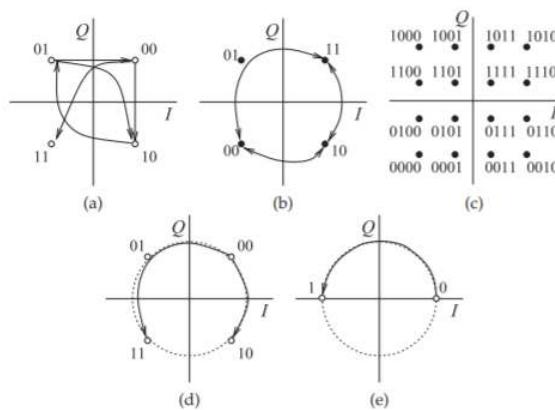


Figure 23.9.18: Constellation diagrams for various modulation formats: (a) OQPSK; (b) GMSK; (c) 16-QAM; (d) SOQPSK (also FOQPSK); and (e) SBPSK.

constellation diagram is shown in Figure 23.9.18(a).

The OQPSK modulator can be implemented using relatively simple electronics with a digital delay circuit delaying the Q bit by half a symbol period and lowpass filters shaping the I and Q bits. The OQPSK scheme is also called **staggered quadrature phase**

shift Keying (SQPSK). Better performance can be obtained by using DSP to shape the I and Q transitions so that they change smoothly and the phasor trajectory nearly follows a circle. Consequently I and Q change together, but in such a manner that the PMEPR is maintained close to 0 dB. Two modulation techniques that implement this are the **shaped offset QPSK (SOQPSK)** and the **Feher QPSK (FQPSK)** schemes. The constellation diagrams for SOQPSK and FQPSK are shown in Figure 23.9.18(d). These are constant envelope digital modulation schemes. As with OQPSK, the Q bit is delayed by one-half of a symbol period and the I and Q baseband signals are shaped by a half-sine filter. The advantage is that high-efficiency saturating amplifier designs can be used and battery life extended. There is a similar modulation format called **shaped binary phase shift keying (SBPSK)** which, as expected, has two constellation points as shown in Figure 23.9.18(e). SOQPSK, FQPSK, and SBPSK are **continuous phase modulation (CPM)** schemes, as the phase

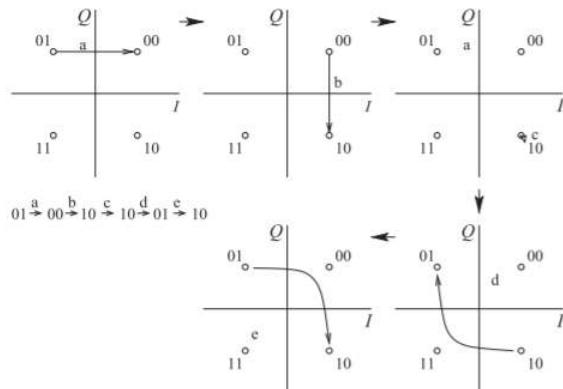


Figure 23.9.19: Constellation diagram of OQPSK modulation for the bit sequence 010010100110.

never changes abruptly. Instead, the phase changes smoothly, achieving high modulation efficiency and maintaining a constant envelope. Implementation of the receiver, however, is complex. CPM schemes have good immunity to interference and are commonly used in military systems.

Example 23.9.3: OQPSK Modulation

Draw the constellation diagrams for the bit sequence 010010100110 using OQPSK modulation.

Solution

The bit sequence is first separated into the parallel stream 01 – 00 – 10 – 10 – 01 – 10. The I bit changes first, followed by the Q bit delayed by half of the time of a bit. Five constellation diagrams are shown in Figure 23.9.19 with the transitions sending the bit sequence.

2.8.7 $3\pi/8$ -8PSK, Rotating Eight-State Phase Shift Keying

The $3\pi/8$ -8PSK modulation scheme is similar to $\pi/4$ -DQPSK in the sense that rotation of the constellation occurs from one time interval to the next. This time, however, the rotation of the constellation from one symbol to the next is $3\pi/8$. This modulation scheme is used in the **enhanced data rates for GSM evolution (EDGE)** system, and provides three bits per symbol (ideally) compared to GMSK used in GSM which has two bits per symbol (ideally). With some other changes, GSM/EDGE provides data transmission of up to **128 kbit/s**, faster than the **48 kbit/s** possible with GSM.

Quadrature modulation schemes with four states, such as QPSK, have two I states and two Q states that can be established by lowpass filtering the I and Q bitstreams. For higher-order modulation schemes such as 8-PSK, this approach will not work. Instead, $i(t)$ and $q(t)$ are established in the DSP unit and then converted using a DAC to generate the analog signals applied

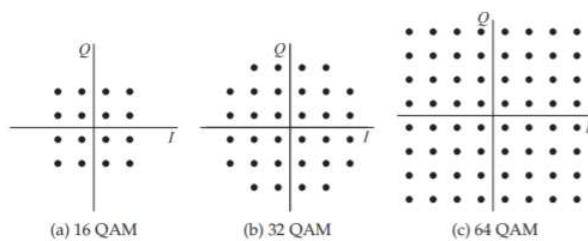


Figure 23.9.20: QAM constellation diagrams.

to the hardware modulator. Alternatively the modulated signal is created directly in the DSP and a DAC converts this to an IF and a hardware mixer up-converts this to RF. QAM

2.8.8 Summary

PSK modulation is implemented in many radio standards including all cellular standards after 2G. There was a 2G system that used $\pi/4$ DQPSK but that is no longer supported. The modern radio standards support many modulation formats but in high interference situations BPSK, QPSK and 8-PSK have the best performance. While QPSK was dismissed in 2G and 3G because of difficulties with carrier recovery, 4G and 5G have another method for implementing carrier recovery which allows QPSK on its own to be used. GMSK is still supported by modern cellular phones but the infrastructure, i.e. basestations, are starting to be retired.

Most of the modulation schemes described in this section were introduced as optimum trade-offs of modulation efficiency, resistance to interference, and hardware complexity. Some, such as BPSK, draw very little power and are suited to the internet-of-things (IoT) applications which must have a battery lifetime of ten years.

Footnotes

[1] This theorem was discovered independently by several people and is also known as the Nyquist-Shannon sampling theorem, the Nyquist-Shannon-Kotelnikov, the Whittaker-Shannon-Kotelnikov, the Whittaker-Nyquist-Kotelnikov-Shannon (WKS), as well as the cardinal theorem of interpolation theory. The theorem states [13]: “If a function $x(t)$ contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced $1/(2B)$ seconds apart.”

This page titled [23.9: Phase Shift Keying Modulation](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.13: Phase Shift Keying Modulation](#) by Michael Steer is licensed [CC BY-NC 4.0](#).

23.10: Quadrature Amplitude Modulation

The digital modulation schemes described so far modulate the phase or frequency of a carrier to convey digital data and the constellation points lie on a circle of constant amplitude. The effect of this is to provide some immunity to amplitude changes to the signal. However, much more information can be transmitted if the amplitude is varied as well as the phase. With considerable signal processing it is possible to reliably use quadrature amplitude modulation (QAM) in which amplitude and phase are both changed.

A **16**-state rectangular QAM, 16-QAM, constellation is shown in Figure 2.8.18(c). Since there are $16 (= 2^4)$ symbols the values of **4** binary bits are uniquely specified by each symbol. In Figure 2.8.18(c) a gray-scale assignment of **4** bit values is shown. Several QAM schemes are shown in Figure 2.8.20. These constellations can be produced by separately amplitude modulating an **I** carrier and a **Q** carrier. Both carriers have the same frequency but are 90° out of phase. The two carriers are then combined so that the fixed carrier is suppressed. The most common form of QAM is square QAM, or rectangular QAM with an equal number of **I** and **Q** states. The most common forms are

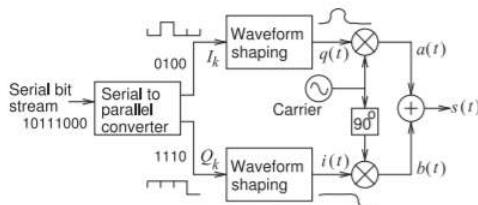


Figure 23.10.1: QAM modulator block diagram. In QAM modulation $i(t)$ and $q(t)$ address the real and imaginary components of a phasor. The wave-shaping block ensures that the symbol has the correct amplitude and phase at each clock tick.

Modulation	bits/s/Hz
BPSK (ideal)	1
BFSK (actual)	1
QPSK (ideal)	2
GMSK (an actual FSK method)	1.354
$\pi/4$ -DQPSK (an actual QPSK method)	1.63
8-PSK (ideal)	3
$3\pi/8$ -8PSK (an actual 8PSK method)	2.7
16-QAM (ideal)	4
16-QAM (actual)	2.98
32-QAM (ideal)	4
32-QAM (actual)	3.35
64-QAM (ideal)	6
64-QAM (actual)	4.47
256-QAM (ideal)	8
256-QAM (actual, satellite & cable TV)	6.33
512-QAM (ideal)	9
1024-QAM (ideal)	10
2048-QAM (ideal)	11

Table 23.10.1: Modulation efficiencies of various modulation formats in bits/s/Hz (bits per second per hertz). The maximum (or ideal) modulation efficiencies obtained by modulation schemes (e.g., BPSK, BFSK, 64- QAM, 256-QAM) result in broad spectra. Actual modulation efficiencies achieved are less in an effort to manage bandwidth. For example, the values for $\pi/4$ -DQPSK and $3\pi/8$ -8PSK are actual. This reduction from ideal arises since symbol transitions are of different lengths and length corresponds to time durations. Since the symbol interval is fixed, it is the longest path that determines the bandwidth required.

16-QAM, 64-QAM, and 128-QAM, in 4G, and 256-QAM additionally in 5G. The constellation points are closer together with high-order QAM and so are more susceptible to noise and other interference. Thus high-order QAM can deliver more data, but less reliably than lower-order QAM.

The constellation in QAM can be constructed in many ways, and while rectangular QAM is the most common form, non rectangular schemes exist; for example, having two PSK schemes at two different amplitude levels. While there are sometimes minor advantages to such schemes, square QAM is generally preferred as it requires simpler modulation and demodulation.

One possible architecture of a QAM modulator is shown in Figure 23.10.1 and this can only be implemented in DSP since it is not sufficient to use analog lowpass filtering to implement the wave-shaping function as the $i(t)$ and $q(t)$ must be precisely the real and imaginary parts of the symbol at each clock tick.

This page titled [23.10: Quadrature Amplitude Modulation](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.14: Quadrature Amplitude Modulation](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.11: Digital Modulation Summary

The modulation efficiencies of various digital modulation schemes are summarized in Table 2.9.1. For example, in **1 kHz** of bandwidth the **$3\pi/8$ -8PSK** scheme (supported in 3G cellular radio) transmits **2700 bits**.

beta.69: It is critical to control interference in digital radio so that the error in digital transmission is no more than one bit per symbol. Error correction can then be used to provide error-free digital communications.

The modulation efficiency of an actual modulation method is less than the ideal (see Table 2.9.1). With digital modulation wave-shaping at baseband is required to constrain the spectrum of the RF-modulated signal. Thus it will take different times for the phasor to make the transition from one symbol to another; to achieve longer transitions in the same time interval requires more bandwidth than that required for shorter transitions. As a result, the modulation efficiency of modulation methods other than binary methods will be less than the ideal. So in a QPSK-like scheme, **2 bits** per symbol are achievable, but the longest transition takes the most time, so the bandwidth needs to be increased so that the transition is completed in time (i.e., in a fixed time equal to one over the bandwidth). Various modulation methods have relative merits in terms of modulation efficiency, tolerance to fading (due to destructive interference), carrier recovery, spectral spreading in nonlinear circuitry, and many other issues that are the purview of communication system theorists.

This page titled [23.11: Digital Modulation Summary](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.7: Digital Modulation Summary](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.12: References

- [1] “American National Standard T1.523-2001, Telecom Glossary 2011,” available on-line with revisions at <http://glossary.atis.org>, 2011, sponsored by Alliance for Telecommunications Industry Solutions.
- [2] S. Boyd, “Multitone signals with low crest factor,” IEEE Trans. on Circuits and Systems, vol. 33, no. 10, pp. 1018–1022, Oct. 1986.
- [3] A. Jones, T. Wilkinson, and S. Barton, “Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes,” Electronics Letters, vol. 30, no. 25, pp. 2098–2099, Dec. 1994.
- [4] M. Steer, Microwave and RF Design, Transmission Lines, 3rd ed. North Carolina State University, 2019.
- [5] D. Porcino and W. Hirt, “Ultra-wideband radio technology: potential and challenges ahead,” IEEE communications magazine, vol. 41, no. 7, pp. 66–74, 2003.
- [6] “FCC (GPO) Title 47, Section 15 of the Code of Federal Regulations SubPart F: Ultrawideband,” www.access.gpo.gov/nara/cfr/waisidx_05/47cfr15_05.html.
- [7] J. Carson, “Notes on the theory of modulation,” Proc. of the Institute of Radio Engineers, vol. 10, no. 1, pp. 57–64, Feb. 1922.
- [8] L. Couch III, Digital and Analog Communication Systems, 6th ed. Prentice-Hall, 2001.
- [9] E. Armstrong, “A method of reducing disturbances in radio signaling by a system of frequency modulation,” Proc. of the Institute of Radio Engineers, vol. 24, no. 5, pp. 689–740, May 1936.
- [10] ——, “Radio telephone signaling,” US Patent US Patent 1 941 447, 12 26, 1933.
- [11] “Armstrong suit over fm settled. r.c.a. and n.b.c. to pay '\$1,000,000' ending action begun by late inventor,” New York Times, Dec. 31, 1954.
- [12] E. Bedrosian, “The analytic signal representation of modulated waveforms,” Proceedings of the IRE, vol. 50, no. 10, pp. 2071–2076, 1962.
- [13] C. Shannon, “Communication in the presence of noise,” Proc. IRE, vol. 37, no. 1, pp. 10–21, 1949.
- [14] J. Costas, “Synchronous communications,” Proc. of the IRE, vol. 44, no. 12, pp. 1713–1718, Dec. 1956.
- [15] F. Hearth, “Origins of the binary code,” Scientific American, pp. 76–83, Aug. 1972.
- [16] F. Gray, “Pulse code modulation,” US Patent US Patent 11 111 111, 03 17, 1953.
- [17] C. Savage, “A survey of combinatorial gray codes,” SIAM Review, vol. 39, no. 4, pp. 605– 629, 1997.

This page titled [23.12: References](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.10: References](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).

23.13: Exercises

1. Develop a formula for the average power of a signal $x(t)$. Consider $x(t)$ to be a voltage across a $1\ \Omega$ resistor.
2. What is the PAPR of a 5-tone signal when the amplitude of each tone is the same?
3. What is the PMEPR of a 10-tone signal when the amplitude of each tone is the same?
4. Consider two uncorrelated analog signals combined together. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = 0.1 \sin(10^9 t)$ and $y(t) = 0.05 \sin(1.01 \cdot 10^9 t)$. The combined signal is $z(t) = x(t) + y(t)$. [Parallels Example 2.2.3]
 - a. What is the PAPR of $x(t)$ in decibels?
 - b. What is the PAPR of $z(t)$ in decibels?
 - c. What is the PMEPR of $x(t)$ in decibels?
 - d. Is it possible to calculate the PMEPR of $z(t)$? If so, what is it?
5. Consider two uncorrelated analog signals combined together. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = 0.1 \sin(10^8 t)$ and $y(t) = 0.05 \sin(1.01 \cdot 10^8 t)$. What is the PMEPR of this combined signal? Express PMEPR in decibels. [Parallels Example 2.2.3]
6. What is PMEPR of a three-tone signal when the amplitude of each tone is the same?
7. What is PMEPR of a four-tone signal when the amplitude of each tone is the same?
8. A tone $x_1(t) = 0.12 \cos(\omega_1 t)$ is added to two other tones $x_2(t) = 0.14 \cos(\omega_2 t)$ and $x_3(t) = 0.1 \cos(\omega_3 t)$ to produce a signal $y(t) = x_1(t) + x_2(t) + x_3(t)$, where $y(t)$, $x_1(t)$, $x_2(t)$ and $x_3(t)$ are voltages across a $100\ \Omega$ resistor. Consider that ω_1 , ω_2 , and ω_3 are 10% apart and that the signals at these frequencies are uncorrelated.
 - a. What is the PMEPR of $x_1(t)$? Express your answer in decibels.
 - b. Sketch $y(t)$.
 - c. The combined signal appears as a pseudocarrier with a time-varying envelope. What is the power of the largest single cycle of the pseudo-carrier?
 - d. What is the average power of $y(t)$?
 - e. What is the PMEPR of $y(t)$? Express your answer in decibels.
9. Consider two uncorrelated analog signals summed together. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = \sin(10^9 t)$ and $y(t) = 2 \sin(1.01 \cdot 10^9 t)$ so that the total signal is $z(t) = x(t) + y(t)$. What is the PMEPR of $z(t)$ in decibels? [Parallels Example 2.2.3]
10. What is the PMEPR of an FM signal at **1 GHz** with a maximum modulated frequency deviation of **$\pm 10\ kHz$** ?
11. What is the PMEPR of a two-tone signal (consisting of two sinewaves at different frequencies that are, say, 1% apart)? First, use a symbolic expression, then consider the special case when the two amplitudes are equal. Consider that the two tones are close in frequency.
12. What is the PMEPR of a three-tone signal (consisting of three equal-amplitude sinewaves, say, 1% apart in frequency)?
13. A phase modulated tone $x_1(t) = A_1 \cos[\omega_1 t + \phi_1(t)]$. What is the PMEPR of $x_1(t)$? Express your answer in decibels.
14. What is the PMEPR of an AM signal with 75% amplitude modulation?
15. Two FM voltage signals $x_1(t)$ and $x_2(t)$ are added together and then amplified by an ideal linear amplifier terminated in $50\ \Omega$ with a gain of **10 dB** and the output voltage of the amplifier is $y(t) = \sqrt{10}[x_1(t) + x_2(t)]$.
 - a. What is the PMEPR of $x_1(t)$? Express your answer in decibels?
 - b. What effect does the amplifier have on the PMEPR of the signal?
 - c. If $x_1(t) = A_1 \cos[\omega_1(t)t]$ and $x_2(t) = A_2 \cos[\omega_2(t)t]$, what is the PMEPR of the output of the amplifier, $y(t)$? Express PMEPR in decibels. Consider that $\omega_1(t)$ and $\omega_2(t)$ are within 0.1% of each other.
16. An FM signal has a maximum frequency deviation of **20 kHz** and a modulating signal between **300 Hz** and **5 kHz**. What is the bandwidth required to transmit the modulated RF signal when the carrier is **200 MHz**? Is this considered to be narrowband FM or wideband FM?
17. A high-fidelity stereo audio signal has a frequency content ranging from **50 Hz** to **20 kHz**. If the signal is to be modulated on an FM carrier at **100 MHz**, what is the bandwidth required for the modulated RF signal? The maximum frequency deviation is **5 kHz** when the modulating signal is at its peak value.
18. Consider FM signals close in frequency but whose spectra do not overlap. [Parallels Example 2.4.1]
 - a. What is the PMEPR of just one PM signal? Express your answer in decibels.

- b. What is the PMEPR of a signal comprised of two uncorrelated narrowband PM signals with the same average power?
19. Consider two nonoverlapping equal amplitude FM signals having center frequencies within 1%.
- What is the PMEPR in dB of just one FM modulated signal?
 - What is the PMEPR in dB of a signal comprising two FM signals of the same power?
20. Consider a signal $x(t)$ that is the sum of two uncorrelated signals, a narrowband AM signal with 50% modulation, $y(t)$, and a narrow-band FM signal, $z(t)$. The center frequencies of $y(t)$ and $z(t)$ are within 1%. The carriers have equal amplitude. Express answers in dB.
- What is the PAPR of the AM signal $x(t)$?
 - What is the PAPR of the FM signal $z(t)$?
 - What is the PAPR of $x(t)$?
 - What is the PMEPR of the AM signal $x(t)$?
 - What is the PMEPR of the FM signal $z(t)$?
 - What is the PMEPR of $x(t)$?
21. Two phase modulated tones $x_1(t) = A_1 \cos[\omega_1 t + \phi_1(t)]$ and $x_2(t) = A_2 \cos[\omega_2 t + \phi_2(t)]$ are added together as $y(t) = x_1(t) + x_2(t)$. What is the PMEPR of $y(t)$ in decibels. Consider that ω_1 and ω_2 are within 0.1% of each other.
22. A radio uses a channel with a bandwidth of **25 kHz** and a modulation scheme with a gross bit rate of **100 kbits/s** that is made of an information bit rate of **60 kbits/s** and a code bit rate of **40 kbits/s**.
- What is the modulation efficiency in bits/s/Hz?
 - What is the spectral efficiency in bits/s/Hz?
23. A cellular communication system uses $\pi/4$ -DQPSK modulation with a modulation efficiency of **1.63 bits/s/Hz** to transmit data at the rate of **30 kbytes/s**. This would be the spectral efficiency in the absence of coding. However, **25%** of the transmitted bits are used to implement a forward error correction code.
- What is the gross bit rate?
 - What is the information bit rate?
 - What is the bandwidth required to transmit the information and code bits?
 - What is the spectral efficiency in bits/s/Hz?
24. A radio uses a channel with a **5 MHz** bandwidth and uses 256-QAM modulation with a modulation efficiency of **6.33 bits/s/Hz**. The coding rate is **3/4** (i.e. of every **4 bits** sent **3** are data bits and the other is an error correction bit).
- What is gross bit rate in Mbits/s?
 - What is information rate in Mbits/s?
 - What is the spectral efficiency in bits/s/Hz?
25. The following sequence of bits **0100110111** is to be transmitted using QPSK modulation. Take these data in pairs, that is, as **0100110111**. These pairs, one bit at a time, drive the **I** and **Q** channels. Show the transitions on a constellation diagram. [Parallels Example 2.8.1]
26. The following sequence of bits **0100110111** is to be transmitted using $\pi/4$ -DQPSK modulation. Take these data in pairs, that is, as **0100110111**. These pairs, one bit at a time, drive the **I** and **Q** channels. Use five constellation diagrams, with each diagram showing one transition or symbol. [Parallels Example 2.7.1]
27. The following sequence of bits **0100110111** is transmitted using OQPSK modulation. Take these data in pairs, that is, as **0100110111**. These pairs, one bit at a time, drive the **I** and **Q** channels. Show the transitions on a constellation diagram.
28. Draw the constellation diagram of OQPSK.
29. Draw the constellation diagrams of $3\pi/8$ -8DPSK and explain the operation of this system and describe its advantages.
30. How many bits per symbol can be sent using $3\pi/8$ -8PSK?
31. How many bits per symbol can be sent using 8- PSK?
32. How many bits per symbol can be sent using 16- QAM?
33. Draw the constellation diagram of OQPSK modulation showing all possible transitions. You may want to use two diagrams.
34. What is the PMEPR of a 5-tone signal when the amplitude of each tone is the same?
35. Draw the constellation diagram of 64QAM.
36. How many bits per symbol can be sent using 32QAM?

37. How many bits per symbol can be sent using 16QAM?
38. How many bits per symbol can be sent using 2048QAM?
39. Consider a two-tone signal and describe intermodulation distortion in a short paragraph and include a diagram.
40. A 16-QAM modulated signal has a maximum RF phasor amplitude of **5 V**. If the noise on the signal has an rms value of **0.2 V**, what is the EVM of the modulated signal? [Parallels Example 2.11.1]
41. Consider a digitally modulated signal and describe the impact of a nonlinear amplifier on the signal. Include several negative effects.
42. A carrier with an amplitude of **3 V** is modulated using 8-PSK modulation. If the noise on the modulated signal has an rms value of **0.1 V**, what is the EVM of the modulated signal? [Parallels Example 2.11.1]
43. Consider a 32-QAM modulated signal which has a maximum **I** component, and a maximum **Q** component, of the RF phasor of **5 V**. If the noise on the signal has an RMS value of **0.1 V**, what is the modulation error ratio of the modulated signal in decibels? Refer to Figure 2.8.21(b). [Parallels Example 2.11.1]

2.14.1 Exercises By Section

§12.21, 2, 3, 4, 5, 6, 7, 8, 910, 11, 12

§12.413, 14, 15, 16, 17, 18, 1920, 21

§12.522, 23, 24

§12.825, 26, 27, 28, 29, 30, 31

§12.932, 33, 34, 35, 36, 37, 38

§12.1139, 40, 41, 42, 43

2.14.2 Answers to Selected Exercises

5. **2.55 dB**

7. (e) **3.78 dB**

8. **0.00022 W**

15. no effect

20. (a) **6 dB**

20. (e) **0 dB**

22. (a) **4 bits/s/Hz**

43. **36.02 dB**

This page titled [23.13: Exercises](#) is shared under a [CC BY-NC](#) license and was authored, remixed, and/or curated by [Michael Steer](#).

- [2.11: Exercises](#) by [Michael Steer](#) is licensed [CC BY-NC 4.0](#).